DESIGNING A MODEL FOR A CORPUS-DRIVEN DICTIONARY OF ACADEMIC ENGLISH

IZTOK KOSEM

Doctor of Philosophy

ASTON UNIVERSITY

April 2010

ASTON UNIVERSITY

Designing a model for a corpus-driven Dictionary of Academic English

Iztok Kosem

PhD in Corpus Linguistics

2010

University students encounter difficulties with academic English because of its vocabulary, phraseology, and variability, and also because academic English differs in many respects from general English, the language which they have experienced before starting their university studies. Although students have been provided with many dictionaries that contain some helpful information on words used in academic English, these dictionaries remain focused on the uses of words in general English.

There is therefore a gap in the dictionary market for a dictionary for university students, and this thesis provides a proposal for such a dictionary (called the Dictionary of Academic English; DOAE) in the form of a model which depicts how the dictionary should be designed, compiled, and offered to students. The model draws on state-of-the-art techniques in lexicography, dictionary-use research, and corpus linguistics.

The model demanded the creation of a completely new corpus of academic language (Corpus of Academic Journal Articles; CAJA). The main advantages of the corpus are its large size (83.5 million words) and balance. Having access to a large corpus of academic language was essential for a corpus-driven approach to data analysis. A good corpus balance in terms of domains enabled a detailed domain-labelling of senses, patterns, collocates, etc. in the dictionary database, which was then used to tailor the output according to the needs of different types of student.

The model proposes an online dictionary that is designed as an online dictionary from the outset. The proposed dictionary is revolutionary in the way it addresses the needs of different types of student. It presents students with a dynamic dictionary whose contents can be customised according to the user's native language, subject of study, variant spelling preferences, and/or visual preferences (e.g. black and white).

Keywords: student, lexicography, EAP, online, dictionary use.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my supervisor Ramesh Krishnamurthy for all the valuable advice and guidance he gave me while I was writing the thesis. A very useful aspect of the collaboration with him were our (sometimes heated) discussions on various topics in lexicography and corpus linguistics, which gave me the opportunity to learn about how dictionaries and corpora are designed and produced, and to argue, and get feedback on, my own opinions and ideas. He also encouraged me to produce papers (some in collaboration with him), and to get involved in various projects and in teaching at Aston University, all of which equipped me with research, teaching, organising, and administrative experience that will without a doubt be very useful in my future career.

I am truly thankful to my beloved wife Karmen for always being there for me. She had complete faith in my ability to complete the thesis. She offered me encouragement when I needed it, and ensured that I took a break whenever the stress caused by working on the thesis started to endanger my health. I also thank my family and friends in Slovenia who were always tremendously supportive.

I would also like to thank many people who have helped me with various technical aspects of my thesis: the Sketch Engine team (Milos Husak, Vojtech Kovar, Jan Pomikálek, Pavel Rychly) for their help with various matters related to Sketch Engine, such as preparing my corpus (e.g. lemmatising, POS-tagging) for Sketch Engine and writing the XML template for exporting corpus data into the TshwaneLex dictionary-writing system. I am grateful to Adam Kilgarriff in particular for reducing the fee for access to my corpus in the Sketch Engine, and for making the TickBox Lexicography function available to me even before it was available in the normal version of Sketch Engine. I am also indebted to the TshwaneLex support team for advising me on the design of the XML template for importing the data from Sketch Engine, Steven (Zhuoxian) Chen for writing the programs for the clean-up of corpus texts, and Patrick Hanks for providing me with access to the database of the Pattern Dictionary of English Verbs before it was publicly available.

Special thanks go to Lizzy Tanguay for her proofreading services in the final stages of the thesis, and to Simon Krek for his comments and provocative questions that inspired several good ideas found in this thesis.

I am very grateful to Aston University (April 2006 to September 2007) and the Arts & Humanities Research Council (October 2007 to March 2009) for funding my PhD research. The funding relieved me of financial worries during my research. In addition, I would like to thank the School of Languages and Social Sciences and the Aston Modern Languages Research Foundation for funding various activities related to my research, such as conference participation, workshops attendance, licence fees for various programs that I used in my research, and the fee for a computer programmer (Steven Chen).

I must not forget to express my appreciation to academic staff and fellow students at the School of Languages and Social Sciences, for their suggestions and comments, for their willingness to help me with my research (for example, when piloting the final version of the questionnaire for the survey of dictionary use), and for sometimes helping me to stop thinking about the thesis for a while. I am also thankful to my colleagues and friends from the University of Birmingham who regularly invited me to their postgraduate seminars, and provided me with the opportunity to present my work outside my university.

Last but not least, I must thank Aston University, especially the support staff at the School of Languages and Social Sciences for their assistance in administrative matters, and all the students and staff that participated in the survey of dictionary use conducted for the purposes of this thesis.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AWL – Academic Word List (Coxhead, 2000)

BOS – Bristol Online Surveys

BSO – Brandeis Semantic Ontology

CPA – Corpus Pattern Analysis

DOAJ – Directory of Open Access Journals

EAP – English for Academic Purposes

EFL – English as a Foreign Language

ESP – English for Specific Purposes

GSL – General Service List (West, 1953)

HESA – Higher Education Statistics Agency

IPA – International Phonetic Alphabet

L1 – level 1 (domain label)

L2 – level 2 (domain label)

L3 – level 3 (domain label)

NNS – non-native speaker

NS – native speaker (or 'native-speaker' if used attributively)

PED – pocket electronic dictionary

UWL – University Word List (Xue & Nation, 1984)

Dictionaries:

AHD1 – American Heritage Dictionary, Second College Edition (1983)

AHD2 – American Heritage Dictionary, 4th edition (2000)

COBUILD – Collins Cobuild Advanced Learner's Dictionary (various editions)

COBUILD1 – Collins Cobuild Advanced Learner's Dictionary, 1st edition (1987)

COBUILD CD-ROM – Collins Cobuild Advanced Learner's Dictionary, 3rd edition (2001) on CD-ROM

CED CD-ROM – Collins English Dictionary, 1st edition (2004) on CD-ROM

CODCE – Compact Oxford Dictionary of Current English (2005)

COEDUCS – Compact Oxford English Dictionary for University and College Students (2006)

DOAE – Dictionary of Academic English (proposed in this thesis)

e-CALD – Cambridge Advanced Learner's Dictionary, 3rd edition (2008) online

e-LDOCE - Longman Dictionary of Contemporary English, 4th edition (2006) online

e-MED – Macmillan English Dictionary (2007) online

e-OALD – Oxford Advanced Learner's Dictionary, 7th edition (2005) online

LED – Longman Exams Dictionary (2006)

LED CD-ROM – Longman Exams Dictionary (2006) on CD-ROM

LDOCE – Longman Dictionary of Contemporary English (various editions)

LDOCE1 - Longman Dictionary of Contemporary English, 1st edition (1978)

LDOCE2 – Longman Dictionary of Contemporary English, 2nd edition (1987)

MEDAL1 – Macmillan English Dictionary for Advanced Learners, 1st edition (2002)

MEDAL2 – Macmillan English Dictionary for Advanced Learners, 2nd edition (2007)

MWCD – Merriam Webster Collegiate Dictionary, 11th edition (2003)

MWCD CD-ROM - Merriam Webster Collegiate Dictionary, 11th edition (2003) on CD-ROM

NODE – New Oxford Dictionary of English (1998)

NODE CD-ROM – New Oxford Dictionary of English (1998) on CD-ROM

OALD3 – Oxford Advanced Learner's Dictionary, 3rd edition (1974)

OALD4 – Oxford Advanced Learner's Dictionary, 4th edition (1989)

OED – Oxford English Dictionary

PDEV – Pattern Dictionary of English Verbs (Hanks, 2008)

W3 – Webster's Third New International Dictionary (2002)

Corpora:

BASE – British Academic Spoken English corpus

BAWE – British Academic Written English corpus

BNC – British National Corpus

CAJA – Corpus of Academic Journal Articles

ICLE – International Corpus of Learner English

LOB – Lancaster-Oslo/Bergen corpus

MICASE – Michigan Corpus of Academic Spoken English

MICUSP – Michigan Corpus of Upper-Level Student Papers

OEC – Oxford English Corpus

PERC – Professional English Research Consortium corpus

RAT – Reading Academic Text corpus

T2K-SWAL – TOEFL 2000 Spoken and Written Academic Language corpus

Glossary

attribute
In corpus tools (e.g. Sketch Engine), attribute is a term used for a property of a token, such as word, lemma, or POS-tag, which can be a part of the concordance display, or can be used as a criterion for selection (e.g. when building wordlists).

collocate
"A word which occurs in close proximity to a word under investigation is called a collocate of it" (Sinclair, 1991:170). (see also **collocation**)

collocation
Collocation is "a recurrent combination of words where one specific lexical item...has an observable tendency to occur with another...with a frequency far greater than chance" (Atkins & Rundell, 2008:302).

concordance
A computer generated display of corpus occurrences of the node (searched item) with context on either side. "The context can be selected on various criteria (for example counting the words on either side, or finding the sentence boundaries)" (Sinclair, 1991:171).

concordance line
A concordance line is each occurrence of the node (searched item) in context, displayed in the concordance.

conversion
A process in corpus compilation where texts in filetypes that are not TXT (e.g. HTML, PDF) are converted into TXT using a conversion program.

corpus
"A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research" (Sinclair, 2005:16).

entry
The term used for a section of a dictionary that contains the information (e.g. senses, examples) about a particular headword.

grammatical relation
A relationship between two words that is based on grammar. An example of a grammatical relation would be that of a verb and an object. In Word Sketch (the function of Sketch Engine), grammatical relations need to be specified before word sketches are built to enable automatic identification.

headword
The term has two meanings in this thesis:

a. In wordlists, headword is the canonical word of the word family. In many cases, this word is also the most frequent of all the word forms in the word family.
b. In dictionary design, headword is a single-word or multi-word item that is assigned an entry in the dictionary.

KWIC
Stands for Key Word In Context. It describes the frequently used concordance display in which concordance lines are aligned centrally around the searched item, which is offered in bold and/or colour to distinguish it from the context, and the context of usually 40 words (20 to the left and 20 to the right) is shown.

lemma
Lemma is the collective term for all the forms of the word (one form is selected as the canonical or lemma form). For example, the lemma JUMP is made of word forms *jump*, *jumps*, *jumping*, and *jumped*. Note that *jump* and *jumps* are word forms of both the verb JUMP and the noun JUMP – thus, lexis and not grammar is the criterion here.

lemma list
A list of lemmas in the corpus (see also **lemmatisation**).

lemmatisation
An automatic process in the preparation of the corpus in which each token is searched for in a prepared list of word forms and their lemmas, and assigned its lemma. This is useful for searches of corpus data – for example, if the corpus is lemmatised, and the concordance search for *jump* is performed, the concordance lines for *jump, jumps, jumping,* and *jumped* are displayed. If a token is not found in the list, it becomes a new lemma.

node
In Sketch Engine, the node is the item that is searched in the corpus, and displayed in the concordance.

part-of-speech tagging (POS-tagging)
An automatic process in the preparation of the corpus in which each token is assigned its word class marker. POS-tagging is often done at the same time as lemmatisation (e.g. with TreeTagger).

subcorpus
A part of corpus that consists of a collection of texts which share a common property or properties, for example domain (e.g. medicine, architecture), discourse (spoken, written), register (e.g. journalism, fiction), year of publication, etc.

sub-entry
The term used for a part of a dictionary entry that deals with a word class of the headword that is found in more than one word class. For example, the entry *attribute* has two word classes (verb and noun), so it contains two sub-entries, one for each word class.

token
This term is used for any string of characters (in the corpus) that is bounded by spaces. The string of characters does not have to be unique – for example, if *jumps* occurs twice in the corpus, it represents two tokens (but only one word form). Token is a term often used for corpus word counts.

tokenisation
An automatic process in the preparation of the corpus where all the tokens are identified, i.e. the boundaries of each word and punctuation mark are marked (Atkins & Rundell, 2008:87).

word form
"This term is used for any unique string of characters, bounded by spaces" (Sinclair, 1991:176), that has meaning (see also **lemma**).

word family
A word family consists of a stem and all closely related affixed forms, i.e. all inflected forms and the most frequent, productive, and regular prefixes and suffixes (Coxhead, 2000). This term is frequently used in association with wordlists that are compiled for pedagogical use. In a wordlist, a word family is represented by a single word (see **headword a.**). For example, in the Academic Word List by Coxhead (2000), the word family with the headword *accompany* contains word forms *accompanied, accompanies, accompaniment, accompanying,* and *unaccompanied.*

1. INTRODUCTION

Over the past three decades, academic English has attracted considerable attention from researchers. As Biber (2006:6) points out, much of the research into academic English

> "...has been motivated by applied concerns: as linguists have come to recognize that language characteristics differ dramatically from one register to the next, they have also argued that we should teach the specific kinds of language that a learner will need."

A great deal of the research has thus been driven by pedagogical motives, which are closely related to the dramatic increase in the number of students at universities where English is the language of instruction. For example, in the academic year 2007/2008, there were approximately 22.5 million students studying at universities in the four largest English-speaking countries (US, UK, Australia, and Canada)[1], which was six times more than in 1955.

Corpora and corpus techniques have played a key role in the research into academic English. Krishnamurthy and Kosem (2007:357) point out that corpora have proved useful in determining the features of academic English, and have stimulated research into register and genre. They go on to say that researchers such as Flowerdew and Peacock (2001), Benesch (2001), and Dudley-Evans and St. John (1998) remark that corpora have also helped to revive some earlier approaches, such as rhetorical analysis and pragmatic analysis. Corpus techniques have also helped researchers to identify not only the differences between different genres of academic English (e.g. Hyland, 1999), but also the differences between academic English and general English (e.g. Biber et al., 1999).

The fact that academic English is different from general English means that students face several difficulties because they need to learn how to use a language they know in new ways. Consequently, disciplines like English for Academic Purposes (EAP) and English for Specific Purposes (ESP) have started to play a more prominent role in higher education in the last few decades (Hyland, 2006). EAP and ESP courses have different names (the courses in the US are called freshman composition classes; the courses in the UK are known as pre-sessional

[1] There were over 18.2 million students in the US (U.S. Department of Education, 2008), over 2.3 million students in the UK (HESA, 2009), 1 million students in Canada (Statistics Canada, 2009), and nearly 1 million students in Australia (Department of Education Science and Training, 2005). The most current data available for Australia is from year 2003, but it can be observed that the number of students has increased steadily between 1996 and 2003.

and in-sessional courses), and while EAP courses focus on vocabulary and phraseology of specific disciplines, ESP courses stress genre-specific linguistic patterns.

But as implied by terms such as 'freshman' and 'pre-sessional', many universities offer EAP and ESP courses to students only at the beginning of their studies (and often not to all students), and in some cases such courses are not offered at all (Biber, 2006). Furthermore, these courses are usually not part of regular study (except in the USA), and students may therefore not be highly motivated to attend the courses to learn *about* language – the focus of students is on the contents of their study. As a result, students are more likely to use language tools which can help them with addressing language problems at the moment of encountering them. One of the most important language tools of this kind is a dictionary.

A great number of monolingual (English) and bilingual (English + second language) dictionaries have been produced, especially in the last two decades, both in terms of number and variety. This is especially true of advanced learners' dictionaries, which have been the source of some important lexicographic innovations, for example new definition styles and a corpus-based approach.

Thus, university students have been provided with valuable tools to tackle English language-related problems during their studies, but, as is argued in this thesis, not specifically for academic English. The problem is that the differences between academic English and general English, pointed out by researchers, are yet to be fully acknowledged by lexicographers. This problem is exacerbated by the unclear status of students as dictionary users – there is a great deal of research in dictionary use that uses students as subjects, but very few studies actually examine the dictionary use of students. As a result, not a lot is known about which dictionaries students use, and how they use – or misuse – them.

Students are therefore forced to use existing dictionaries that, although they may contain some useful information on the use of words in academic English, remain focused on the uses of words in general English (see 2.2.5.1). It seems that the time has come to recognise students as dictionary users in their own right, and provide them with a dictionary that is built with their needs in mind.

This thesis aims to provide a proposal for such a dictionary (A Dictionary of Academic English; DOAE hereafter) in the form of a model which will depict how the dictionary should be designed, compiled, and offered to students. The Model will draw on state-of-the-art techniques in lexicography, dictionary-use research, and corpus linguistics. An outline of how

the aims of the thesis will be achieved is offered in the following paragraphs, which provide a summary of each chapter.

Chapter 2 takes a closer look at the linguistic characteristics of academic English, and gives an overview of research into the registers, genres, vocabulary, and phraseology of academic English. The focus then shifts to the dictionaries currently available to students, and their suitability for the needs of students is evaluated in terms of their representative coverage of academic English. Next, the available research into the dictionary use of students and related types of users, is examined for its usefulness for the Model for DOAE proposed in this thesis. Similarly, an overview of existing corpora of academic English is undertaken, and their lexicographic potential analysed.

In Chapter 3, the plan for the design of the Model is outlined, followed by a discussion of the methodology used for designing the Model. First, the procedures and tools used to administer the survey conducted to compile the user profile are presented, including the questionnaire used in the pilot survey. Next, the design and compilation of the Corpus of Academic Journal Articles (CAJA hereafter), a corpus built especially for the purposes of the Model, is described in detail. This is followed by an overview of secondary data that was consulted in certain parts of the analysis, namely other academic and general corpora, and a selection of dictionaries currently used by students. Section 3.2.2.3 is dedicated to Corpus Pattern Analysis, which served as a source for evaluating the results and, more importantly, provided a framework for the analysis of word meanings. The chapter moves on to introduce two lexicographic programs used for the analysis, the corpus tool Sketch Engine (3.3.1) and the dictionary-writing system TshwaneLex (3.3.2). The functionality of both programs is looked at, with the stress being put on the features that were extensively used during the analysis. The chapter concludes by depicting the corpus-driven approach, and the rationale for choosing this particular approach for this Model.

The designer of any new dictionary needs to know its potential users, and this issue is addressed in Chapter 4 where the user profile is compiled. The user profile is based on the results of the online survey into the dictionary use of students at Aston University. Especially relevant are findings on dictionaries used, preferred dictionary formats, frequency of use of different microstructural features, and activities that dictionaries are used for. After the user profile is developed, its implications for the dictionary Model are briefly discussed.

Drawing on the user profile, Chapter 5 focuses on the macrostructure of the proposed dictionary, first explaining the choice of online (web-based) format. The remainder of the chapter discusses the building of headword list, explaining the criteria for selection of single-word headwords and multi-word headwords. Various factors in building the headword list are considered, for example the treatment of proper names, variant spellings, and homographs. Comments are also made on how to deal with problematic items such as rare words and items caused by errors in tokenisation. The final section of the chapter offers a brief discussion on the additional material for the dictionary.

The microstructure of the Model, the most important aspect of the dictionary design, is dealt with in Chapter 6. Before focusing on the analysis of data, the selection of the sample entries is described, and some of the key microstructural elements from the database are presented. The difference between the database entry and the dictionary entry is also explained. Then, the building of the database entry, consisting of three stages, is described in detail. The three stages are recording basic information, meaning analysis, and compiling the dictionary entry. The description of each stage is supported with examples taken from the sample entries. Moreover, an insight is offered not only into what information is recorded in the database, but also how it is recorded. Also, in the section on meaning analysis, an overview is given of issues that lexicographers need to be aware of when performing an analysis with Word Sketch (this Sketch Engine function is presented in detail in 3.3.1.2.2). The section on steps in compiling the dictionary entry includes a discussion of various microstructural features, dedicating considerable attention to decisions related to definitions and examples. The last two sections of the chapter explain the potential role of existing dictionaries and corpora in the creation of DOAE, and discuss the American English version of the proposed dictionary[2] respectively.

Chapter 7 contains the aspect of the proposed Model that is the most relevant for students as future users, namely the dictionary output. The solution proposed is to utilise the customisability and flexibility of TshwaneLex software, and the richness of information recorded in the database, and create different outputs (called style sets) for different types of students, with the main distinguishing characteristics of the students being their native language and their subject of study. In addition, four style sets with different options of style and formatting are proposed. Other customisable features of the proposed dictionary are discussed,

[2] British English version has been selected as the default version because the survey conducted for developing the user profile used UK students.

as well as the possibilities of allowing students to shape the dictionary contents for their specific purposes.

Chapter 8 offers a discussion of the research conducted in this thesis. The advantages of the Model are pointed out, and suggestions on how the Model could be published are made. Formats other than the online format are also briefly considered. The Model is also evaluated by making a comparison of the sample entries with the entries in existing dictionaries and in the Corpus Pattern Analysis database. The review of the Model concludes by indicating some potential enhancements to the Model. Next, the methodology used to design the Model is reviewed, pointing out both advantages and disadvantages. Following that is a discussion on implications that the Model is likely to have on lexicography and pedagogy. Lastly, recommendations are made for further research in areas identified as being particularly poorly covered until now.

In Chapter 9, conclusions are drawn which reflect on the features of the Model, and the most important findings made during the process of designing the Model. An awareness that the Model is not perfect is shown by identifying some of the shortcomings of the Model. Then, the Louvain EAP dictionary (Granger & Paquot, 2010) project, which is currently under way at the Université catholique de Louvain, is presented and compared to the Model for DOAE. This is followed by the discussion of how the outcomes of this research can help us to envisage the future of lexicography. The thesis concludes by considering whether its main aim – offering lexicographers a model to produce a dictionary that would meet the needs of students – has been successfully achieved.

2. LITERATURE REVIEW

This chapter aims to establish why there is a need for DOAE by looking at the characteristics of academic English, and by discussing why existing dictionaries fail to reflect these characteristics. Furthermore, the chapter looks at what role can existing research into dictionary use and corpora of academic English play in the development of the Model for DOAE.

2.1 Academic English

This section gives an overview of the existing research in academic English, focussing on registers and genres, vocabulary, and phraseology. Studies that examine the differences between academic English and general English, and linguistic differences between different academic disciplines are also presented. The section concludes by attempting to define a user of academic English, partly by trying to answer the question of whether there is such a thing as a native speaker of academic English.

2.1.1 Genres of academic English

Studies of different varieties of academic English frequently use terms such as 'register' and 'genre', but no consensus has yet been reached on the distinction between these two terms (see Lee, 2001 for a detailed discussion on the use of these terms in previous research). As a result, researchers prefer to consistently use one of these terms; for example, 'genre' is used by linguists such as Swales (1990; 2004), whereas 'register' is favoured by authors such as Biber and Conrad (e.g. Biber et al., 1999; Conrad, 2001; Biber, 2006). As works of both groups of authors are cited in this study, both terms are used.

There are two caveats regarding the use of 'register' and 'genre' in this study. Firstly, although Biber and others use the term 'register' also for broader categories of text, such as the two modes of communication in academic language (academic writing and academic speech), the term 'discourse' is preferred in this study. Secondly, when 'register' and 'genre' are used together, 'register' is considered to be a superordinate term of 'genre', again on account of the fact that 'register' tends to have a broader meaning.

2.1.1.1 Written discourse

Research into academic English has largely focussed on written academic discourse. Until recently, researchers have mainly studied genres produced by academics, and the genre of the research article in particular. Studies of research articles have explored evaluation and stance – especially hedging devices and classes of verbs – (e.g. Hunston, 1993; Hyland, 1994; Hunston, 1995; Hyland, 1996a; Hyland, 1996b; Hyland, 1998; Hyland, 2001; Flowerdew, 2002a; Hyland, 2002; Silver, 2003; Stotesbury, 2003; Tucker, 2003; Lewin, 2005), grammatical constructions (e.g. Swales et al., 1998; Hewings & Hewings, 2001; Koutsantoni, 2004), vocabulary (e.g. Nation, 1990; Schmitt & McCarthy, 1997; Coxhead, 2000; Schmitt, 2000), and phraseology (e.g. Gledhill, 1996; Biber et al., 1999; Charles, 2003; Swales, 2004; Groom, 2005). Similar linguistic features have recently been studied in other academic genres.

Some of the above-mentioned studies have used corpora that contained other genres in addition to research articles, for example books and book reviews, but all of these genres, while read by students, are primarily targeted at other academics. Researchers have therefore become interested in genres that specifically target students, arguing that the language in those genres is what students will encounter and need help with (Biber, 2006). One of the most comprehensive studies has been the TOEFL 2000 Spoken and Written Academic Language (T2K-SWAL) project that aimed to determine whether the TOEFL[3]exam tasks are representative of university language (Biber et al., 2004b). The written part of the corpus compiled for the purposes of the project consists of the following register categories (examples of genres are provided in brackets): textbooks, course packs (lecture notes, study guides, descriptions of assignments written by the instructor, and photocopies of articles and book chapters), course management (syllabi, written assignments, exams), and institutional writing (academic programme brochures, university catalogues, student handbooks, university magazine articles)[4].

More recently, an increasing number of studies have focussed on genres that students have to produce at universities. Many studies have researched a single genre, for example undergraduate essay (e.g. Myers, 2001), research proposal (e.g. Cadman, 2002), short answer (e.g. Drury, 2001), PhD thesis (e.g. Charles, 2003; Swales, 2004; Thompson, 2005), etc. One of the first comprehensive studies that attempted to identify various genres of student writing is *An Investigation of Genres of Assessed Writing in British Higher Education* (Heuboeck et al.,

[3] TOEFL or Test of English as a Foreign Language is the essential exam for non-native speakers when applying to universities in the United States.

[4] The information on the contents of the T2K-SWAL corpus can be found in Biber (2006b), Biber et al. (2004b), and Biber et al. (2002).

2008)[5]. Using the British Academic Written English (BAWE) corpus, thirteen 'genre families' were identified:

- case study
- critique
- design specification
- empathy writing
- essay
- exercise
- explanation
- literature review
- methodology recount
- narrative recount
- problem question
- proposal
- research report.

Each genre family contains several genres, the total number of genres being fifty. The main distinguishing characteristic is the discipline the genre is found in. For example, the genre family 'design specification' contains the following genres that are predominantly found in Computer Science and Engineering (Physical Sciences disciplines): application design, building design, database design, game design, label design, product design, system design, and website design.

Written academic genres are also internally complex, i.e. they contain different sub-genres. Some of the genre families and genres identified in the BAWE corpus can in fact be either a genre or a sub-genre (e.g. literature review can be a complete text, or it can be only a part of a text, such as a book or an article). Examples of sub-genres that have been studied are research article introduction (Gledhill, 2000), discussion section of a research article (Martinez, 2003), introductory book chapter (Freddi, 2005), discussion chapter of PhD thesis (Bitchener & Basturkmen, 2006), and conclusion chapter of PhD thesis (Bunton, 2005).

2.1.1.2 Spoken discourse

In comparison to written academic discourse, spoken academic discourse has been somewhat neglected by EAP/ESP researchers. Several studies in spoken genres have focussed on lectures (e.g. DeCarrico & Nattinger, 1988; several papers in Flowerdew, 1994; Thompson, 2003; Camiciottoli, 2004). But as with written discourse, researchers have started to acknowledge that other spoken academic genres also deserve attention. The T2K-SWAL project was one of the first projects that analysed a variety of spoken academic genres. The spoken part of this corpus contained categories such as class sessions, classroom management, labs/in-class groups, office hours, study groups, and service encounters (Biber et al., 2004b).

[5] Similar investigation of genres of student writing in the US is currently underway, based on the Michigan Corpus of Upper-level Student Papers (MICUSP).

The MICASE (Michigan Corpus of Academic Spoken English) project is the first comprehensive attempt to describe spoken genres in academic discourse, focusing on spoken academic discourse at US universities[6]. Several genres have been identified (in the project corpus, they are called speech events), and grouped into classroom events and non-classroom events:

a) classroom events:
   a. small lectures
   b. large lectures
   c. discussion sections
   d. lab sections
   e. seminars
   f. student presentations

b) non-classroom events:
   a. advising sessions
   b. colloquia
   c. dissertation defenses
   d. interviews
   e. meetings
   f. office hours
   g. service encounters
   h. study groups
   i. tours.

The BASE (British Academic Spoken English) project is the British equivalent of MICASE. However, the BASE project collected only two genres, lectures and seminars. Together though, the MICASE corpus and the BASE corpus have provided researchers with an excellent resource to analyse spoken academic discourse.

2.1.1.3 Variation among academic genres

Besides studying individual genres, researchers have been interested in differences and similarities between different genres. Studies of this type can be divided into three groups: studies that compare genres of the same discourse, studies that compare spoken genres with written genres, and studies that compare characteristics of a genre or genres in different disciplines.

As the focus of research has long been on written discourse, there are naturally more studies that compare written genres than studies that compare spoken genres. Studies have compared, and found differences between, published genres (e.g. Hyland, 1999; Groom, 2005), student writing and expert academic writing (e.g. Hewings & Hewings, 2002), and even student university writing and writing tasks at the IELTS test (Moore & Morton, 2005). Groom (2005) finds differences between the use of two grammar patterns *it v-link ADJ that* and *it v-link ADJ to-inf* in research articles and in book reviews. Hyland (1999) compares textbooks and research articles, two genres with different target audience (students and academics respectively), and

[6] The project has already led to numerous publications and conference presentations that focus on a certain aspect of academic speech (see http://micase.elicorpora.info/micase-publications-and-presentations for an up-to-date list of MICASE-based publications and presentations).

finds significant differences in the frequency of textual devices such as logical connectives (e.g. *in addition, but*) and referencing (e.g. *X argues, according to Y*), and of evaluative devices such as hedges (e.g. *perhaps, might*) and attitude markers (e.g. *surprisingly, I agree*). In their study of metadiscoursal *it*-clauses (hedges, attitude markers, emphatics, and attribution), Hewings and Hewings (2002) remark that students make greater use of *it*-clauses than experts. Moore and Morton (2005:63) find "important differences between the writing required at university and that required to pass the IELTS test."

Multi-dimensional studies have made a considerable contribution to describing and comparing genres of the same discourse. The multi-dimensional analysis compares various registers in terms of dimensions, where each dimension represents a set of specific features likely to co-occur in a text (for more on multi-dimensional analysis, see Biber, 1988; Biber, 2006). Multi-dimensional analysis of the BAWE data, while pointing to some similarities, has revealed significant differences between different genres of student writing (Nesi, 2009). On the other hand, multi-dimensional studies that involved analysis of spoken academic registers (e.g. Biber et al., 2002; Biber, 2006) have shown that "all spoken registers are similar in their typical linguistic characteristics" (Biber, 2006:223).

Multi-dimensional analysis has also been used to compare spoken academic discourse with written academic discourse. Biber (2006) makes a comprehensive comparison between spoken and written academic registers, and concludes that "spoken registers are systematically different from written registers, with respect to a wide range of vocabulary characteristics and lexico-grammatical features" (ibid.:177). One of the few registers where some similarities are found is student management registers, namely spoken classroom management and written course management.

Another area where multi-dimensional analysis has provided useful insights is disciplinary variation among academic registers. Biber (2006) compares five discipline categories in the T2K-SWAL corpus (Business, Engineering, Humanities, Natural Science, and Social Science), and finds that there are similarities between Social Science and Humanities (Biber names them 'academic disciplines'), and between Business and Engineering ('professional disciplines'), but remarks that these two groups of disciplines differ significantly in the use of several linguistic features. Natural Science disciplines are found to be somewhere in between, being similar to both groups, but with these similarities lying in different linguistic features. Similar differences between disciplines have also been found by the BAWE project team in their corpus of student writing (Nesi, 2009).

An even more important finding of the BAWE project is the difference in distribution of genre families across different discipline categories. Essay is the dominating genre family in the corpus, but dominates only in Arts and Humanities, and Social Sciences, whereas in Life Sciences and Physical Sciences genre families are more evenly distributed (Nesi, 2009). Furthermore, certain genre families are specific to one or more discipline categories. For example, design specifications are predominantly found in Physical Sciences disciplines. On the other hand, case studies are not found in Arts and Humanities[7].

Many studies have focussed on comparing linguistic features across disciplines in a single genre. Hewings and Hewings (2001) compare academic articles from four disciplines (Astrophysics/Astronomy, Business Administration, Geography/Environmental Sciences, and History) and find significant differences in the use of *it*-clause hedges. Charles (2003) reports on a difference in the selection and frequency of stance nouns in PhD theses in politics/international relations and material science. Groom (2005) points to phraseological differences between articles and book reviews in History and Literary Criticism. Hyland (2008) analyses sequences of four words in his corpus of articles, PhD theses, and MA dissertations from four different disciplines (Electrical Engineering, Business Studies, Applied Linguistics, and Biology), and while he identifies four four-word sequences that occur frequently across all the disciplines, he comments that most sequences are discipline-specific.

Academic language is linguistically very diverse across discourses, genres, and disciplines. Most significant differences in the use of academic language are found between spoken genres and written genres, between student writing and expert writing, and between certain disciplines. One of the rare similarities found is the one across all spoken academic genres.

2.1.1.4 Academic English compared to general English

Most studies of academic language have focussed on particular registers, genres, and disciplines. Those studies have not considered academic language as a whole, and compared its linguistic characteristics with other general registers of English. This has been done by Biber et al. (1999) in *Longman Grammar of Spoken and Written English*, a corpus-based study of four major registers of English: conversation, fiction, newspapers, and academic writing.
Especially noteworthy are the following findings (summarized from Biber et al., 1999; Biber, 2006):

[7] It needs to be acknowledged that some of these findings can be affected by availability of data during collection.

a) nouns, adjectives, and prepositions are very frequent in academic writing, more than in the other three registers;
b) grammatical features especially characteristic of academic writing are for example nominalizations, stance noun + *of*-phrase, noun phrases with (multiple) modifiers, extraposed *that*-clauses, and relative clauses with the relative pronoun *which*.
c) although verbs and adverbs are not very frequent in academic writing, certain categories of verbs and adverbs are most frequently found in academic writing, such as copula *be*, passive voice, present tense simple, linking adverbials (e.g. *however*);
d) several grammatical features are frequent only in academic writing (e.g. verbs with inanimate subjects, *that/those* + *of*-phrase, quantifier *each*, coordination tag).

The importance of Biber et al.'s (1999) study for academic English lies not only in the fact that the study describes grammatical features that are especially frequent in academic language, but also that it shows that academic English is different from other registers of English. The differences between registers are likely to present difficulties to students. For example, grammatical features that are frequent in academic language, but infrequent in other registers may be problematic due to students' lack of familiarity with their use and functions. On the other hand, students may overuse some grammatical features that are not common in academic language.

2.1.1.5 Academic genres - summary and implications

Academic English is far from homogenous. Both spoken academic discourse and written academic discourse consist of many genres and sub-genres. The genres and sub-genres are very varied in terms of their linguistic characteristics. The main differences between the genres are observed at mode level (spoken vs. written), disciplinary level, and proficiency level (student vs. expert). Such variation in academic language demands a range of linguistic skills from students.

Another difficulty for the students is that academic English is very different from registers of general English the students may have previously encountered. Several grammatical features, such as nominalization, are very specific to academic language, and not being familiar with their form and function can seriously impede comprehension or production of academic texts. Comprehension of an academic text also becomes problematic when students encounter unfamiliar words. This has been the focus of vocabulary studies, which are discussed next.

2.1.2 Vocabulary

Vocabulary and its acquisition has always been one of the main focus areas of research in academic English (Nation & Waring, 1997; Schmitt & McCarthy, 1997; Schmitt, 2000; Nation, 2001). The majority of the research is based on the notion that students need to understand 95% of the words in the text to gain reasonable comprehension (Nation & Waring, 1997; Nation, 2001). It is expected that students attempting to achieve this comprehension threshold will know high frequency words before starting their studies, and that they will acquire technical vocabulary during lectures and ESP courses. Researchers have therefore concentrated on studying academic vocabulary, namely vocabulary which "is common to a wide range of academic texts, and not so common in non-academic texts" (Nation, 2001:189). This has led to the production of many different corpus-based wordlists, the most recent and most often cited wordlist being Coxhead's Academic Word List (Coxhead, 2000).

Before taking a closer look at the research in academic vocabulary, it is important to discuss some other groups of words in the English lexicon. The reason for this is that these groups are often used as exclusion criteria when lists of academic words are being built. Coxhead and Nation (2001) divide the English lexicon into four groups: high frequency words, academic vocabulary, technical vocabulary (different for each domain), and low frequency words.

High frequency words comprise 2,000 word families and are based on West's (1953) General Service List (GSL). These words cover around 80% of words in an academic text (Nation, 2001).

Technical vocabulary, according to Coxhead and Nation (2001), does not exceed 1,000 words in any individual subject area, and represents 5% of words in an academic text. Indeed, Nation (2001) observes that technical dictionaries usually contain around 1,000 entries. However, Chung and Nation (2003) argue that a) technical vocabulary is much larger than 5% (according to their study, it can exceed 30% of words in a text), and b) that the percentage of technical vocabulary varies across disciplines.

Low frequency words, representing the final 5% of the words in an academic text, may be text-specific and/or occur only once or twice in a general corpus. The importance of low frequency words for students is considered negligible, as knowledge of the other three groups is expected to ensure reasonable comprehension of an academic text. Nevertheless, the term 'low frequency words' should be used very carefully. Coxhead and Nation (2001) give an example of

a five-million-word corpus compiled by Carroll, Davies, and Richman (1971) in which 40.4% of the words occurred only once. Of course, those 40.4% of the words could be called 'low frequency' only for that particular corpus. It is likely that many of those words were either technical or simply less specific to the topic of the texts in the corpus.

Academic vocabulary, covering approximately 10% of words in an academic text (Coxhead & Nation, 2001), is considered key to students' success in understanding academic language (Jordan, 1997; Nation, 2001; Hyland & Tse, 2007). The wordlists that have attempted to capture academic vocabulary are now presented and the value of wordlists for the purposes of this research discussed.

2.1.2.1 Academic vocabulary

The two most often mentioned lists of academic words are Xue and Nation's (1984) University Word List (UWL) and Coxhead's (2000) Academic Word List (AWL). The UWL consists of 836 word families[8] that are not in the first 2,000 word families of West's (1953) GSL. The authors of the UWL have actually combined four previous word lists made by Campion and Elley (1971), Praninskas (1972), Lynn (1973), and Ghadessy (1979). Campion and Elley, and Praninskas, identified words that occurred in several disciplines by using corpora, while the approach used by Lynn and Ghadessy involved "tracking student annotations above words in textbooks" (Coxhead, 2000:14). By combining these four studies, the UWL inherited their weaknesses, and created a new one, namely inconsistency in the methodology. The UWL was later replaced by the AWL (Coxhead, 2000) and even Nation himself admitted that Coxhead's word list is currently the best of all academic word lists (Nation, 2001).

The AWL consists of 570 word families and is based on the 3.5-million-word Academic Corpus. A word family had to fulfil three conditions to be included in the list (Coxhead, 2000). Firstly, it had to be outside the first 2,000 words in West's General Service List. Secondly, a word form in a family had to occur at least ten times in each of the four top-level domain categories (see section 2.4.2.1) and in fifteen or more of the twenty-eight subject areas. Thirdly, each word form in a word family had to occur at least 100 times in the Academic Corpus.

Coxhead offers coverage information in support of the AWL: 570 word families account for 10% of the total tokens in the Academic Corpus and only 1.4% of the tokens in the Coxhead 3.7-million-word corpus of fiction (consisting of fifty texts from the Project Gutenberg website).

[8] See Glossary (page 27) for the explanation of the term 'word family'.

According to Coxhead (2000), this proves that the word families in the AWL are typical for academic writing.

Coxhead's wordlist has several major weaknesses, some of which are shared with other similar wordlists, so we address them in more detail:

a) *Data.* A number of texts in the Academic Corpus were taken from four other corpora: the Learned and Scientific section of the Brown corpus (Francis & Kucera, 1979), the Learned and Scientific section of the Wellington Corpus of Written English (Bauer, 1993), the Learned and Scientific section of the Lancaster-Oslo/Bergen (LOB) corpus (Johansson, 1978), and the Academic Texts section of the MicroConcord Academic Corpus (Murison-Bowie, 1993). As a result, the Academic Corpus contains many old texts (e.g. the Brown corpus and the LOB corpus consist of texts from 1961) and incomplete texts (114 texts, accounting for 6% of the Academic Corpus).

   Furthermore, 64% of the texts originated in New Zealand, 20% in Britain, 13% in the US, 2% in Canada, and 1% in Australia. Although Coxhead (2000) argues that at least some authors might not have come from the country they published in, these percentages are unlikely to reflect the up-to-date situation in global academic publishing.

   The Academic Corpus has also been criticized for being biased towards certain disciplines, such as law and business (Hyland & Tse, 2007).

b) *Excluding 2,000 most frequent words.* Just as Xue and Nation (1984) did, Coxhead excluded the first 2,000 word families from the West's GSL. They all assumed that students should know these 2,000 word families before starting their studies, and that EAP courses need to focus on academic vocabulary. Yet, this approach is problematic as studies indicate that (non-native speaker) students may know only around 1,000 word families from the GSL (Nurweni & Read, 1999; Cobb & Horst, 2001; Ward, 2005: cited in Ward, 2009). Furthermore, several studies (Paquot, 2007; Hancioglu et al., 2008; Martínez et al., 2009) have identified a number words from the GSL which are "academic", i.e. are predominantly used with an academic meaning. Hanciouglu et al. (2008) also argue that the GSL has its own weaknesses (e.g. it is old, contains archaic words, does not contain all word forms), which affect the validity of any wordlist that excludes the word families from the GSL.

c) *Word families*[9]. Although the AWL is a list of 570 words, each word is a headword of a word family, and a word family is defined as "a stem plus all closely related affixed forms", where affix covers "all inflections and the most frequent, productive, and regular prefixes and suffixes" (Coxhead, 2000:218). The actual number of 'words' (i.e. word forms) in the AWL is therefore about 3,000 (Ward, 2009). Grouping word forms into word families raises several issues. Morphology is given priority over frequency, one of the main criteria in creating a wordlist. Some word families, frequent enough because of the frequency of all the word forms in the family, can thus contain infrequent word forms. For example, the word *validly* from the word family represented by headword *valid* occurs less than twice per million words in the 100-million-word British National Corpus (BNC), and six times per million words in the 600-million-word Corpus of Contemporary American English (COCA). Secondly, the most frequent word form in the word family is not always the headword. For example, *appendix* is the most frequent word form of the word family whose headword is *append.* Finally, as Ming-Tzu and Nation (2004) point out, it is often the case that word forms from the same word family are not related in meanings, for example *consist* (meaning 'contain') and *consistent* (meaning 'unchanging').

d) *Based on word forms only.* The AWL, as many other wordlists, does not provide information about meanings and word classes of words. This presents EAP teachers and students with a further task of identifying the relevant meanings of words before teaching/learning them. This becomes even more problematic considering that often words have different meanings in different disciplines (Hyland & Tse, 2007; Martínez et al., 2009). Also, students may choose to ignore some words on the list, because they have encountered them before and are not aware that the words are used with a different meaning in academic English.

e) *Based on single words.* One criticism of wordlists in general is that they simply list words, thus failing to take phraseology into account (Hyland & Tse, 2007; Paquot, 2007; Hancioglu et al., 2008; Durrant, 2009). Phraseological items such as multi-word units, collocation[10], and pre-fabricated sequences are undoubtedly an important characteristic of academic English, and are discussed in more detail in the next section (2.1.3).

---

[9] The list of word families in the AWL was obtained from http://www.uefap.com/vocab/select/awl.htm (Gillett, 2009).

[10] See Glossary (page 25) for the explanation of 'collocation'.

f) *Unsuitable for productive purposes.* Paquot (2007) argues that the AWL is more suitable for receptive purposes than for productive purposes. Paquot (Paquot, 2007) created a productively-oriented wordlist of 838 lemmas[11], and comparing 237 nouns (out of total 298 nouns) on her own wordlist with the GSL and the AWL, Paquot discovered that two-thirds were found in the GSL, and only one-third in the AWL.

By using a more up-to-date corpus, namely the MicroConcord corpus of Academic Texts (Scott & Johns, 1993), not excluding the top 2000 word families from the GSL, and using the notion of lemma rather than a word family for selection of the items on the list, Paquot addressed some of the weaknesses of the AWL.

Paquot did use some of Coxhead's methodology (e.g. frequency, range, and evenness of distribution) to refine her selection of lemmas, but her main criterion for selection was keyness (Scott, 1997). Key (academic) lemmas, extracted using the WordSmith Tools version 3 (Scott, 1999), were lemmas found significantly more frequently in the corpus of academic texts (the MicroConcord corpus) than in a large corpus of fiction writing.

Nonetheless, Paquot's wordlist shares some of the weaknesses of the AWL, for example it is based on a small corpus which contains incomplete texts. In addition, Paquot does not provide any explanation for her hypothesis that academic words are likely to be under-represented in the genre of fiction writing, which constituted a reference corpus for key lemma extraction.

g) *Uneven coverage across disciplines.* Although the AWL, combined with the GSL, is supposed to account for around 90% of the words in an academic text, recent studies suggest that the coverage of some disciplines is better than of others. Hyland and Tse (2007), for example, discovered that sciences are poorly covered, the GSL and the AWL accounting for only 78% of the words in their science subcorpus. Similarly low coverage has been discovered by Martinez et al. (2009) for the agricultural sciences. Hyland and Tse (2007) also point out that a high percentage of words (94%) in the AWL are irregularly distributed across their three subcorpora.

Despite many criticisms, the AWL remains useful in EAP courses because it helps the teaching of academic vocabulary to students of different disciplines. The usefulness of the AWL to ESP courses, however, is much more limited because the wordlist lacks sufficient focus on individual disciplines (Martínez et al., 2009). This has prompted calls for the creation

[11] See Glossary (page 26) for the explanation of 'lemma'.

of discipline-specific academic wordlists (e.g. Chen & Ge, 2007; Vongpumivitch et al., 2009), and so far, several discipline-specific wordlists have already been built, for example for medicine (Wang et al., 2008), engineering (Mudraya, 2006; Ward, 2009), and agricultural sciences (Martínez et al., 2009). In all four studies mentioned here, the overlap between the discipline-specific wordlist and the AWL was not very high (the lowest overlap was 26% in Ward's study, and the highest 62.5% in the study by Martinez et al.).

2.1.2.2 Final remarks on wordlists and academic vocabulary

There is little doubt that wordlists of academic words have some pedagogic value. But because wordlists do not provide any information about the meaning, word class, and phraseology of the words they contain, they are of little value to students. Wordlists can thus be considered merely a point of departure for EAP/ESP teachers who have limited time to try and teach students how to understand and use academic language.

Academic vocabulary is a notion that further attempts to narrow down the words that are common to many different disciplines, and is of particular interest to EAP teachers. ESP teachers, on the other hand, are more interested in discipline-specific academic vocabulary which represents frequent words typical of the discipline.

One of the major weaknesses of all the different types of wordlists that have been produced (academic wordlists, productively-oriented wordlists, discipline-specific academic wordlists, collocation wordlists) is that the methodology used to create them is aimed at excluding certain words, normally high frequency words and less frequent, technical words. Omission of high frequency words is based on the assumption that students 'know' those words already. Considering what knowing a word can entail (i.e. knowing all its meanings, collocations, patterns, synonyms, etc), it is highly unlikely any student will possess such knowledge. It is more likely, however, that students will know core meanings and patterns of the words, i.e. meanings which are normally most frequent in general English. Consequently, if a word has a different meaning or different distribution of meanings in academic English, it will probably present difficulties to students.

It could be argued that all the searches for the words that are most 'academic' and thus most useful to students have picked the wrong starting point. Instead of taking a list of words found in academic texts and attempting to narrow it down, it would be better to perform a semantic analysis of the words, also considering the differences between their roles in general

English and academic English. Such tasks are normally left to lexicographers and the helpfulness of existing dictionaries is discussed in Section 2.2.

2.1.3 Phraseology

Phraseology has become an important area of research in the past two decades (Cowie, 1998). Phraseology can be broadly defined as the study of multi-word expressions. Researchers have studied various types of multi-word expressions, from widely-accepted types such as pre-fabricated sequences, idioms and collocations, to more specifically-termed notions such as 'recurrent word combinations' (Altenberg, 1998), 'phraseological units' (De Cock, 1998), 'phrasal lexemes' (Moon, 1998b), 'lexical bundles' (Biber et al., 1999), 'formulaic sequences' (Wray, 2000), and 'lexical phrases' (Nattinger & DeCarrico, 1992). Most studies of phraseology in academic English have focussed on lexical bundles and collocations.

The term 'lexical bundle' was first used in Biber et al. (1999), and means a frequently occurring sequence of words which is usually not idiomatic in meaning, and is usually not a complete structural unit (Biber, 2006)[12]. Many studies have researched the use of lexical bundles in various discourses and academic registers (e.g. Biber & Conrad, 1999; Biber et al., 2003; Biber et al., 2004a; Cortes, 2004; Biber, 2006). Biber and Barbieri (2007) provide a comprehensive summary of the findings of these studies, supplemented with their own research; lexical bundles are described as occurring in all academic registers, but are more frequent in spoken academic registers than written ones. Biber and Barbieri however, point out that the distribution of types of lexical bundles, grouped according to their discourse function (stance expressions, discourse organizers, and referential expressions[13]), is considerably more varied, and depends not only on mode but also on communicative purpose. Comparing these findings with the findings of multi-dimensional analyses of academic registers (see 2.1.1.3), Biber and Barbieri (2007:282) argue that "lexical bundles are fundamentally different from other lexico-grammatical features in their patterns of use."

Further variation in the use of lexical bundles is found on the disciplinary level, and on the proficiency level. Hyland (2008) analyses the use of lexical bundles in four different disciplines and finds that more than 50% of lexical bundles are discipline-specific. Cortes

[12] A sequence of words needs to occur at least 40 times per million words, and in at least five different texts to be called a lexical bundle (Biber, 2006b).
[13] See Biber et al. (2004a) for a detailed description of the three types of lexical bundles.

(2004) shows that students and academics ('published authors') use lexical bundles differently; students use them less frequently and for different purposes.

Lexical bundles have discourse functions in academic language, and are therefore considered important for comprehension/production of academic language (Biber, 2006). This suggests that students need to be familiar with the functions of lexical bundles to be successful in their studies. One of the advantages of lexical bundles in terms of pedagogy is that they can be easily identified by using corpus techniques. But considerable variation in the use of lexical bundles means that teaching them would need to be very discipline- and/or register-specific. Furthermore, because lexical bundles are rarely complete units (e.g. *on the basis of, in relation to the*; as opposed to *on the other hand*), it is questionable whether students would find learning them useful.

Lexical bundles have been criticized for their narrow approach to the notion of collocation. Lexical bundles miss out on collocational relationships where the positioning of words is not fixed. Collocational studies thus attempt to fill this gap and enhance our understanding of the use of phraseology in academic language.

Many studies of collocation in academic language have focussed on discovering collocational patterns in a specific discipline and/or genre. For example, Gledhill (1996; 2000) studies Cancer research articles, Ward (2007) Engineering textbooks, Marco (2000) Medical research papers, and Williams (1998) Plant Biology research articles. Their findings are very similar; collocational patterns are pervasive in academic writing, and each academic discipline has many collocational patterns which are specific to that discipline, or are more frequent in that discipline than in other disciplines.

Some of the other studies that have looked at the use of collocations in various disciplines and registers have concentrated on particular phrases or particular types of phrases, *it*-clauses in particular (Hewings & Hewings, 2001; Hewings & Hewings, 2002; Oakey, 2002; Groom, 2005). These studies show that certain types of *it*-clauses are more common in some disciplines than others. More importantly, the same lexical phrases have different functions in different disciplines, or are used differently by students and academics. This is similar to the findings reached by studies of lexical bundles, which is not that surprising since *it*-phrases can be seen as semi-fixed word combinations, a stage between a lexical bundle and collocation.

Research suggests that collocations are very much domain-specific, so students from different disciplines need to be taught different collocational patterns. Recently, however, Durrant (2009) has built a list of collocations that are common to most academic disciplines.

Durrant's list consists of 1000 'academic collocations', namely collocational pairs "which appear significantly more frequently in academic than in non-academic texts" (Durrant, 2009:162). The list can be regarded as a phraseological equivalent to Coxhead's AWL. Weaknesses of the list, also acknowledged by Durrant himself, are that the functions of collocational pairs are not examined, and that it is limited to two-word collocations. Furthermore, Durrant finds that academic collocations on the list are far less useful to Arts and Humanities students, which partly supports the claim of the discipline-specific nature of collocations.

Researchers have also been interested in non-native speaker (NNS hereafter) collocational performance, arguing that collocations are the most common source of learner error (Howarth, 1996). Howarth (1996; 1998) compares NNS academic writing (postgraduate level) with native-speaker (NS hereafter) expert academic writing, and concludes that, on average, NNSs use around 50% less restricted collocations than NSs. Similarly, Granger (1998b) finds that NNSs use fewer collocations than NSs, underusing native-like collocations and overusing atypical collocations.

One of the more interesting findings of Howarth's (1996) study is that atypical collocations are also found in NS data. Howarth calls such occurrences deviations, but they should perhaps be regarded as errors because that is how they would be described if found in NNS data. Howarth acknowledges this by saying that even NSs have problems with phraseology in specialist areas; and academic language (considering all the disciplines and genres) can be regarded as an example of a specialist area.

Phraseology plays an important role in academic language. There is considerable variation in the use and discourse function of phraseology across academic disciplines and registers. Phraseology is another aspect of academic language that students are likely to have difficulties with, because even though they may encounter familiar words, those words will often be in unfamiliar combinations (Oakey, 2005). There are indications that NSs may not be immune to errors in phraseology in academic language. This suggests that NSs also need help with academic language, bringing the status of the NS as a model for student academic writing into question.

2.1.4 Users of academic English

Academic English, as discussed in the Introduction, is used by approximately 22.5 million students in the four largest English-speaking countries alone. NNS students are often

perceived as those in need of lessons in how to use academic English. NS students, on the other hand, normally receive less attention, and their writing is even used as a benchmark for analysis of NNS student writing (Howarth, 1996; Granger, 1998a). It is argued here, however, that no distinction should be made between NS students and NNS students, as they are all learners of academic English.

There is already evidence that the academic community is reconsidering the distinction between NS students and NNS students. Authors such as Hyland (2006) and Jordan (1997) have started to acknowledge that the needs of NS students should not be overlooked. What is more, several authors suggest that the needs of NS students are no different to the needs of NNS students (Biber, 2006; Hyland & Tse, 2007). In her study of the students' use of phraseological items in academic writing, Römer (2009) discovered little difference between NS students and NNS students.

Also, calls have been made to use expertise and not nativeness as the measurement of proficiency in academic language (Gabrielatos, 2005; Leńko-Szymańska, 2008; Tribble, 2008). It is suggested that the terms 'apprentice writers' or 'novice writers' or 'trainee academics' should be used for students, and 'expert writers' or 'established writers' for academics who have had their papers published in peer-reviewed publications.

US universities have already made a step towards equal treatment of NSs and NNSs by introducing a language course (called freshman composition or English composition) which is compulsory for all the students (McCreary & Dolezal, 1999; Biber, 2006). Unfortunately, such courses are currently of little benefit to the students because they emphasize personal narrative and personal opinion writing, which do not really reflect the language required in university courses (Biber, 2006).

Still, it would be wrong to suggest that the native language of a student does not play a role in learning academic language. Corson (1997) points out that because many academic words are Greek and Latin in origin, students with non-Romance native language background will have more difficulties learning how to use them. At the same time, Corson admits that "Graeco-Latin words in English tend to be opaque, even for most L1 language users" (ibid.:696). Of course, many would argue that NNS students are at a disadvantage due to their inadequate knowledge of linguistic features such as collocations. Yet, one disadvantage of NS students is that they are likely to be more confident of their language proficiency, and thus (mistakenly) think they need little or no help with academic English.

In sum, all students should be viewed as users or learners of academic English. Their native language background cannot be the deciding factor on whether they need help with academic language or not; a student's native language should be considered as just one of the factors that influence student performance, such as student's experience in (academic) writing, previous education performance, and subject of study. Such an approach is already found in certain corpora of student texts (e.g. BAWE, MICASE) which EAP/ESP practitioners can now use to identify common problems of students (both NSs and NNSs), and then tailor language instruction accordingly.

2.1.5 Academic English - summary and implications

Academic English is linguistically very complex, and this is evident in its vocabulary and phraseology. Differences in linguistic features are found not only between written academic discourse and spoken academic discourse in general, but also between written genres and, to a much lesser extent, between spoken genres. Linguistic variation is also manifested at discipline level. Words are not only used differently, they are often used in different functions/patterns.

Academic English is different from *general* English in terms of vocabulary and grammar. Certain words (see Coxhead, 2000) and grammatical features (see Biber et al., 1999) are much more frequent in (written) academic English than in general English, and are often considered characteristic of academic English due to their frequency in academic English, and relative infrequency in general English.

Because academic English has many different genres with different linguistic features, and because it is different from general English, students have many problems with comprehension/production of academic language. This is true for both NS students and NNS students as they all face using the English language in new ways.

EAP/ESP teachers attempt to address language needs of students by teaching them how to use academic language, but with limited success. The problem is that students need/want to focus on their studies rather than to learn academic language. There is also a problem of finding a balance between teaching students academic language skills and academic conventions, such as referencing.

Students will therefore need access to language tools to help them during their studies. One such important tool is the dictionary. Students can choose from a number of dictionaries, which come in many shapes and sizes. However, a wide range of dictionaries does not

necessarily mean they are valuable, especially as far as problems of academic English are concerned. In order to determine the actual usefulness of dictionaries to university students, the dictionaries available and some of their features need to be examined.

2.2 University students and dictionaries

There is a plethora of monolingual English dictionaries on the market, however only a few of them claim to specifically target university students. Each dictionary for university students selects its own set of features, which have ostensibly been designed specifically for students; however, one feature that is found in almost every student dictionary is a section on academic writing. Of course, many non-student dictionaries, e.g. advanced learner's dictionaries, may also contain a section on academic writing, or other features that may be useful for students, but the main target users of these dictionaries are not specifically university students.

2.2.1 Dictionaries for university students

American lexicography has been producing dictionaries for students – called college dictionaries – since late 19th century[14]. These dictionaries display the encyclopaedic features of other American dictionaries, but focus on the present day language (Béjoint, 2000) and include many scientific and technical terms. The most widely-used college dictionaries are the Merriam Webster Collegiate Dictionary (MWCD), currently in its 11th edition, and the American Heritage Dictionary (AHD2), currently in its 4th edition. College dictionaries are aimed at NS college students, and have many of the features of NS dictionaries in general. Nevertheless, at US colleges and universities, these dictionaries are equally used by NNS students, but not necessarily by choice (McCreary & Dolezal, 1999).

In Great Britain, on the other hand, dictionaries for university students have a relatively short tradition[15]. Unlike their US counterparts, UK publishers believe that there are differences between the needs of NS students and NNS students, and produce dictionaries aimed at each group. The latest additions to dictionaries for university students are the Compact Oxford English Dictionary for University and College Students (2006; COEDUCS), aimed at NSs, and the Longman Exams Dictionary (2006; LED), aimed at NNSs. LED represents a milestone in

[14] The first edition of the Merriam-Webster Collegiate Dictionary was published in 1898 (the information was obtained from http://www.merriam-webster.com/info/reform-timeline.htm).
[15] These dictionaries should not be confused with students' dictionaries, which have been around for some time, but are targeted at students in secondary schools.

lexicography, as it is the first dictionary to make use of an academic word list (the AWL by Coxhead, 2000).

Recently published, the Cambridge Academic Content Dictionary (2009) also points out the word families from the AWL to the users, but cannot be described as a dictionary for university students since it is targeted at secondary school students. And although some form of academic language is used in secondary education, the language demands of the students are mainly connected with technical vocabulary. Such a dictionary will therefore be of limited use at university level.

Since students constitute a large market, and are very likely to buy a dictionary, it is not surprising that other types of monolingual dictionary are promoting themselves as useful to students. For example, the latest edition of Oxford Dictionary of English (2005) is advertised as "ideal for anyone who needs a comprehensive and authoritative dictionary of current English; for professionals, students, academics, and for use at work or at home" (OUP website[16]). As such dictionaries are often more established and widely available, one cannot expect that students will confine themselves to dictionaries for university students. In fact, the reality, especially in the UK, is quite the opposite – students use anything but the dictionaries targeted at them. This is discussed in the next section.

2.2.2 Which dictionaries are students actually using?

Very little is known about the dictionary ownership of students in universities where English is the language of instruction. This is especially true of students at US universities. However, some conclusions can be drawn from studies involving US students that actually investigate more specific aspects of dictionary use (McCreary & Dolezal, 1999; McCreary, 2002; McCreary & Amacker, 2006). Students in the US are likely to own, but not necessarily regularly use, at least one college dictionary as "English departments in the US typically require that students … buy a mandated dictionary, or choose one from a list of dictionaries, commonly known as 'college dictionaries', for the freshman (1st year) composition course" (McCreary & Dolezal, 1999:108). The list of dictionaries often includes comprehensive NS dictionaries, such as the American Heritage Dictionary. It is noteworthy that no distinction is made between NS students and NNS students.

[16] The quote was obtained from http://ukcatalogue.oup.com/product/9780198610571.do on 25 September 2009.

More is known about the dictionary preferences of UK students, mainly owing to the studies by Hartmann (1999) and Nesi and Haill (2002). Hartmann (1999) provides a comprehensive survey of the dictionary use of students at a UK university. His subjects were 710 students, 579 NSs and 131 NNSs. NNS students came from 29 different language backgrounds. 70.4% of subjects were undergraduate students, and 13.4% were postgraduates (Masters and PhD students)[17]. Nearly half of the subjects were language students, studying English (20%) or Modern Languages (29.4%). Other disciplines listed were Engineering and Computer Science (18.3%), Education (16.9%), Business & Economics (8.5%). 5.9% of the subjects were attending courses at the English Language Centre.

Nearly all the students in Hartmann (1999) reported owning a general dictionary, 77.2% owned a bilingual dictionary, 66.2% a thesaurus, and 37.8% a specialist dictionary. The general dictionary was reported as used most frequently by 50.4% of the subjects, the bilingual dictionary by 39.8% of the subjects, while the thesaurus and specialist dictionary were used most frequently by only 5.1% and 3.7% of the subjects respectively.

These results should be interpreted with caution. For example, the question asking the subjects about their dictionary ownership did not include a separate option for a learner's dictionary. It is therefore possible that many subjects owning a learner's dictionary selected one of the other options, most likely 'general dictionary'. In addition, the results have undoubtedly been affected by the high percentage of language students. This is evident from the answers such as 77.2% of the subjects reporting ownership of a bilingual dictionary and 48% of the subjects claiming to own more than four dictionaries.

Whereas Hartmann's subjects were both NSs and NNSs, Nesi and Haill (2002) used only NNSs in their study. The strength of Nesi and Haill's study is its method; they attempted "to monitor dictionary use under somewhat more natural conditions" (Nesi & Haill, 2002:278) by allowing their subjects a free choice of dictionary, text, time of consultation and the words consulted[18].

Nesi and Haill's subjects (all students of Oxford Brookes University) were studying, or were about to study, a range of disciplines[19], and had different language backgrounds. Out of 89 subjects, 37 were either first-year international undergraduates or SOCRATES exchange students, while 52 were from the International Foundation programme. 39 students were from

[17] 16.2% of students are classified as 'other', but no explanation is provided about the type of students that fall into this category.
[18] As a consequence, the study took 3 years to complete.
[19] Although Nesi and Haill do not provide this information, the findings on the dictionaries used by the students suggest that Business, Science, Technology, Mathematics and Law were among the areas of study.

Asia, 22 from the European Union, 14 from Eastern Europe, 8 from the Middle East, 4 from South America, 1 from Africa and 1 from the USA.

Most of the students reported owning more than one dictionary. Out of 63 students who reported on their frequency of dictionary use, 60 claimed to use their dictionaries very frequently. The most frequently mentioned titles by Nesi and Haill (2002) are listed in Table 1.

Table 1. The most frequently mentioned dictionaries in the study by Nesi and Haill (2002).

Oxford Advanced Learner's Dictionary (5th and 6th edition)
Concise Oxford Dictionary (various editions)
Collins Cobuild Dictionary (1st and 2nd edition)
Longman Dictionary of Contemporary English (2nd and 3rd edition)
Longman Dictionary of English Language and Culture (1st and 2nd edition)
Collins English Dictionary (various editions)
Longman Language Activator (various editions)
Oxford Wordpower (1993 and 1997)
Chambers English Dictionary (1989 and 1996)

Nesi and Haill discovered two other frequently consulted dictionaries, namely the New Oxford Dictionary of English (1998), and the Oxford English Reference Dictionary (1996), by obtaining a list of dictionaries borrowed from the library.

Nesi and Haill (2002) find that although NNS students prefer a learner's dictionary, many of them (also) use a general NS dictionary. A small number of technical dictionaries are mentioned (e.g. Business and Science) but Nesi and Haill do not provide any specific details. The list of dictionaries does not contain any dictionaries for university students. This is not surprising because a) the study was conducted before the publication of COEDUCS and LED, and b) college dictionaries are targeted at US market and thus not promoted/popular in the UK.

Béjoint's (1981) study is one of the rare studies that focuses on the use of monolingual English dictionaries by students at a university in a non-English speaking country. Béjoint's subjects were 122 French students of English at the University of Lyon, a much more homogeneous group than the ones used by Nesi and Haill. The results showed that 96% of the students possessed at least one monolingual dictionary, with 40% using it at least once a day, and 52% at least once a week. The most popular monolingual dictionaries were OALD3 (45% of the students owning a monolingual dictionary had it), LDOCE1 (27%), and the Concise Oxford Dictionary (14%). The students reported a slight preference for EFL over NS dictionaries; however, Béjoint stresses the fact that not many students answered this question. Exhaustive coverage was said to be one of the advantages of NS dictionaries.

2.2.2.1 Secondary dictionaries

The focus of this thesis is mainly on dictionaries that cover a wide range of vocabulary because its aim is to identify a single dictionary that would meet the needs of students. Nonetheless, studies show that students also use other types of dictionaries, such as technical dictionaries, quite often. These types of dictionaries can be regarded as secondary lexicographic resources, as they present an addition to general-purpose dictionaries rather than an alternative. A quick overview of most frequently used secondary dictionaries is offered here.

2.2.2.1.1 Bilingual dictionaries

Dictionary-use research shows that bilingual dictionaries are still preferred, and used more frequently by NNS students, especially the ones with less advanced language skills (Tomaszczyk, 1979; Baxter, 1980; Bensoussan et al., 1984; Battenburg, 1989; Atkins & Varantola, 1998). Nevertheless, despite being used less, monolingual dictionaries are held in high esteem (Tomaszczyk, 1979; Béjoint, 1981; Nesi & Haill, 2002). Also, "monolingual dictionary use increases and bilingual dictionary use decreases with increasing linguistic sophistication" (Nesi, 2000:39). This is supported by the research of Tomaszczyk (1979), Neubach and Cohen (1988) and Atkins and Varantola (1998).

The studies mentioned so far share two main problems, the methodology used and the selection of subjects. The methods used by the studies have often probably affected the results, either for being unreliable (e.g. questionnaire), or for favouring a specific type of dictionary (e.g. a reading test puts more emphasis on decoding skills, which means there is less need to use a monolingual dictionary). The subjects of the studies are often language students (Tomaszczyk, 1979; Atkins & Varantola, 1998), which means that the findings are always going to have limited implications for all the students. Furthermore, several studies (Tomaszczyk, 1979; Atkins & Varantola, 1998; Hartmann, 1999) include translators, who are by definition more likely to use bilingual dictionaries.

2.2.2.1.2 Technical dictionaries

Technical dictionaries deal with the vocabulary of one or more disciplines. They are most often produced by terminologists or subject experts. Technical dictionaries usually cover only a few thousand words. Most of the entries are nouns as, according to Baudot and Clas (1984:49), "it is known that approximately 80% of the terminology of a special field is composed of noun groups."

Technical dictionaries differ from general dictionaries not only in coverage, but also in the manner of defining the words. In the words of Svensén (1993:22),

> "A definition of a concept in a technical dictionary is often more detailed than that of the same concept in a general dictionary, and experts often do not like the 'imprecision' of the definitions of technical terms as given by general dictionaries."

Technical dictionaries are virtually devoid of information needed for encoding. They do, however, often provide illustrations, which can be effectively used for both decoding and encoding purposes. Technical dictionaries have an educational function as well, however not a linguistic one – for example, by looking up the term *chi-square test*, the user will know how to make the necessary calculations.

The authors of technical dictionaries determine their target users by considering their level of subject expertise, rather than their language proficiency. Students are named as the target readership particularly often. Nevertheless, due to the complexity of the definitions, it can be expected that students, especially NNSs, will find technical dictionaries quite difficult to use.

Selecting the vocabulary of a particular subject field is by no means an easy task. Some terms are used only by experts, while others are part of everyday language – dictionary-makers have to be careful when determining the criteria on which words to include in the headword list. This also applies to any terms from other neighbouring subject fields that might be looked up by the user. For example, Business students are likely to encounter many terms from Economics. If they realize that many of the terms cannot be found in the dictionary of Business, they will be inclined to replace it with a general dictionary.

2.2.2.1.3 Thesauri

Thesauri or dictionaries of synonyms – although some lexicographers do not perceive these two terms as equivalent[20] – are onomasiological, i.e. based on the meaning of the words. There are two types of thesaurus: Rogetian and non-Rogetian (Knowles, 1988). The distinguishing characteristic of a Rogetian thesaurus is that it is based on concepts and not words. As Foxley and Gwei (1989:113) say,

> "it takes a concept at a time and lists terms which express that concept. There is one paragraph for each concept; each concept paragraph is sub-divided into different parts of speech, and within a part of speech, words are grouped together by the degree of synonymity between them."

[20] According to Svensén (1993:28), in a thesaurus, "words are grouped together according to conceptual connection, regardless of part of speech."

Despite having many advantages, Béjoint (2000:15) points out that a Rogetian thesaurus is very demanding on the user, as "it is based on an organization of human knowledge that is bound to vary from author to author."

The non-Rogetian type of thesaurus is simpler to use and probably more often encountered by students. Thesauri of this type list synonyms of each word (headwords are in alphabetical order), sometimes also providing antonyms and examples. Lists of synonyms tend to be quite long and it is usually the case that the level of synonymy decreases with each word on the list.

A thesaurus is an encoding tool and thus very useful for students who have to produce a great deal of written material during their studies. Using a thesaurus requires a relatively high level of language proficiency; the user needs to know the meaning(s) of the looked up words and of all the synonyms, their grammatical and collocational characteristics, and the subtler differences in meaning usage between the looked-up word and any of its synonyms. Without this knowledge, the user may need to look up the words and synonyms in a dictionary to obtain this information. This has prompted publishers to start including a thesaurus and a dictionary within a single book.

2.2.2.1.4 Specialist dictionaries

There are also dictionaries available that deal with a specific part of the lexicon or that focus on particular kinds of encoding information. These include dictionaries of idioms, dictionaries of phrasal verbs, dictionaries of collocations, production dictionaries, pictorial dictionaries, and others still.

Dictionaries of collocations have appeared only recently. The most widely known are the BBI Dictionary of English Word Combinations (Benson et al., 1997) and the Oxford Collocations Dictionary for Students of English (Lea, 2002). The importance of collocation information for language learners is evidenced by extensive treatment of phraseology in dictionaries for advanced learners, however, a dictionary that focuses solely on collocations is, as Béjoint (2000) and Klotz (2003) argue, a valuable resource for both NSs and NNSs.

The only representative of production dictionaries is the Longman Language Activator (2002). Rundell (1998:327-328) describes it as

> "a conceptually organized reference resource designed specifically to meet the encoding needs of learners (in this case, fairly advanced learners) It is organized around what cognitive psychologists would call 'basic-level concepts' (about 1000 of them), and the

look-up process essentially involves deciding on a broad meaning area, selecting the 'keywords' to which it is related, and then browsing and comparing sets of near synonyms."

The use of this resource requires not only advanced language skills but also advanced dictionary skills. The Activator's structure differs considerably from the structure of an average dictionary, which makes the search process quite demanding.

2.2.2.2 Dictionaries used by students – summary

Research indicates that, in the US, dictionaries for university students are owned, but not necessarily used, by many students (see 2.2.1). Elsewhere, dictionaries for university students do not seem to be used at all by students, who prefer general-purpose dictionaries such NS dictionaries (used by NS and NNS students) and learners' dictionaries (used by NNS students only). The finding that many NNS students also use NS dictionaries might be an indication that they find advanced learners' dictionaries lacking.

Students also use other types of dictionaries such as bilingual dictionaries, technical dictionaries, and thesauri (see 2.2.2.1), which may be an indication that general-purpose dictionaries do not completely meet their needs. But before we can start arguing the case for dictionaries for university students, we need to identify any differences between dictionaries for university students and their general-purpose competitors (NS dictionaries and learners' dictionaries).

2.2.3 Unique features of existing dictionaries for university students

Dictionaries for university students seem to be relatively unknown to students in general, with the exception of the USA, where they are 'introduced' to students by their universities. But the question is whether these dictionaries *should* be used by students. They are supposedly targeted at students, but do their macro- and microstructures reflect that? What distinguishes dictionaries for university students from general-purpose dictionaries?

In order to determine that, entries from three dictionaries for university students (MWCD CD-ROM, COEDUCS, and LED CD-ROM) were compared with corresponding entries in general-purpose dictionaries. The general-purpose dictionaries were from the same publisher as it was believed that the publishers, in the interests of economy, re-used some of their existing material (i.e. material in general-purpose dictionaries) when compiling dictionaries for university students.

Firstly, MWCD CD-ROM (2002), representing American college dictionaries, was compared with the unabridged edition of Webster's Third New International Dictionary (2002; W3). The noun entry for *progress* was selected for this comparison. The MWCD CD-ROM entry (Figure 2) is merely a shortened version of the W3 entry (Figure 1); certain (sub)senses have been omitted, a few definitions have been shortened and all the examples have been removed (greyed out text). The only new text is a slightly modified definition of subsense b under sense 1 (underlined).

Figure 1. The noun entry *progress* in W3.

1 a (1) : a royal journey or tour marked by pomp and pageant <a staff of clerks accompanied the king on his *progresses* -- F.M.Stenton> (2) : a state procession <at last all was ready for my *progress* -- George VI> b : an official journey or circuit <these men of law ... on a *progress* from court to court -- Van Wyck Brooks> c : a journeying forward : an expedition, journey, or march through a region : TOUR <balls, dinners and crowds of beautiful women attended his *progress* -- *Time*>
2 a : an advance or movement to an objective or toward a goal : purposeful getting or going ahead <when impeded in their *progress*, these people suddenly ceased muttering -- E.A.Poe> <a fishing boat made a slow *progress* -- Elizabeth Bowen> <*progress* to the presidency and chairmanship of the board -- *Current Biography*> b : a movement onward (as in time or space) : a forward course : PROGRESSION <the daily *progress* of the sun> <the *progress* of a disease> <we make *progress* -- we pass from night to morning -- Edmund Wilson>
3 *Scots law* : succession in right to a feudal estate : the abstract of title with the deeds evidencing such succession
4 a : the action or process of advancing or improving by marked stages or degrees : gradual betterment; *especially* : the progressive development or evolution of mankind <there was a general belief in inevitable and universal *progress* -- John Berger> <found in civil law principles ... the analogies that were needed to smooth the path of *progress* -- B.N.Cardozo> b : a theory that change from old to new is essential to progress
- in progress : going on : **OCCURRING** <entertained troops ... while the fighting was still *in progress* -- *Current Biography*> <with the beginning of healing already *in progress* -- Morris Fishbein>

Figure 2. The noun entry *progress* in MWCD CD-ROM.

1 a (1) : a royal journey marked by pomp and pageant (2) : a state procession b : a tour or circuit made by an official (as a judge) c : an expedition, journey, or march through a region
2 : a forward or onward movement (as to an objective or to a goal) : ADVANCE
3 : gradual betterment; *especially* : the progressive development of mankind
–in progress : going on : OCCURRING

Whereas MWCD CD-ROM derives its entries from a much larger W3 (MWCD – 165,000 entries; W3 – 470,000 entries), COEDUCS and its apparent source, the Compact Oxford Dictionary of Current English (CODCE), contain practically the same number of 'words, phrases, and definitions'[21]. Figure 3 and Figure 4 show the entry *choice* in both dictionaries (greyed out text in Figure 4 is shared by both dictionaries).

Figure 3. The entry *choice* in CODCE.

noun 1 an act of choosing. **2** the right or ability to choose. **3** a range from which to choose. **4** something chosen.
adjective 1 of very good quality. **2** (of language) rude and abusive.
— ORIGIN from Old French *chois*, from *choisir* 'choose'.

Figure 4. The entry *choice* in COEDUCS.

noun *1 an act of choosing. 2 the right or ability to choose. 3 a range from which to choose:* ***a menu offering a wide choice of dishes.*** **4** a person or thing that has or can be chosen: ***this disk drive is the perfect choice for your computer.***
adjective *1 of very good quality. 2 (of language) rude and abusive.*
– PHRASES **of choice** chosen as one's favourite or the best: ***champagne was his drink of choice.***
– *ORIGIN Old French chois.*

The entries contain identical definitions, with the exception of the definition for sense 4 under the noun part of the entry. COEDUCS offers its users additional, seemingly useful features, such as examples and a phrase. And while this deserves praise, a closer analysis of the examples reveals that they do not reflect typical language use. For example, *the perfect choice for your computer* in the example under sense 4 cannot be found in the BNC or the Bank of English, while an internet search produced only 14 hits[22], ten of them with the word *computer* being a modifier not a head noun, as in *the perfect choice for your computer needs/repair/configuration*, and three of them from websites quoting the dictionary entry. In addition, the BNC and the Bank of English do not contain any examples of the phrase *the perfect choice for* followed by a possessive adjective.

[21] COEDUCS contains over 144,000 words, phrases, and definitions; CODCE contains 145,000 words, phrases, and definitions (www.askoxford.com).
[22] The internet search was done on 1 October 2009.

Another potential issue is the distribution of examples – why are examples not offered for the first two noun senses? The first senses are presumably more important (why would they otherwise be offered first?) and, in this particular case, the abstract definition under noun sense 1, "an act of choosing", could definitely benefit from an example or two.

Finally, LED CD-ROM (2006) was compared to the Longman Dictionary of Contemporary English (2006; e-LDOCE). The selection of the entry for comparison was slightly less random than in the previous two comparisons. As LED CD-ROM advertises the use of the AWL (Coxhead, 2000) as one of its unique features, the entry for the verb *abandon* (Figure 5), the first word labelled as academic in the dictionary, was chosen for the comparison of the two dictionaries.

Figure 5. The verb entry *abandon* in LED CD-ROM.

abandon[1] *verb* W3 AC

1 to leave someone, especially someone you are responsible for → **abandoned:**
How could she abandon her own child?
2 to go away from a place, vehicle etc permanently, especially because the situation makes it impossible for you to stay SYNONYM leave; → **abandoned:**
We had to abandon the car and walk the rest of the way.
Fearing further attacks, most of the population had abandoned the city.
3 to stop doing something because there are too many problems and it is impossible to continue:
The game had to be abandoned due to bad weather.
*They **abandoned** their **attempt** to recapture the castle.*
*Because of the fog they **abandoned** their **idea** of driving.*
a refusal to abandon nuclear arms development
4 to stop having a particular idea, belief, or attitude:
They were accused of abandoning their socialist principles.
*Rescuers had **abandoned** all **hope** of finding any more survivors.*
Education leaders do not want to abandon California's commitment to affordable college education.
5 **abandon yourself to something** *literary* to feel an emotion so strongly that you let it control you completely:
She abandoned herself to grief.
6 **abandon ship** to leave a ship because it is sinking
—**abandonment** *noun* [uncountable]

The entry in LED CD-ROM is almost an exact copy of the entry in e-LDOCE (greyed out text is shared by both dictionaries), with the following minor additions/changes: the label AC (denoting academic use), two links to a related word (*abandoned*), the label 'synonym' (in e-LDOCE, the sign = is used), and two additional examples. Examples were added to senses 3

and 4; if this is an indication of the importance of those senses in academic English, why are those senses not offered as senses 1 and 2? Also, if those are the academic senses of the verb *abandon*, the label AC should be placed only in front of those senses.

These comparisons point to the fact that dictionaries for university students are nothing more than a poorly adapted version of general-purpose dictionaries. An indication of the differences between both dictionary types is found in the publishers' promotional material, which puts more emphasis on additional material (sections on academic writing, how to write a CV, etc), rather than on dictionary macrostructure and microstructure.

2.2.4 Which dictionaries should students (not) be using?

The previous sections have shown that students use different types of monolingual dictionaries, and that the only dictionaries for university students seemingly used by students are American college dictionaries. Considering that dictionaries for university students are not that different from general-purpose dictionaries, students using general-purpose dictionaries do not seem to be missing much.

Several studies have attempted to determine whether the dictionaries that students use are actually suitable for students, and/or have provided advice on which dictionaries students should be using.

McCreary and Dolezal (1999) tested the usefulness of one of the college dictionaries, the American Heritage Dictionary, 2nd College Edition, for NNS students. Their aim was to determine the legitimacy of the prescriptive stance adopted by American universities regarding dictionary use. A vocabulary test was administered to 74 undergraduates at the University of Georgia. The test included 17 questions ("hard words" picked by a group of NNS students during the pre-testing process) and the subjects were asked to select one of the synonyms (13 questions) or antonyms (4 questions) offered. Five answers were available for each question, including an "I don't know" option to eliminate guessing. McCreary and Dolezal report on the difficulties that some students had in understanding the dictionary definitions, and acknowledge that the shortcomings of the dictionary may have affected the results. The authors then examine some of the definitions in advanced learners' dictionaries and find them easier to comprehend. In their conclusion, they suggest that the practice of prescribing dictionaries at US universities should be abandoned not only for NNSs but also for NSs,

> "There is no educationally sound reason for any English department to continue to require any of the current crop of American college desk dictionaries. American students who are native speakers of English may also benefit from the US versions of the monolingual English learners' dictionaries originally developed in the UK. These students may yet discover that learners' dictionaries, perhaps in an electronic format inside a word processor, are, in fact, the best avenue for building their vocabulary skills." (McCreary & Dolezal, 1999:134)

A similar test to the one used in McCreary and Dolezal (1999) was administered to NS students at US universities in studies by McCreary (2002, 2008) and McCreary and Amacker (2006). The studies compared the usefulness of college dictionaries and advanced learners' dictionaries. The findings confirm McCreary and Dolezal's (1999) claim that NS students would also benefit from using US versions of advanced learners' dictionaries. The three main problems of college dictionaries are identified as:

> "the use of low frequency words in the defining language, the ordering of senses, and the density of the column with many typefaces, abbreviations, multiple pronunciations, and etymological information." (McCreary & Amacker, 2006:875)

Other issues mentioned include the use of tildes, and the non-use of corpus data (especially for ordering senses).

The latest research by McCreary (2008) is another study with NS students as subjects that used a test with 'hard words'. What makes it different from the other studies is that its comparison of dictionaries includes a NS dictionary (the New Oxford American Dictionary, 2nd edition, 2005), in addition to college dictionaries (MWCD and AHD1), and an advanced learner's dictionary (MEDAL1). The results indicate that the New Oxford American Dictionary is the most useful dictionary for students, followed by MEDAL1, whereas the two college dictionaries are found to be significantly less useful. McCreary is especially critical of college dictionaries for their historical ordering of senses, complex definition language, and lack of examples and collocations.

The testing method employed by McCreary and Dolezal (1999), McCreary and Amacker (2006) and McCreary (2002; 2008) is also open to criticism. The researchers examined the items, marked by the subjects during the pre-test, and labelled them as "hard words", assuming those words to be the ones that the subjects would be most likely to look up. It is noteworthy that none of the hard words used in the studies appears on the AWL (Coxhead, 2000). This is perhaps an indication that the text used in the study was not a typical example of academic English.

Furthermore, the vocabulary test examined only the decoding skills of the subjects. The multiple-choice format of the test suited the use of college dictionaries, as they often use synonyms in their definitions – and McCreary and Dolezal (1999) admit that some of the potential synonyms in the multiple-choice test were indeed taken from the definitions in the college dictionary.

Other studies have focussed exclusively on NNSs. Williams (2006) examines the value of (advanced) learners' dictionaries for NNS students of science subjects. Looking at the treatment of technical vocabulary, Williams offers examples of inconsistent labelling, poor exemplification (or the lack of it), and inappropriate definition style. He concludes that the needs of NNS students of science subjects are not catered for. These findings were confirmed by Williams' follow-up study (Williams, 2008), which focussed on the treatment of scientific verbs in an advanced learner's dictionary.

Williams also raises an important issue about the corpora used by the compilers of advanced learners' dictionaries by pointing out that,

> "all the modern learner dictionaries are... based on general language reference corpora. The aim of the reference corpus is to give a representative picture of the language at a given time. This means covering a wide variety of genres so as to give a balanced coverage. Consequently, sciences and technical fields are represented as genres rather than by field." (Williams, 2006:797)

Williams calls for the use of corpora with a deeper coverage of scientific language that would give us a better idea about the role that scientific words should have in advanced learners' dictionaries.

Complementary to Williams' research is a study by De Cock (2006), who compares the treatment of business vocabulary in five advanced learners' dictionaries and two specialised learners' dictionaries. Examining the treatment of 17 items, De Cock reveals a lack of consensus in labelling among the advanced learners' dictionaries. Moreover, some technical items receive less prominence by being presented as sub-entries. On the other hand, the definitions, examples, and collocations in the two specialised dictionaries display more comprehensive lexicographic treatment which, combined with a better coverage, makes the dictionaries suitable for both encoding and decoding purposes.

Despite proving that the treatment of technical vocabulary is better in the specialised dictionaries, De Cock recommends that NNS (business) students should use a general purpose learners' dictionary instead, explaining that they will also need to look up a lot of 'general'

words, which makes specialised dictionaries less suitable for their needs. In addition, she makes a call for an electronic learner's dictionary that would offer exhaustive treatment of both general and specialised English.

One major weakness of De Cock's study is that she focuses on a subject field (Business) whose vocabulary is full of terms that occur in everyday language. It is therefore more likely that such terms will occur more frequently in a corpus, and receive better treatment in a learner's dictionary. The study of a subject field with a higher proportion of words that do not occur in everyday language would probably yield less favourable results for learners' dictionaries, and this would, of course strengthen De Cock's call for a new type of (learner's) dictionary.

Another study often cited in the EAP literature is West's (1987) survey of 19 English Language Teaching dictionaries available at that time. The aim of the survey was to review and compare the dictionaries, and determine which ones NNS students should use. The reviews were conducted by English teachers from Britain and overseas. Categories of information analysed were: UK price, format and date, number of pages, level/coverage, workbooks, pronunciation, ease of use, definitions, grammatical assistance, illustrations, and appendices.

The dictionaries recommended for purchase were LDOCE1 for advanced learners, the Longman Active Study Dictionary for intermediate learners, and the Oxford English Picture Dictionary for elementary learners. The Harrap's Mini Pocket English Dictionary and the Oxford-Duden Pictorial English Dictionary were thought to be the best among pocket dictionaries and ESP dictionaries respectively.

The main value of West's survey lies in the fact that the reviewers included non-content factors in the dictionary criteria, and thus adopted a more user-like approach. Price, visual image, and paper quality perhaps do not mean a great deal to lexicographers and researchers, but these characteristics frequently play a key role in the process of purchasing a (paper) dictionary.

The selection of the pictorial dictionary as the best ESP dictionary points to the value of illustrations in explaining terminology. As West (1987:65) argues, "as most specialist words are nouns, these can usually be explained more effectively by illustration than by definition."

Although NS dictionaries were not included in the survey, West reports that it was thought that students – and not only very advanced ones – should also have access to a native speakers' dictionary. The Collins English Dictionary (1979) was recommended due to its coverage, reflection of modern language, and the use of the International Phonetic Alphabet for

pronunciation (whereas US dictionaries and UK school dictionaries tend to favour 'respelling' to represent pronunciation). Interestingly enough, the Collins English Dictionary shared some features (e.g. encyclopaedic entries, coverage) with American college dictionaries, which are targeted at NS students.

The reviewers saw bilingual dictionaries as a hindrance, feeling that they encourage students to believe that each word has a direct equivalent in another language. This view is echoed in the research of other authors, including Baxter (1980), Béjoint (1981), Ard (1982), Koren (1997), and Rundell (1999).

The major drawback of West's survey is that the information is now out of date; the dictionaries reviewed were published between 1972 and 1984. Furthermore, the reviews were based on opinions that were not backed up any research but are rather based on personal intuition. What is more, West fails to provide any details about the number of teachers involved in the survey.

The research review suggests that general NS dictionaries and advanced learners' dictionaries are recommended to students. The use of dictionaries for university students, especially college dictionaries, is not advised. It is noteworthy, however, that most of the studies mentioned above compared dictionaries and thus aimed to determine which dictionaries are *more* suitable for students, rather than to test whether the dictionaries were even suitable for students. And the findings of the studies show that all the types of dictionaries, recommended or not, have certain problems in meeting the needs of students. These problems are discussed in the next section.

2.2.5 Problems of existing dictionaries that students use

The following section takes a closer look at some of the main macrostructural and microstructural features of three types of dictionaries used by students: dictionaries for university students, NS dictionaries (general-purpose dictionaries targeted at NSs), and learners' dictionaries (general-purpose dictionaries for NNSs). Comments are also made about any dictionary features that specifically target students.

2.2.5.1 Corpus-based

Corpus-based dictionaries have now become the norm in lexicography (Kilgarriff, 2000). Lexicographers use corpora to obtain information on the frequency of words and phrases, to discover their meanings, and to find examples of authentic usage. All three types of dictionaries

analysed here are based on corpora of general English. And although academic texts are found in any general English corpora, they are often incomplete, and suffer from a lack of representativeness (Thompson, 2006). As a result, academic words (or to be more exact, academic word meanings) have a less prominent role.

2.2.5.2 Coverage

The coverage in general NS dictionaries is very comprehensive – this can be positive for students as technical terms and meanings are included. However, the negative aspect is that the dictionaries include words and meanings for many non-academic communications which can hinder navigation through the entries.

The coverage in learners' dictionaries is significantly reduced, when compared with general NS dictionaries. The focus is on more frequent words and meanings. Rarer words and meanings are excluded, which is beneficial to students. Exclusion of many technical terms and meanings is much more problematic as students are likely to encounter terminology frequently during their studies. Furthermore, as frequency is likely to be a key factor in deciding which word or sense to exclude, many academic uses of the words may not be covered by a learner's dictionary due to the non-academic nature of most of the corpus data.

The coverage in dictionaries for university students resembles the coverage in learner's dictionaries, meaning that many rarer words and technical words are excluded. The exception is college dictionaries which resemble NS dictionaries in their comprehensive coverage of terminology.

2.2.5.3 Sense ordering

Considering the 'choose the first definition' strategy often adopted by users (Mitchell, 1983; Tono, 1984; Neubach & Cohen, 1988; McCreary, 2002; Nesi & Haill, 2002), word senses in any dictionary for students should be ordered by their frequency in academic language. But senses in these types of dictionaries are ordered according to information from general corpora, which does not constitute a user-friendly feature for students.

Students thus face potentially long searches through entries for the relevant academic senses of the words. Another related problem is that the position of a sense in the sense ordering plays an important role in deciding the amount of space the sense gets in the entry; so if an academic sense of the word is not that frequent, it receives less comprehensive treatment (e.g. fewer examples).

Sense ordering is problematic in all three types of dictionaries discussed here. But college dictionaries such as MWCD and AHD1 are especially unfriendly to student users, because they order senses historically, with the oldest sense given first.

2.2.5.4 Definitions

Definitions in general NS dictionaries are relatively short and contain rarer words, and therefore require high(er) language proficiency. This is likely to present problems to students, especially NNSs, whose vocabulary knowledge is not at the level expected from the target users of these dictionaries.

Much less challenging are the definitions in learner's dictionaries, which often use a restricted defining vocabulary. An especially informative type of definition is the full-sentence definition (first introduced by COBUILD), which provides encoding information in addition to explaining the meaning.

In comparison to definitions in general-purpose dictionaries, definitions in dictionaries for university students are left almost unchanged. Whenever a change is made, definitions usually become shorter. Students are therefore seen as having the same language competence, if not even higher, than the NS users of general-purpose dictionaries.

2.2.5.5 Examples

In NS dictionaries, examples are rare and usually consist of short phrases rather than full sentences. On the other hand, learners' dictionaries provide plenty of examples, using them to not only support the definition (i.e. provide additional decoding information), but to demonstrate the usage and phraseology of the word (i.e. provide encoding information). Another advantage of examples in learners' dictionaries is that they are normally provided as full sentences.

Exemplification is one of the features where dictionaries for university students demonstrate the highest degree of departure from their general-purpose counterparts (see 2.2.3). MWCD CD-ROM tends to omit all the examples found in W3, something which is difficult to understand, especially in view of the fact that the definitions are already rather brief.

COEDUCS, on the other hand, has enriched CODCE entries with examples. Unfortunately, the examples seem to be poorly selected, which decreases their value. In addition, the practice of adding examples to some senses, but not the others (normally more important/prominent ones is highly questionable.

LED is already based on a dictionary that provides plenty of examples, so its practice of adding examples to the entries, especially academic senses of the entries, is commendable. Unfortunately, additional examples are offered at the end of the sense, so they become useful to the users only if they bother to read all the examples under the sense. Furthermore, dictionary examples are rarely taken from academic discourse, a weakness that applies to all the used by students.

2.2.5.6 Features targeted at students

A few dictionaries for NNSs offer features targeted specifically at university students (additional material such as a section on how to write an academic essay is not counted here). Such features are labelling academic words (words from AWL), found in LED, and 'Academic writing' boxes, found in MEDAL1.

LED labels its headwords as academic if they occur in Coxhead's AWL, to alert students to the academic status of these words. It is laudable that a product of EAP research has been incorporated into the dictionary. However, simply labelling words as academic creates a number of problems. As was already remarked in 2.2.3, the label AC is placed at the beginning of the entries rather than specifically in front of the academic senses. In addition, students are led to believe that 'academic' words are more important than other words, and may overuse them. Also, AWL itself has several shortcomings, such as automatically excluding the 2,000 most frequent words (see 2.1.2.1 for more).

MEDAL1 provides Academic Writing boxes which are effectively Usage boxes for words frequently encountered in academic writing. Academic Writing boxes provide more detailed information about how to (correctly) use the word, offer a list of alternative words (synonyms), and describe the subtler differences between the word and some of its (near)-synonyms. Students will find such information very useful when writing academic texts. Unfortunately, Academic Writing boxes are not a standard entry feature, i.e. they are not offered at many entries.

The above analysis of a few core dictionary features has shown that none of the three types of dictionaries are very suitable for students. NS dictionaries are (too) comprehensive in coverage, contain linguistically-demanding definitions, and offer little encoding information. Learners' dictionaries are much more user-friendly, offering less complex definitions and a great deal of encoding information. Shortcomings of learners' dictionaries are small coverage, especially of terminology, and their focus on NNSs (NSs are unlikely to use them). Finally,

dictionaries for university students are no more than a spin-off of general-purpose dictionaries. It often seems that no significant lexicographic effort has been put into the creation of the dictionary for students. And when any features designed especially for students are added (e.g. labelling academic words), their weaknesses tend to outweigh the benefits.

The main problem of all three types of dictionaries is that they are based on corpora of general English, rather than academic English. As a result, the entries do not reflect the semantic and lexico-grammatical properties of words as used in academic English. Also, general senses of the words are given more prominence than academic senses.

Publishers and lexicographers clearly do not consider university students as a group of dictionary users with their own specific needs. The same criticism can be levelled at researchers who have studied dictionary use; they often use students as their subjects, but seldom focus on the needs of students as users of academic English.

Publishers and lexicographers have thus yet to incorporate the findings of EAP/ESP researchers which have started to reveal the linguistic and genre variations in academic English, and the differences between academic English and general English. As a result, there is currently no dictionary on the market that offers a comprehensive description of academic English.

2.2.6 Calls for a dictionary of academic English

It is important to note that the idea of a dictionary dealing with academic language is not new. More than two decades ago, Hollósy (1988) acknowledged the need for such a dictionary after analysing learners' dictionaries, NS dictionaries, thesauri, and dictionaries of collocations. He pointed to the lack of examples from academic publications, and to the unsatisfactory treatment of some common academic words. Furthermore, Hollósy commented on the fact that existing manuals on writing for scientists focused on how to structure an essay or article, and on the writing process (drafting, etc.) rather than on the academic vocabulary required.

While Hollósy's remarks on the situation of academic English in dictionaries were insightful, the same cannot be said for his proposed solutions. He suggested that a dictionary of academic English was needed, and that the dictionary should consist of three parts: an alphabetical bilingualised dictionary (with native language equivalents accompanying definitions), a conceptually organized lexicon or thesaurus, and a monolingual guide to academic writing. There are four problematic issues in Hollósy's proposals:

1) The proposed dictionary would have only 3,000 entries. This means that students would still need to consult other dictionaries for both high frequency words and many technical words.
2) The dictionary would be bilingualised, giving an advantage to larger languages, as smaller languages might be less commercially viable for the publishers.
3) The dictionary of three parts would be too demanding for the user, as three different sets of skills would be required.
4) A sample entry for *information* in Hollósy (1988) demonstrates a significant focus on collocations, however exemplification is predominantly provided by the use of phrases, as there is only one full-sentence example in the entire entry.

More recently, Williams (2006) and De Cock (2006), primarily concerned with NNS students of specialist subjects, have also made calls for a more academically-oriented dictionary. By suggesting that there is a need for a dictionary that would address the needs of non-language specialist students (i.e. whose primary goal is not language learning), Williams (2006) points to a vacant niche in the dictionary market, namely EAP/ESP. Two of the features proposed for this new type of dictionary include using a 'tweaked' form of the Hanks' (1994; 2000) model of definitional prototypes (i.e. that every word has a prototypical meaning, to which all its other meanings are related) and offering more scientific examples.

After finding that existing advanced learners' dictionaries and specialist dictionaries are unsuitable for students due to limited coverage and a lack of examples and information on phraseology, De Cock (2006) suggests the creation of a learner's dictionary that would provide exhaustive treatment of both general and specialised English. De Cock also specifies that the new dictionary should be electronic, highlighting the important role that dictionary format can play in students' decisions about which dictionary to use.

So while De Cock and Williams see the solution in adapting existing advanced learner's dictionaries, Hollósy is the only one, at least to this author's knowledge, who proposes the creation of a completely new dictionary of academic English. It should be pointed out though that all three authors focus on the needs of NNS students. Another common feature of their proposals is the focus on a restricted part of the vocabulary. Therefore, the idea of a comprehensive dictionary of academic English for all students is yet to be explored.

2.2.7 Summary

Students often rely on a dictionary for help with the language they are encountering during their studies. The dictionaries mainly used are general NS dictionaries and learners' dictionaries. The use of dictionaries for university students is limited to the students at US universities, where the selection of dictionaries is imposed on students. But none of these types of dictionaries are very helpful as they reflect general English rather than academic English.

Many lexicographers and researchers have pointed to the shortcomings of existing dictionaries in terms of student needs, but very few have called for a completely new dictionary for students. Instead, most studies have provided suggestions for improving existing dictionaries, or offered advice on which existing dictionaries are more suitable for students.

There is clearly a need for a dictionary that would meet the needs of students, both native and NNSs. Such a dictionary needs to incorporate EAP/ESP research on academic English. Most importantly, the dictionary needs to be based on a corpus of academic language, and not general language. But before the suitability of existing academic corpora is discussed, an overview of dictionary use by students needs to be undertaken.

2.3 Dictionary-use research

Section 2.2.2 examined the variety of dictionaries that students are currently using. But which features are most often consulted? What are the most common errors made by the users? These are only some of the questions that researchers have tried to answer over the years, in an attempt to provide dictionary-makers with valuable information on how dictionaries could be improved.

The majority of the research into dictionary use has focused on the needs of language learners. Consequently, advanced learners' dictionaries have received a great deal of attention from the researchers. This is not surprising, as these dictionaries have introduced the greatest number of innovations (among all types of dictionary) in order to meet the demands of this fast-growing market.

There are only a few studies that focus on the needs of students. These studies give valuable insights into certain aspects of students' dictionary use that are not covered by other studies. Studies with NS students as subjects are especially rare, and thus even more valuable.

Yet, university students are often used as subjects in dictionary research, partly because they are more likely to regularly consult dictionaries than any other user group. The other

reason is probably more practical in nature: since many researchers work at universities, there is no need to go very far from their offices to obtain the data, and the students are much more inclined to participate in a study conducted by someone they know. In any case, this means that a lot of the findings apply directly to the current study.

Hartmann (1987) proposed the following classification of research into dictionary use:

1) research into the information categories presented in dictionaries ('dictionary typology')
2) research into specific dictionary user groups ('user typology')
3) research into the contexts of dictionary use ('needs typology')
4) research into dictionary look-up strategies ('skills typology').

Hartmann's categories were not constructed with students in mind and none of his categories would cover research into the types of dictionaries used. Furthermore, certain aspects of research into students' dictionary use have already been covered in the previous section (e.g. which dictionaries students are using, and how often). Therefore, a slightly adapted version of Hartmann's classification, with several sub-categories, has been created:

1) role of dictionary format
2) research into dictionary features
3) purpose of dictionary use
4) research into dictionary look-up strategies
5) role of language and cultural background.

The aim of this section is twofold: a) to get an understanding of which features students use, and how and for which activities they are using them, and b) to identify any gaps in research.

2.3.1 Role of the dictionary format

In the past decade, the selection of dictionaries on the market has been given another dimension, with the advent of electronic dictionaries. As computers became a virtual necessity, dictionary users had to decide not only which type of dictionary to buy, but also in which format. University students have been particularly attracted by new dictionary formats. There is however little empirical evidence to support this claim, as researchers have shown little interest in studying the role of dictionary format in dictionary selection.

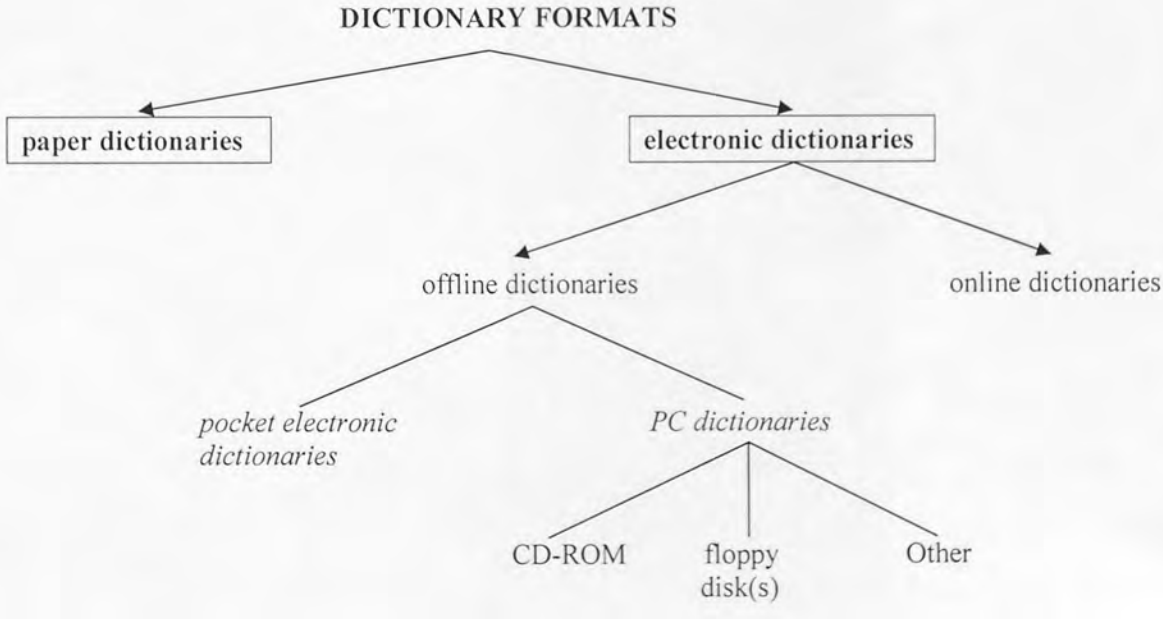
Hartmann's (1999) survey of dictionary use at a UK university is one of the few studies that provides information on the dictionary formats used by students. Nearly 30% of the students reported owning an electronic dictionary (the majority on a PC, and some in the form of a pocket calculator), and nearly two thirds reported that they did not own an electronic dictionary. These findings are now out-of-date as technological progress has not only increased the popularity of electronic dictionary formats, but introduced new formats (e.g. online dictionaries). Nevertheless, the findings do show that even in their early days, electronic dictionaries quickly became popular among students.

The importance of the electronic dictionary format for students has been acknowledged by De Cock (2006) who, in her call for a new dictionary for (NNS) students, proposes that such a dictionary should be electronic. Her argument is that the electronic format has advantages over the paper format due to its 'flexibility' (by this, De Cock probably means customizability, quicker searches, etc.).

In his comprehensive analysis of dictionary formats, de Schryver (2003) provides an extensive list of pros and cons for each dictionary format, supported by comments made by researchers and lexicographers. In addition, de Schryver points out the benefits of the electronic format for users, lexicographers, publishers and researchers.

De Schryver's point of departure is the division of dictionaries into paper and electronic dictionaries, followed by a further sub-grouping of electronic dictionary formats (see Figure 6).

Figure 6. Simplified typology of dictionary formats from de Schryver (2003)



A summary of the advantages of the various formats listed by De Schryver is provided below. First, the advantages that paper dictionaries still have over electronic ones are presented (Table 2). Then, the benefits of using handheld dictionaries (Table 3), the group comprised of paper dictionaries and pocket electronic dictionaries (PEDs), are considered. This is followed by a look at the advantages common to electronic dictionaries (Table 4). Finally, the advantages of dictionaries on CD-ROM and online dictionaries are listed in Table 5 and Table 6 respectively.

Table 2. Advantages of paper dictionaries

- familiar
- books have symbolic value as a physical object
- 'cosy' – easy to browse, can be read recreationally
- eyes get less tired than by looking at computer screen
- specific information can be annotated or underlined
- durable (can be carried around without fear of serious damage or loss of functionality)
- independent (not dependent on a machine)

Table 3. Advantages of handheld dictionaries (paper dictionaries and PEDs)

- usable immediately after purchase
- usable anywhere
- portable (less true of paper dictionaries)
- accessible offline
- PEDs often considered to be mere gadgets or status symbols (but this might increase the motivation to use them!)

Table 4. Advantages common to electronic dictionaries

- novelty value sometimes motivates people to use them
- no paper used (ecologically sound)
- users no longer tied to alphabetical order
- hypertext, menus, etc. eliminate linear text restrictions
- powerful search capabilities (cross-referencing, "sounds like" searches, etc.)

Table 5. Advantages of dictionaries on CD-ROM

- minimal physical storage space required
- audio files can be included: recorded pronunciation, sound files, and record-and-compare feature
- quick data retrieval
- computer graphics: good for presenting data (e.g. the use of coloured network to show synonyms or antonyms, or the degree of synonymity)
- variety of look-up routes
- flexibility (user customization, user-friendly interfaces, etc.)
- copy and paste feature
- linkability with other software

Table 6. Advantages of online dictionaries

- large available collection of data; corpus access sometimes available
- many dictionaries can be consulted simultaneously
- can include video sequences and animation
- linkability with other software
- some available for free

While de Schryver (2003) speaks very favourably of electronic dictionaries, he does point out some of the shortcomings these dictionaries currently exhibit. The main one is that electronic dictionaries seem to be the result of the mere transfer of paper dictionaries into electronic format. While features such as hyperlinks, menus, and user customizability increase user-friendliness and functionality, the content remains more or less the same. This view is shared by Nesi (2000:110) who proposed a potential solution by saying that "electronic dictionaries would be most effective if they were designed from scratch with computer capabilities and computer search mechanisms in mind". This approach has also been advocated by lexicographers such as de Schryver (2003) and Pajsz (2009).

Electronic dictionaries, especially online dictionaries, offer the possibility of unobtrusively monitoring dictionary use by using log files. The findings can then be used to improve dictionaries by enhancing existing features, or adding new ones. Examples of improvements include identifying unsuccessful headword searches and adding new entries, and identifying frequent misspellings in searches and re-routing the user to the correct entries (and potentially alerting the user to the correct spelling) (de Schryver & Joffe, 2004). This method of improving a dictionary, while frequently advocated, is still rarely used by "real world dictionaries" (de Schryver & Joffe, 2004:187), but recent studies (e.g. de Schryver et al., 2006; Keselj & Keselj, 2006) suggest that makers of (online) dictionaries are slowly becoming aware of its potential.

PEDs are a special phenomenon within electronic dictionaries. Although these dictionaries seem to offer the fewest advantages of all electronic formats, they are quite popular, especially among students from Japan, China, Hong Kong, Taiwan and other Asian countries (Nesi, 1999). PEDs are especially valued for their portability and ease of use (Taylor & Chan, 1994), but they are relatively expensive and their content is often of poor quality (Nesi, 1999). However, their provision of access to more than one dictionary, and additional features (such as calculators, calendars, and music players) seem to more than make up for their deficiencies, as far as the users are concerned.

Despite the numerous advantages that electronic dictionaries have to offer, they receive considerably less attention from lexicographers, reviewers and teachers than paper dictionaries. According to Atkins and Rundell's (2008) description of the dictionary-making process, lexicographers are not normally involved in designing electronic dictionary software. Reviewers and teachers are more familiar with paper dictionaries and often prefer using them. However, the fact of the matter is that people, or in our case, students, are using all of the dictionary formats, so ignoring the advantages of electronic dictionary formats means that their full potential may never be realized. Sticking to the same old formula for producing dictionaries – making the paper version first and then thinking about the other possible formats – may be more to the publishers' liking, but it certainly does not benefit dictionary users.

2.3.2 Research into dictionary features

Having examined various dictionary formats, it is now time to turn to the dictionary's macrostructure and microstructure. These two elements can make a dictionary stand out from the rest. They present the essence of the lexicographer's work, and are most scrutinized by the reviewers.

Surveys into dictionary use are a highly important source of feedback for lexicographers and publishers. The findings can reveal a) which features should receive the most attention from dictionary-makers, and b) which features should perhaps be added, omitted, or made less prominent. Most of the research discussed in the following sections refers to paper dictionaries, which highlights the need for more research into the use of electronic dictionaries, and online dictionaries in particular.

2.3.2.1 Which information is most frequently consulted?

Research (Quirk, 1975; Tomaszczyk, 1979; Béjoint, 1981; Jackson, 1988; Battenburg, 1989; Harvey & Yuill, 1997; Hartmann, 1999) shows that meaning (definitions) and spelling are the most frequently consulted microstructural features, both by NS and NNS students. Synonyms are also consulted quite frequently by both groups of students (Béjoint, 1981; Harvey & Yuill, 1997; Hartmann, 1999). NNS students also look up grammatical and collocational information (Béjoint, 1981; Harvey & Yuill, 1997), and examples and idioms[23] (Béjoint, 1981) quite often. Some of these findings are contradicted by Nesi's (2000) study,

[23] It is not clear from Béjoint's article whether his definition of 'idioms' included phrases and fixed expressions.

whose students ignored grammatical and collocational information. On the other hand, features such as encyclopaedic information and pronunciation (Hartmann, 1999), and etymology (Tomaszczyk, 1979; Béjoint, 1981; Hartmann, 1999) are rarely consulted.

In his seminal study, Béjoint (1981) obtained valuable information about NNS students' look-up habits. 66% of his subjects reported never looking up common words, a finding confirmed by the studies of Hatherall (1984) and Bogaards (1998). Similarly, Nesi and Haill (2002) report that their subjects (NNS students) mainly looked up rare words. Another interesting aspect of students' dictionary look-up is provided by Hartmann (1999) whose subjects reported that they found technical terminology, and idioms and phrases most difficult to find, followed by common words used in a specialist area. Such findings are a good reminder that no matter how well individual dictionary features are designed, their quality is always measured by how useful they are to the target user.

2.3.2.2 How useful are the individual features?

In the research mentioned in the previous section, the information was obtained mainly by administering a questionnaire and asking the subjects to reflect on their dictionary use. Another method often used is to focus on a specific dictionary feature and test its usefulness. Many researchers approach this in a theoretical manner, but it is the studies that involve users that provide the most relevant information for this thesis.

Not surprisingly, definitions have always received most attention. MacFarquhar and Richards (1983) examined users' preference of defining style. The subjects were 180 students at English language courses at the University of Hawaii. Three types of definition were offered: two from advanced learners' dictionaries (LDOCE1 and OALD3), with the difference being that LDOCE1 used a restricted defining vocabulary[24]; and one from a NS dictionary (Webster's New World Dictionary). More than half of the students (51.5%) preferred the defining style used in LDOCE1, 28.5% liked OALD3's definitions best, and 20% reported their preference for the definitions in Webster's New World Dictionary.

The problem with MacFarquhar and Richards' method is that their users did not perform any tasks, but simply reported on their overall impression of the look of the entry (Nesi, 2000). This particular deficiency was overcome by Cumming, Cropp and Sussex (1994) whose subjects (85 students with English as a second language) had to produce sentences without copying too much from the definitions or examples. COBUILD and LDOCE definition styles

[24] OALD did not use a restricted defining vocabulary until its fifth edition, published in 1995 (Kirkness, 2006)

were used in this study, with the subjects being given four variations: COBUILD definitions and examples, COBUILD definitions only, LDOCE definitions and examples, and LDOCE definitions only.

The subjects showed a clear preference for COBUILD's full-sentence definitions (with or without examples), while LDOCE definitions with examples were preferred to those without examples.

Nesi (2000) analysed the effectiveness of definitions in three advanced learners' dictionaries (COBUILD1, LDOCE2, and OALD4). 52 subjects were used, all overseas students at Warwick University, which makes the study especially relevant for the EAP context. 18 words from groups 5 and 6 of Nation's University Word List (Xue & Nation, 1984) were selected, and each word was paired with a high frequency word (e.g. *abnormal – shoe*). The subjects had to use the pairings in sentences without copying too much from the definitions or examples. The findings show no significant difference between the three dictionaries, regarding the number of words consulted, or the number of correct sentences produced by the subjects. Nesi does, however, make the comment that the proportion of incorrect sentences was very high, and that OALD4 look-ups produced a significantly higher number of semantic errors.

Other important findings were made in the study, for example common look-up strategies and mistakes in the look-up process made by the students. This is discussed in more detail in section 2.3.4.

Nesi's test was administered on a computer, which was somewhat revolutionary. A specially written computer program recorded and timed instances of definition look-up, and recorded the subjects' own language production (Nesi, 2000:79).

The test administered by Nesi is not without flaws. The pairing of a target word and a high-frequency word is questionable – it would be worth checking in a corpus how frequently the paired words co-occur, as the difficulty of producing a sentence increases if the words do not occur together often in everyday language. In addition, the definitions and/or examples cover a great deal of collocational information about the target words, therefore the subjects had a difficult, and unnatural task in trying to think of a completely different sentence.

Nesi reports on the problem of achieving a high degree of agreement and consistency between the raters, partly stemming from the complex rating system designed for the task which, as Nesi admits, did not always work. Another issue was that the subjects were offered only the entries of the target words and thus prevented from making any further searches. Finally,

COBUILD's extra column was not included, which might have affected the results, as the grammatical information from the other two dictionaries *was* provided.

Black (1986) and Nesi (2000) both examined the role of examples in dictionaries and came to the same conclusion: that the value of examples as an addition to the definition is not that important. Nevertheless, Nesi does mention that a higher percentage of correct sentences were produced by her subjects when consulting the entries with examples. Nesi also criticizes Black's research, saying that it involved a rather substantial amount of guesswork, but her own method, which was in fact similar to the one used in the abovementioned study of definitions (Nesi, 2000), is also not without its shortcomings.

COBUILD's extra column containing grammatical information[25] is a unique feature in lexicography. The authors have separated grammatical information from the rest of the dictionary text in order to give the user the facility to read the main column as a 'normal text', with no abbreviations or codes, with the option of consulting the extra column when the need arises. Studies have shown either that the extra column is rarely consulted (Bogaards & van der Kloot, 2001) or that it is not particularly helpful when consulted (Harvey & Yuill, 1997), which raises questions about its value.

Clearly, some dictionary features are of more interest to lexicographers than to users. However, as Quirk points out, this does not necessarily mean that these features should be omitted, as the users perceive the dictionary as the most comprehensive reference work of all.

2.3.3 Purpose of dictionary use

Hartmann's (1999) study finds that students consult dictionaries mainly when writing or reading. Most frequent activities during which dictionaries are used are working on a written assignment (reported by 91.2% of the students), and reading textbooks (68.3%) and (somewhat less often) academic journals (39.1%). Students rarely use dictionaries during listening, and hardly ever while speaking.

The use of a dictionary for decoding or for encoding are two completely different activities. The user's point of departure is different, as are the specific dictionary features which are consulted. Sections 2.3.3.1 and 2.3.3.2 look into this in more detail, presenting both the processes and the difficulties involved.

---

[25] Extra column also contained lexical information and information on pragmatics, but it is particularly known for its presentation of grammatical information.

2.3.3.1 Dictionary use in decoding

Both reading and listening fall under the label of 'decoding', but reading is regarded as the archetypal decoding activity. In his article, Scholfield (1999) makes a valid point about the use of dictionary during listening:

> "If one is unable to stop the speaker, then the constraints of 'real time' processing make dictionary use difficult. If one is able to stop the speaker, then one would probably ask him or her about the problem word rather than consult a dictionary." (Scholfield, 1999:13)

Nevertheless, the arrival of PEDs has changed this scenario, as many EAP/ESP teachers will undoubtedly confirm.

Scholfield (1999) identifies the following five main steps in dictionary use for decoding:

1) the user identifies an unknown word or phrase
2) the user decides to use a dictionary
3) the user needs to locate an entry
4) the user needs to locate the right part of an entry
5) the user needs to successfully use the information found.

Therefore, the dictionary consultation starts at step 3. But as Scholfield goes on to point out, there are several things that the user has to be able to do just to arrive at the relevant entry:

a) identify the look-up form of the word (e.g. *tried* – the user needs to look up the entry *TRY*)
b) use English alphabetical order effectively (which is especially problematic for users whose native language does not use the Latin alphabet)
c) spell the word (if the user has only heard it)
d) look up the right entry when compounds, phrases, etc. are encountered (this, of course, depends on their treatment in a specific dictionary, which makes this step even more difficult)
e) know what to do if the entry is not found at the location where it is initially sought (maybe it is listed as a sub-entry, maybe the user is redirected by a cross-reference within that initial entry to a different location, or maybe it is not in the dictionary)

This list applies mainly to paper dictionaries, because electronic dictionaries and their advanced search facilities offer the user considerable help in overcoming several of the above mentioned problems. Item c) remains problematic but is limited to listening. However, even this

problem is addressed by some electronic dictionaries, which offer the "sounds like" feature in their searches.

As far as dictionary features are concerned, Scholfield stresses the importance of definitions, sense division and ordering, examples, and signposts (and other such useful guides) for the user. Scholfield also calls for some changes in the treatment of phraseology in some of the existing advanced learners' dictionaries. He thinks that the more frequent phrases should be provided earlier, and not together at the end of an entry:

> "The placement of phrases all in one place, just because they are phrases, though linguistically admirable, only helps the user to find a phrase if he or she is aware that what they are looking up is a phrase, which it may be over-optimistic to assume."
>
> (Scholfield, 1999:27)

When using a dictionary for decoding purposes, the user is confronted with an unknown word or phrase, and the dictionary provides an explanation using vocabulary that the user is, or is at least supposed to be, familiar with. The true decoding value of a dictionary is determined by two factors: 'findability' and 'comprehensibility' (Bogaards, 1996).

2.3.3.2 Dictionary use in encoding

In decoding, the user's search starts with form and ends with meaning. In encoding, the process is reversed: the user knows what he/she wants to say or write, and is looking for the appropriate word/phrase. Using a dictionary for production is more demanding than using it for reception. Similarly, trying to provide encoding information in a dictionary is a more challenging task than providing information for decoding. However, Rundell (1999) reminds us that the distinction between receptive and productive skills is not that clear-cut, pointing out that even when encoding you need to use some decoding skills (e.g. locating the right entry).

Rundell (1999:37) lists the following categories of information required in the majority of production tasks:

1) syntactic behaviour
2) collocational preferences and selection restrictions
3) sociolinguistic features (including register and regional variety)
4) semantic features
5) contextual effects.

Rundell adds that some of the problems are shared with decoding, albeit having a different role in the search process. Also, while electronic dictionaries can make a big difference with

decoding searches, they are not as advantageous in encoding as it is not easy to search for a meaning.

Whereas Rundell agrees that monolingual dictionaries are much more useful for decoding than encoding, he points to several encoding-related improvements in advanced learners' dictionaries. These include comprehensible definitions, common syntactic patterns, lexical sets (e.g. usage notes explaining differences between near-synonyms), help with common errors, and the greater provision of information about phraseology, collocations, illustrations, and frequency.

2.3.4 Research into dictionary look-up strategies

No matter how user-friendly a dictionary claims to be, if you cannot find the information you need in it, the dictionary is of little use to you. Lexicographers need to know as much as possible about their users and their look-up practices. Researchers have managed to identify some common strategies adopted by users, as well as some of the common problems that the users encounter.

2.3.4.1 'Choose the first definition' strategy[26]

This very common strategy is the cause of many unsuccessful searches, and it is particularly important because most dictionary users have not been given any training in dictionary use. Mitchell (1983), Tono (1984), Neubach and Cohen (1988), McCreary (2002), and Nesi and Haill (2002) all report that their subjects encountered problems during searches on account of using this strategy.

One way of addressing this issue, as well as several other problems faced by dictionary users, is by training the users in how to use a dictionary efficiently. However, such calls have been made in the past and the situation has not changed over the years. Hence, Rundell (1999) makes a valid point when he says that lexicographers should not count on someone to teach the users how to use dictionaries efficiently, but should rather try to produce dictionaries with straightforward structure and content that do not require a great deal of expertise on the user's part.

If users do often look only at the first sense, dictionary-makers are right to order senses by frequency. Thus, the most common sense, which is also the most likely to be encountered by

[26] The term used by McCreary (2002).

the user, is offered first. On the other hand, it could be argued that the user is more likely to know the most common sense(s), and it is precisely the less frequent ones that they will be looking up. However, as there is no way of knowing which sense will be most likely to be looked up, ordering senses by frequency remains the best option.

The implications of this practice are different for students. Word senses in general-purpose dictionaries may be ordered by frequency, however the frequency information is based on general language corpora. Word senses with a higher frequency in academic language will therefore be offered later in the entries. In view of the fact that the 'choose the first definition' strategy is supposedly used very often, students need a dictionary that focuses on words and senses in academic language.

2.3.4.2 'Kidrule' strategy

When using this strategy, "a short familiar segment of the dictionary definition is taken out of context as an equivalent for the unknown headword" (Nesi & Haill, 2002:285). A good example is found in Nesi's (2000) study 3, in which one of the subjects seemingly selected the word *different* from the LDOCE2 entry for *version* (1. a slightly different form, copy or style of an article) and produced the following sentence: *I will begin new job that is version.*

The 'kidrule strategy' was first mentioned by Miller and Gildea (1987) who carried out research involving 10- and 11-year-old children. Later research has shown that the strategy is also used by students and adults (Harvey & Yuill, 1997; McCreary & Dolezal, 1999; Nesi, 2000; McCreary, 2002; Nesi & Haill, 2002). However, the language proficiency of the subjects in these studies was still quite low. The best way to prevent users from employing this strategy is by producing clear and comprehensible definitions.

An extended version of the kidrule strategy, called 'the superficial cognate rule', was observed by McCreary (2002, 2008) in research on NS students using dictionaries while completing a production task with 'hard words'. The subjects

> "think of a more familiar word (which may not necessarily be in the entry) that at first glance appears to be similar to the test word (because of similarity in the spelling, similarity in the suffix, or assumed similarity in the sound) and transfer that more familiar sense to the test word; in that supposed sense, the test word is then inserted in the target sentence." (McCreary, 2002:16)

As there is no other evidence to support this, it is possible that this strategy may be limited to the specific group of students or the type of production task used by McCreary.

2.3.4.3 Problems with locating the right entry or sense

In section 2.3.3.1, Scholfield's (1999) list of the skills the user needs just to get to the appropriate entry was mentioned. The user's language background can play a crucial role, especially if the user is not familiar with the Latin alphabet. For example, Bensoussan et al. (1984) suspected that some of their subjects had problems in locating the entry because of the difference between the English alphabet and the Hebrew and Arabic alphabets.

Whereas the problem with the alphabet can be solved by learning the English alphabet, there is no such simple solution for the occasions when the user simply gives up the search. This can be caused by many factors: the user may be looking at the wrong place, or may be discouraged by the length of an entry, the dictionary may not contain the particular entry sought, and so forth.

The problem of facing numerous senses at entries for polysemous words can not only make the user give up the search, but may also lead to the user selecting the incorrect sense, or employing the 'choose the first definition' strategy. The introduction of features like signposts and menus has made a significant contribution to addressing this issue.

2.3.4.4 Identifying the wrong grammatical class of a word

At some entries, the user is required to make an additional decision, namely to select the grammatical class of the word before proceeding to the selection of the appropriate sense. Nesi and Haill (2002) report that this was the most common source of the errors made by their subjects. In most cases, the subjects confused nouns, verbs and adjectives.

Although Nesi and Haill's study is one of the rare ones to mention this problem, the fact that it was conducted at an English-speaking (British) university with (NNS) students makes it highly relevant for this research.

2.3.4.5 Incorrect use of information found in the dictionary

Once the users successfully navigate through the dictionary to find the right entry and the right sense, they need to correctly use the information offered. This applies to both decoding and encoding, with encoding being the more demanding of the two activities, as it involves not only the understanding of the information, but also its application. Dictionaries contain many helpful features that try to offer guidance to the user on how to use the word correctly. These include collocations, examples, syntactical information, and frequent phrases.

2.3.5 Role of language and cultural background

The learner's language background plays an important role in the learning of English. The greater the difference between the learner's native language and English, the more severe the problems that can be anticipated. Of course, this does not necessarily mean that a student from China will be less proficient in English than a student from Germany, but it is likely that a student from China will encounter more difficulties in the learning process, and will thus have to learn English longer to reach the same level as a student from Germany.

But even some NSs are affected by their language and cultural background. This refers to NSs who were born in an English-speaking country, or have lived in an English-speaking country since their early childhood, but are of a different ethnic origin. These NSs are normally bilingual as they are encouraged to learn the language and customs of their parents' culture as well. The number of students that belong to this group of NSs is not negligible – for example, in the academic year 2007-08, 17% of entrants into UK universities (first-year undergraduates) came from minority ethnic groups (Universities UK, 2009), and 19.1% of entrants into universities in the USA were Hispanic, Asian/Pacific Islanders, or American Indian/Alaskan natives (U.S. Department of Education, 2008).

Researchers have shown that language background also plays an important role in successful dictionary use. Ard's (1982) study focused on the use of bilingual dictionaries by subjects from two different language backgrounds (Japanese and Spanish), and the conclusion reached was that students with a native language 'closer' to English (i.e. Spanish) were more likely to make successful (bilingual) dictionary searches.

Meara and English (1987) examined lexical errors in Cambridge First Certificate examination papers. The subjects used the Longman Active Study Dictionary during their examination, and it was established that the dictionary was far more effective with students with certain language backgrounds (e.g. Greek, Finnish) than others (e.g. Chinese, Swahili).

In Nesi's (2000) research, the subjects came from only two language backgrounds, Portuguese and Malay. The subjects used LDOCE2 in a production task. The students with a Malay background performed less well in all respects (they looked up more words, their lookups took longer, and they obtained lower marks for the sentences produced), despite having displaying a larger vocabulary size. Nesi's conclusion was that language background does indeed affect dictionary use and the success of lookups. While Nesi attributes the better results of the Portuguese-speaking subjects to the closer relation between their native language and

English, she admits that the Portuguese students probably had more experience in using dictionaries as well.

On the other hand, Battenburg (1989; 1991) did not find language background to have any influence on dictionary use. The majority of the subjects in his study had Chinese or Arabic as their native language. As both languages are very different from English, this might have contributed to Battenburg's divergent findings. Also, Nesi's (2000) criticism of Battenburg's research refers to his use of questionnaires, asking the subjects to report on their dictionary use instead of observing them during the actual process.

2.3.6 Dictionary-use research – summary and implications

Students use dictionaries mainly during writing and reading activities, which by nature allow more time for dictionary consultation. It is noteworthy that writing and reading differ in their purpose; the former is a productive activity while the latter is a receptive activity. And dictionary skills used during productive activities and receptive activities are quite different, partly because different dictionary features are consulted.

The purpose of dictionary use dictates the relevance of different dictionary features, so it is normal that students consult some features more frequently than others. Meaning and spelling are looked up most frequently, by all students, whereas other features (e.g. collocation and examples) are looked up more frequently by NNSs than by NSs.

When using dictionaries, students encounter various problems which hinder or even stop their searches. Many problems are caused by the students' lack of dictionary skills and/or low language proficiency. But the problems of the users during dictionary use can also be interpreted as a sign that the dictionaries have certain shortcomings.

One improvement that dictionary-makers can provide to their users is a greater choice of dictionary formats. Electronic formats (especially the online format) offer many advantages, not only to the users, but also to dictionary-makers who can improve dictionaries quickly and according to the needs of (individual) users. However, lexicography has yet to take full advantage of the potential of electronic dictionary formats.

This section has pointed to two gaps in research on students' dictionary use. Firstly, there is no up-to-date information on the dictionary formats that students prefer. Secondly, there is a lack of research that compares the dictionary use of NS students and NNS students. This information would be of crucial importance when designing DOAE.

2.4 Academic English and corpora

DOAE should be based on a corpus of academic English. Corpus linguistics has had a great impact on the work of EAP researchers and practitioners, and many corpora of academic English have been created and used for both research and teaching. The existing corpus resources for academic English are analysed here, in an attempt to determine whether any of them can be used as a basis for the design of DOAE.

2.4.1 Academic English in general corpora

Although corpora of academic English as such did not emerge until the 1980s, academic discourse has always formed some part of all general corpora. The 1-million-word Brown corpus (Francis & Kucera, 1979) of the 1960s, for example, has a category specifically labelled 'Learned', consisting of 80 written texts (approx. 160,000 words). The BNC, which is 100-times larger, has a very similar percentage (15.8%) of academic texts, including some transcripts of lectures. The 2000 version of the Bank of English corpus included 6 million words of American academic textbooks and many texts in the 44-million-word British books subcorpus (Krishnamurthy, 2001). The latest (and the largest) general corpus of English, the Oxford English Corpus (OEC), contains over two billion words and although the website (http://www.askoxford.com/oec) mentions that academic papers and journals are included, no specific figures are offered.

Furthermore, as Thompson (2006) notes, academic texts in general corpora are often incomplete because of the use of text extracts rather than complete texts in corpus compilation. In addition, there are issues with representativeness, as certain disciplines are sometimes better represented, which is probably the result of the availability of data rather than corpus design principles. Of the general corpora mentioned above, the OEC might be of interest for DOAE, as it may contain a great number of academic books and journals that the Oxford University Press has access to. Unfortunately, the corpus is not available to the public.

General corpora can therefore not be used, at least not directly, for the purposes of DOAE. Corpora of academic English might be a better option, and some of them are examined more closely in the next section.

2.4.2 Corpora of academic English

Attempts to classify corpora of academic English (e.g. Flowerdew, 2002b; Thompson, 2006) have usually grouped them into English for General Academic Purposes (EGAP) and English for Specific Academic Purposes (ESAP) corpora. Normally, EGAP corpora contain texts from several (significantly) different disciplines, and ESAP corpora contain texts from only one discipline, or two or more similar disciplines. However, there is a lack of agreement regarding the classification of specific corpora. For example, Flowerdew (2002b) lists the Jiaotong Daxue English for Science and Technology (JDEST) corpus under ESAP corpora, while Thompson (2006) puts it in his EGAP group; Flowerdew (2002b) discusses learner corpora under EGAP, whereas Thompson (2006) makes them into a separate group altogether.

Although studies are not always in agreement on the classification of corpora of academic English, they do provide an up-to-date inventory of existing corpora. For the purposes of DOAE, information on corpus contents (e.g. discourses, genres, and disciplines covered; expert and/or learner authors) is of particular importance.

The list of corpora of academic English in Table 98 in Appendix 1 does not include many ESAP corpora that were compiled by individuals for their own research, as the interest of this thesis is in the accessibility of the corpora, and not merely in their existence (for more exhaustive lists of ESAP corpora, see Aston 1997 and Flowerdew 2002).

In comparison to general corpora, corpora of academic English are much smaller in size. A major problem with academic texts is that publishers and authors are reluctant to grant copyright permission to corpus compilers who wish to make their corpora more widely available, because academic books and journal articles can remain relevant for a long time. The Professional English Research Consortium (PERC) corpus is one of few successful attempts to make a large collection of published academic texts available to the public[27]. Another issue is funding: while large general corpus projects are usually funded for several years, corpora of academic English are compiled by academics who can only afford a limited investment of time.

T2K-SWAL is the only corpus on the list that could be considered a reference corpus of academic English, because it contains both written texts and speech events. The corpus is reasonably well-balanced in terms of spoken and written discourse, as well as in terms of genres and disciplines. Nevertheless, T2K-SWAL does not contain any academic journal articles which are a major element in university study.

---

[27] The users of the corpus must sign the end-user agreement, and the use of texts in the corpus is still subject to copyright law (i.e. only segments of limited length can be extracted).

The number of corpora of academic English focusing on student writing is on the increase. The latest projects in collecting academic texts written by students are the BAWE corpus (Nesi et al., 2005), MICUSP, and the Reading Academic Text (RAT) corpus. BAWE and MICUSP have collected student texts (awarded high grades) at both undergraduate and postgraduate level, and RAT has focused on PhD theses written by staff or students.

Corpora of student writing can make an important contribution to DOAE. They can be used with a similar purpose to learner corpora in EFL, namely to identify differences between student writing and expert writing, and highlight common errors made by students. Learner corpora such as the International Corpus of Learner English (ICLE) can also be of some use, but it should be stressed that ICLE contains mainly argumentative essays, which are not typical of academic writing – the topics are broad and/or not particularly academic (e.g. "Crime does not pay", "Europe"), and show an absence of common academic conventions such as referencing.

Flowerdew (2002b:110) complained about the lack of spoken corpora for academic purposes. This is now no longer the case. The T2K-SWAL corpus, the Hong Kong Corpus of Spoken English (HKCSE), MICASE, the BASE corpus, and the Corpus of English as Lingua Franca in Academic Settings (ELFA) (Mauranen, 2003) have all been compiled in the intervening years. However, T2K-SWAL is not publicly available. HKCSE and ELFA include NNS speech, and have problems of representativeness. MICASE and BASE, built to a common design, can be accessed online, and complement each other in terms of representing American and British academic speech respectively.

The variety of corpus data of academic English is wide indeed. Written data ranges from articles, essays, theses, monographs and textbooks, to course packs and laboratory manuals. Spoken data includes a number of different speech events, from lectures and seminars to tutorials and student presentations. The authors of texts can be lecturers or students, native or non-native speakers of English. Hence, there are substantial amounts of academic corpus data in existence. The question is whether they can be used in the design of DOAE. This is discussed next.

2.4.2.1 Problems of existing corpora of academic English

The majority of existing corpora of academic English have one of three major problems that make them unsuitable for DOAE. These problems are contents, size, and availability.

a) Contents

A corpus for the purposes of DOAE would need to contain target texts, namely published academic writing (articles, monographs, etc.) and lectures, seminars and other speech events from a wide variety of disciplines. However, the corpora of written academic English found in Table 98 that contain published academic texts often focus only on particular disciplines.

Several corpora of academic English (e.g. BAWE, RAT, MICUSP) consist of student writing. BAWE and MICUSP display lexicographic potential as they include student texts that were awarded high grades. Nevertheless, these texts are probably half way between the complete beginner stage of academic writing and the target texts (published material). Somewhat closer to the target texts are the contents of the RAT corpus (PhD theses), but the corpus lacks sufficient coverage of disciplines.

More potential is exhibited by corpora of spoken academic English, specifically BASE and MICASE. Both corpora are publicly available and contain a variety of speech events in different disciplines. Furthermore, the compilers of both corpora have used the same classification system. These two corpora could provide information about the characteristics of spoken academic English and the differences between spoken American and spoken British academic English.

b) Size

It has been mentioned that corpora of academic English are much smaller than general corpora. This is especially true of corpora of written academic English that contain target texts. In Table 98, Coxhead's (2000) Academic Corpus is the largest existing corpus with coverage across many different disciplines, containing around 3.5 million words. Coxhead uses a classification with four top-level categories (Arts, Commerce, Law, and Science), each divided into 7 subject areas (making 28 subject areas in total). Therefore, if the corpus data is equally distributed, each subject area contains 125,000 words. Such an amount of data would not meet lexicographic needs, as any claim that DOAE based on this corpus made would carry little weight.

The PERC corpus (17 million words) is in fact the largest corpus in Table 98 of target academic texts, containing approximately 770,000 words per discipline, which would be a more acceptable size for a lexicographic project. However, the corpus has problems in terms of contents as the texts are taken only from science and technology disciplines.

ESAP corpora prove that even EAP researchers prefer to base their findings on larger amounts of data. For example, Gledhill's (1996; 2000) corpus consisted of 500,000 words, despite focusing on a single topic within discipline, or to be more precise, on a specific genre within the topic (cancer research articles).

Corpora of spoken academic English would probably meet the demands of DOAE. A corpus of spoken academic English would be especially valuable for determining the differences between spoken academic English and written academic English.

c) Availability

Availability is without a doubt one of the main problems related to many corpora of academic English, especially those containing published material. The problem is that EAP researchers compile their corpora for their own research and do not obtain copyright permission. Hence, a great deal of effort put into the compilation of the corpora has benefits only for the compiler(s).

Existing corpora of spoken academic English have fewer issues concerning copyright and are easier to obtain. For example, BASE and MICASE are available both online and on a CD-ROM.

But even when a corpus of academic English is available, the access is often offered only via specific software which has not been designed with lexicographers in mind. As a result, the software lacks many features required for comprehensive lexicographic analysis.

2.4.3 Overview of academic corpora – summary and implications

This analysis has shown that the suitability of existing corpora for lexicographic purposes is limited. The main shortcomings of corpora of academic English are the lack of appropriate contents (the target texts), their small size, and their unavailability. All this suggests that a completely new corpus of academic English will need to be compiled for the purposes of DOAE. Also, as corpora of spoken academic English exhibit far fewer of the general shortcomings of academic corpora, they could form a very useful part of the new corpus of academic English.

2.5 Conclusions

It has been demonstrated that, despite the plethora of dictionaries available, students lack a single dictionary that would address their needs. They are forced to use at least two

dictionaries, if not more. Research shows that students often resort to using general-purpose dictionaries, although these dictionaries are not designed for them. In addition, the practice of consulting only the first definition seems to be quite widespread, which can be a major problem for students, considering the fact that existing dictionaries are based on general language corpora and therefore the first definition is likely to be for general usage, not academic. The purpose of this research is therefore to answer calls for an academically-oriented dictionary (made by Hollósy 1988, Williams 2006, and De Cock 2006), and to produce a model for DOAE.

3. METHODOLOGY

The Model for DOAE, which will be designed in this thesis, aims to provide lexicographers with a proposal on how to compile a dictionary that would cater for the needs of university students. It is important to design a model first, rather than directly designing the dictionary itself, because a model can address each stage of dictionary production in detail without being limited by some of the constraints of dictionary-making, such as time and cost.

A user profile is a prerequisite for any dictionary, and the user profile for this dictionary Model was expected to utilize past research into dictionary use. Past research into dictionary use has indeed provided some useful insights into the dictionary use of students. There are, however, some important gaps in the research (e.g. format preference, differences/similarities between NS and NNS student dictionary use) that prevented us from developing an exhaustive user profile for the dictionary. Thus, a specially-designed survey of dictionary use will be conducted as part of the current research to fill these gaps.

In order to be able to describe the data analysis and demonstrate how the dictionary entries could be compiled, the Model needs to create a number of sample entries. This requires access to a suitable a corpus of academic English, but none of the available corpora is adequate for the compilation of the proposed dictionary. Hence, a corpus of academic English will be compiled specially for the purposes of this Model. The corpus will contain written texts only, as the existing corpora of spoken academic English spoken data are adequate (see 2.4.2).

The Model will have credibility only if it uses the software and tools that are used by current lexicographers. State-of-the-art lexicographic tools will thus be used throughout the design of the Model – from data analysis, creation of the dictionary database, and input of information into the database, to the compilation of dictionary entries, and the presentation of dictionary entries.

This chapter presents the data and tools used in the design of the Model for DOAE. The structure of the chapter follows the order in which a dictionary is created; the first section focuses on the method used to obtain more information about potential users, the second section presents the corpus used for analysis and ancillary resources used for consultation purposes, the third section presents the tools used for analysing the corpus and compiling a dictionary, and the last section discusses the approach used to analyse the corpus data.

3.1 Questionnaire – creating a user profile

A questionnaire was used to create the profile of the potential users of DOAE. The questionnaire aimed to identify the needs and preferences of university students. The questions focused on obtaining the following information:

- *Which monolingual English dictionaries are used by students?* The intention was to analyse the most frequently used dictionaries to identify any useful features that could be used in DOAE.

- *Which dictionary format do students prefer?* It was very important to obtain some information on the dictionary format preferred by the students, as there is currently very little research on this topic. This information will also play an important role in choosing the main format of DOAE.

- *When and for what purposes are dictionaries used?* This information was needed to establish the importance that students attribute to decoding and encoding activities when consulting a dictionary.

- *What microstructural information is consulted and how often?* This information would help to determine which features should be included in the dictionary, and which features should receive more attention than others (in terms of space in the dictionary entry).

In addition, the results were analysed to obtain the following information:

- *Are there any differences in the dictionary use of different groups of students, e.g. between NSs and NNSs?* These findings would be used when designing customizable features in the dictionary.

The questionnaire was designed by consulting questionnaires used in previous dictionary-use research (e.g. Hartmann, 1999), and using Lew's (2002) checklist of dos and don'ts for questionnaire design (see Figure 7). The implementation involved piloting the questionnaire, making revisions, piloting the revised version, making revisions to the revised version, and administering the final version of the questionnaire.

Figure 7. Checklist for questionnaire design (Source: Lew, 2002).

Do:	Don't:
• write your questionnaire in the subjects' native language • pay attention to clean, unambiguous graphical layout • consider in each case whether multiple choice or open-ended or mixed question format is most appropriate • decide before the design is complete how the results will be coded and processed • screen your questions and multiple-choice answers for possible bias • ask a colleague or two to read through a draft of your questionnaire • pilot your questionnaire • allow appropriate time for your questionnaire to be completed	• use technical language that subjects might not understand • use complex syntax • use negatives in questions • let page breaks split questions • put nonessential questions in the questionnaire just because others had them • give away your own position or preference in any way

3.1.1 Pilot survey

The questionnaire used for the pilot survey (see Appendix 2) administered in summer 2006 had ten items, and was designed to obtain general information about the subjects (pre-sessional students), their history of learning English, their dictionary use, information about the types and format of dictionary they used, and what kind of information they searched for in dictionaries. The questions were mainly of the multiple-choice type, some requiring additional explanation of the answer, and only the last question was completely open. The purpose of the pilot survey was to identify any shortcomings or problems of the questionnaire itself, and at the same time to obtain information for training sessions on dictionary use, which were later conducted with the subjects.

The subjects were 111 international students attending pre-sessional courses at Aston University. 55 students attended a 10-week pre-sessional course (Group 1) and 56 a 5-week pre-sessional course (Group 2) in summer 2006. They came from the following language backgrounds: Chinese (64), Thai (20), Japanese (6), Korean (5), Mandarin (4), Italian (2), Greek (2), Arabic (2), Spanish (1), Russian (1), Kazakh (1) and Bangla (1)[28].

A vast majority of the students (97%) were about to start postgraduate studies and most of them had been conditionally accepted onto Master's courses at Aston Business School, with

[28] Two students did not provide information about their native language.

MSc in Marketing Management (48 students), MSc in Human Resource Management and Business (12), and MBA (12) being the most popular courses.

A gap of several weeks between the start of the 10-week and the 5-week pre-sessional course provided an opportunity to test two methods of administering the questionnaire. Group 1 was given the questionnaire during a lesson and allowed 10 minutes to complete it. The analysis revealed that the students had difficulties in remembering the titles and publishers of the dictionaries they owned. Thus, the students in Group 2 were allowed to complete the questionnaire at home. There was, however, no significant improvement in the amount of information provided (in comparison with Group 1).

In both cases, the questionnaire raised problems which prompted the decision to change the format from paper to online. Firstly, on seeing a four-page questionnaire, many students seemed to lose interest. Secondly, the analysis first required manual input of the data into a computer program, in this case Microsoft Excel, which was a time-consuming process. For this reason, it was decided to use an online questionnaire for the main survey. This enabled distribution to a larger group of students, quicker completion of the questionnaire, and quicker processing of the data.

Another problematic feature of the pilot questionnaire was the amount of information it tried to obtain. This is especially the case at Question 7 (see Appendix 2, page 374). For example, there is no need to know how frequently each dictionary format is used; partly because the student might have only a single format, and partly because this requires a great deal of reflection by the student. Thus, the preference of format and the frequency of use of individual formats would be addressed in separate questions (see Question 2 and Question 3 in Appendix 2).

Another aspect that required improvement was the use of clearer and more appropriate terminology. Expressions I had used such as "general monolingual English dictionary", "learner's monolingual English dictionary" and "dictionary version" were found to be rarely used in the lexicographic literature, and needed reformulating. This problem was solved to a large extent by the decision to abandon pre-selecting types of dictionary for the students, and rather to allow the students to provide as much information as they can about the dictionaries they use.

It was also decided that the questionnaire would focus solely on monolingual English dictionaries. Whereas the information about bilingual dictionaries and thesauruses was useful for dictionary training sessions, it did not have any great value for creating a user profile for

DOAE. This solution therefore followed Lew's (2002) advice about not putting nonessential questions in the questionnaire.

As the results for Question 8 (What information do you usually look in a dictionary for?) were not conclusive, some of the approaches taken by other researchers were examined. Two options were then considered: one was to ask the students to name the three most often (and three least often) consulted features on a list provided. However, the results would be difficult to analyse and would not provide any information about the use of the remaining two features. Hence, this option was not used. The second option included changing the selection of answers. "Always" and "never" seem to be too extreme and "sometimes" covers a wide scope of answers – from "not always" to "extremely rarely" (but not never). In the end, a decision was made to change the answers, but to keep four different options (*almost always – often – rarely – almost never*).

The final changes made to the questionnaire included changing some of the existing questions about the students, and adding new ones. The students in the pilot survey were a very homogenous group, and a lot of their background information was already provided by the director of pre-sessional courses at Aston University. For the main survey it was expected, and intended, that the students would be much more heterogeneous in terms of native language, subject of study, and level of study.

3.1.2 Main survey

The survey was piloted in October 2007 among six members of staff (three lecturers, two PhD students, and an Academic Support Officer) at the Aston School of Languages and Social Sciences. The subjects were asked to complete the survey, and then comment on the questions, the layout, and the time needed to complete the survey. Their comments were then considered when making the final version of the survey.

The main survey (see Appendix 3) was administered online in the period between November 2007 and March 2008. The email invitation to participate in the survey was sent to anyone with an Aston University email account, so it included distance students as well as on-campus students. The survey was completed by 620 students from all four Aston Schools[29] –

[29] The total number of responses was 901 – 136 were immediately excluded because the questionnaire was not fully completed, and another 145 were not included in the main analysis as they were completed by members of staff.

449 NSs and 171 NNSs. The students ranged from first-year undergraduates to PhD researchers, and the proportion of NS and NNS respondents was fairly similar for each of these levels.

Both the pilot survey and the main survey were administered using the Bristol Online Surveys (BOS) tool[30]. BOS enables the user to design, administer, and even analyse a survey via the internet. BOS offers a list of templates which are based on surveys already conducted by organisations such as the Institute for Learning and Research Technology at the University of Bristol. The user can use the design tool to create any of the following six types of questions: selection list (drop-down menu), multiple-choice, multiple answer, single line (open-ended, only one line of text allowed), multiple lines (open-ended, multiple lines of text allowed), and date (e.g. for date of birth). These six types of questions can also be combined into GRID format questions (e.g. several multiple-choice questions with the same optional answers – see Question 1 in Appendix 3, page 378, for an example). Also, follow-on questions can be used.

BOS provides basic statistics of the results, such as the distribution of answers in numbers and percentages, median rank, mean rank, variance, and standard deviation. These statistics are available for selection lists, multiple-choice questions and GRID questions. The analysis of open-ended questions is left to the user, although a quick overview of the frequency of words or numbers in the answers can be obtained by using the Word Cloud feature (Figure 8).

Figure 8. Main survey: Word Cloud for question 16.

Other tools, such as Microsoft Excel (2002) and the Statistical Package for Social Sciences (version 12, SPSS Inc., Chicago, USA) software also used for the analysis of the results. Microsoft Excel was used to analyse open-ended questions because students' answers often required manual corrections due to minor errors, such as typos and different ways of providing information. For example, some students reported that the name of their dictionary was 'Oxford Advanced Learner's Dictionary', while others wrote 'Advanced Learner's Dictionary', and typed 'Oxford' in the 'Publisher' row. Both students obviously had the same dictionary in mind, so the latter answer was modified to 'Oxford Advanced Learner's Dictionary'. Such errors and inconsistencies were corrected to improve the accuracy of the

[30] All the information about the Bristol Online Survey was obtained from http://www.survey.bris.ac.uk/.

statistics. In addition, Microsoft Excel was used to store a copy of all the results of the survey, and to export data to SPSS. SPSS was used to compare NS students with NNS students using the Mann-Whitney U test[31], a non-parametric test of significance.

3.2 Data for the Model for DOAE

The language data used for the purposes of this research can be divided into primary data and secondary data. There is only one source of primary data, the CAJA corpus, which was compiled specifically for this research due to lack of (lexicographically) suitable corpora of academic English (see 2.4.2.1), and was used as the basis for designing the Model for DOAE.

Secondary data such as other corpora, and dictionaries, were used for identifying any differences between meanings/uses of the words in academic English and in general English, and any differences between meanings/uses of the word in expert academic writing (found in CAJA) and in other academic discourses (e.g. spoken academic English). In addition, dictionaries have been consulted as a possible source of features that could be useful for DOAE.

3.2.1 Primary data

3.2.1.1 Corpus of Academic Journal Articles (CAJA)

CAJA was compiled between November 2007 and September 2008 and contains 83,554,346 words of written academic discourse. The corpus consists of 13,116 texts from 28 different disciplines. As the corpus is to be used as the main source for designing the Model for DOAE, it is important to explain in detail its design and contents, and the decisions taken.

3.2.1.1.1 Mode

CAJA contains written texts only; the texts could even be categorized as electronic written texts as they were downloaded from online journals. However, every effort was made to remove any information from the text that was part of the electronic version, but would not be in a paper version (e.g. menus, hyperlinks).

[31] "This is a frequently used test to determine whether two independent groups belong to the same population. It is considered to be one of the most powerful non-parametric tests available and it is appropriate for the data that is at least ordinal." (Lewis-Beck et al., 2004:606)

It was decided not to collect any spoken data, partly because of the availability of existing spoken corpora of academic English (e.g. the BASE corpus), and partly because compiling a spoken corpus is very time-consuming, as it consists of recording and transcribing data. Furthermore, the average user of the proposed dictionary is expected to use the dictionary when reading or writing academic texts, and not when speaking. Nonetheless, the difference(s) between the use of a word or phrase in written and spoken academic discourse may be worth pointing out in some cases, and this will be exemplified by comparing CAJA with the BASE and MICASE corpora.

3.2.1.1.2 Varieties of English

All the texts are in English and represent written academic English, i.e. English used in written academic discourse. Academic English is an international phenomenon, and we can expect to find many different varieties of English in the corpus texts. Here are some of the factors that can affect the variety of English in an academic text:

a) Native language of the author(s). If the authors are from an English-speaking country, their writing will contain features of their variety of English. If the authors are NNSs of English, their writing will show the influence of the variety of English they were taught.
b) Location of the place of work/study of the author(s). For example, if the authors work at a university in the US, they are more likely to be influenced by American English.
c) Location of the journal publisher or the journal. For example, if the publisher or journal is based in the UK, the authors probably need to follow spelling conventions used in British English. Nonetheless, it should be borne in mind that reviewers of submitted articles are likely to be based all over the world, so they may be indifferent to the variety of English used in the texts.

However, it would clearly be impossible to compile all the necessary information about the individual authors of the texts; the location of the workplace is an increasingly problematic factor, as academic staff become more mobile; so the location of the journal publisher seems the most reliable indicator of variety. Judging purely by the location of the publishers, it seems clear that the majority of the texts are likely to be written in British English or American English. The contributors to these journals are likely to be academic staff at leading universities, the vast majority of which, irrespective of the academic discipline, are based in the UK and the US (Graddol, 2006).

This problem of determining the variety of English in academic texts could present difficulties if the aim was to compile a corpus containing an equal amount of each variety of English. This is less problematic for lexicographers working on a global dictionary who can focus on the meanings of the words first, and only then look at the characteristics of the texts the meanings come from.

3.2.1.1.3 Domain categories

Domain categories are a crucial part of the corpus design as they are expected to present a basis for domain labels in the dictionary. However, there is no generally-accepted list of academic disciplines. Moreover, as Krishnamurthy and Kosem (2007:363) point out, "there is a lack of consensus among academic institutions and librarians, as well as within the corpus community [about the classification of academic subjects]". As is evident from the table Krishnamurthy and Kosem provide (see Table 7), institutions tend to use more categories than corpus developers.

Dictionaries, especially for NSs, use a large number of domain labels, and understandably so. Domain labels are used to signal terminology, and must be quite specific. Broad domain labels such as Biological and Health Sciences (a category used by corpus compilers of BASE, MICASE and BAWE) would not be sufficiently accurate for the dictionary user. Learners' dictionaries use far fewer domain labels, but they are still very narrow. The reason for using fewer labels lies in the fact that these dictionaries contain only terminology that is often found in the general language, for example business and medical terminology.

These findings led to the conclusion that a completely new list of academic subjects would need to be created for the corpus. The question that was used as the point of departure was: what classification, with categories narrow enough to be used as domain labels in an academic dictionary, would enable students to identify their subject of study without any problems? To provide the best answer to this question, the following steps were taken:

a) visiting the websites of 10 UK universities[32], ranging from small to very large in terms of student population, and making a list of subjects offered. While smaller universities do not offer as many subjects as the larger ones, they tend to specialize in certain fields and were expected to use a more detailed classification.

[32] The universities were: Aston University (9,555 students), Lancaster University (17,410), the Open University (176,560), the University of Birmingham (30,415), the University of Cambridge (28,775), the University of Durham (17,410), the University of London (c. 125,000), the University of Manchester (39,165), the University of Oxford (24,640), the University of Worcester (7,750). Source: (HESA, 2008).

Table 7. Major categories of academic subjects adopted by some academic institutions, librarians, and corpus compilers.

Joint Academic Classification of Subjects (Higher Education Statistics Agency)	Dewey Decimal Classification System	Brown corpus (Francis & Kucera, 1979)	The Academic Corpus Coxhead (2000)	MICASE, BASE, and BAWE (Nesi et al., 2005; MICASE Manual, 2003)
1. Medicine & Dentistry 2. Subjects allied to Medicine 3. Biological Sciences 4. Veterinary Sciences, Agriculture and related subjects 5. Physical Sciences 6. Mathematical and Computer Sciences 7. Engineering 8. Technologies 9. Architecture, Building and Planning 10. Social Studies 11. Law 12. Business and Administrative studies 13. Mass Communication and Documentation 14. Linguistics, Classics, and related subjects 15. European Languages, Literature and related subjects 16. Eastern, Asiatic, African, American and Australasian Languages, Literature and related subjects 17. Historical and Philosophical studies 18. Creative Arts and Design 19. Education	1. Generalities 2. Philosophy & Psychology 3. Religion 4. Social Sciences 5. Language 6. Natural Sciences & Mathematics 7. Technology (Applied Sciences) 8. The Arts 9. Literature & Rhetoric 10. Geography & History	1. Natural Sciences 2. Medicine 3. Mathematics 4. Social and Behavioral Sciences 5. Political Science, Law, Education 6. Humanities 7. Technology and Engineering	1. Arts 2. Commerce 3. Law 4. Science	1. Biological & Health Sciences 2. Physical Sciences & Engineering 3. Social Sciences & Education 4. Humanities & Arts

b) comparing the subjects offered by the universities. If the subject was found to be common to many universities, a category was created.

c) comparing the list of categories thus created with the categories used by the Higher Education Statistics Agency (HESA).

d) comparing the categories obtained so far with the domain labels in three NS dictionaries (NODE CD-ROM, CED CD-ROM, and the Chambers 21st Century Dictionary Online) and in two advanced learner's dictionaries (MEDAL1, and LED CD-ROM). This often helped with naming problematic categories; for example, the universities are not consistent in classifying Anthropology: the Universities of Cambridge and Oxford classify it under Anthropology & Archaeology, the University of Durham has it as Anthropology including Human Sciences, whereas the University of London had two kinds of Anthropology – Biological and Social. The NS dictionaries are unanimous in labelling; they contain separate labels for Anthropology and Archaeology.

The result of this process was a list of 33 domain categories (see Table 99 in Appendix 4 for the comparison of the domain categories found in 10 UK universities, HESA subject classification, and the domain labels in five dictionaries), which was reduced to 28 during the downloading process as 5 categories did not have sufficient online electronic journals available. Journals from the 5 omitted categories were included within other categories:

- some journals from the Classics and Ancient History category were included within History, Philosophy, and Linguistics.
- History of Art, Creative Arts (Craft & Design), and Drama, Theatre and Dance were merged into a single category Arts and Art History.
- Cultural and Media Studies journals were included within Social Sciences.

Every attempt was made to make the corpus as balanced as possible. The target size for each domain category was 2-3 million words. After examining a sample of texts, it was decided that 400 texts per domain would be downloaded. However, in some categories (e.g. Biochemistry), it was noticed that the texts are considerably shorter; hence, the target number of texts to download was increased accordingly.

Table 8. CAJA: Number of journals, articles, percentage of texts, number of texts per journal, and average words per text by domain subcorpora.

SUBCORPUS	Number of journals	Number of texts	%	number of texts per journal	Average words per text
Anthropology	45	348	2.7	7.7	7941
Archaeology	28	312	2.4	11.1	7112
Architecture	35	387	3.0	11.1	4845
Arts and Art History	57	350	2.7	6.1	7061
Biochemistry	104	808	6.2	7.8	4976
Biology	99	748	5.7	7.6	5364
Business and Management	130	492	3.8	3.8	7096
Chemistry	86	664	5.1	7.7	4482
Computer Science	100	534	4.1	5.3	8286
Economics	106	431	3.3	4.1	7049
Education	154	496	3.8	3.2	6524
Engineering	75	581	4.4	7.7	4689
Finance	88	460	3.5	5.2	6528
Geography, Earth and Env. Studies	74	400	3.0	5.4	6087
History	143	505	3.9	3.5	8866
Law	52	405	3.1	7.8	8819
Linguistics	123	473	3.6	3.8	8262
Mathematics	74	386	2.9	5.2	6690
Medicine and Health Sciences	70	460	3.5	6.6	5023
Music	37	355	2.7	9.6	8046
Philosophy	78	522	4.0	6.7	7767
Physics	49	356	2.7	7.3	6405
Politics, Government & Int. Relations	77	382	2.9	5.0	7150
Psychology	67	343	2.6	5.1	7260
Social Sciences	64	336	2.6	5.3	6581
Sports	40	398	3.0	10.0	4205
Theology and Religion	56	453	3.5	8.1	7109
Veterinary Science	47	731	5.6	15.6	3415
TOTAL	2,158	13,116	100	6.1	6559

Table 8 above provides detailed statistics about the number of journals and texts by domain category. On average, 4 to 5 texts were downloaded from each journal. In some categories (e.g. Archaeology), the average number of texts per journal ended up significantly higher, mainly due to limited number of journals available online.

Word counts by domain category in Table 9 below indicate that the balance has been more or less successfully achieved. There are few significant deviations from the average size

of 3.34 million, however this was unavoidable as the final subcorpus size was obtained only after the clean-up of the texts (after the downloading process was completed).

Table 9. CAJA: Word counts and percentages of the corpus size by domain subcorpora.

SUBCORPUS	WORDS	% of words
Anthropology	2,763,491	3.3
Archaeology	2,218,977	2.7
Architecture	1,875,173	2.2
Arts and Art History	2,471,518	3.0
Biochemistry	4,020,387	4.8
Biology	4,012,438	4.8
Business and Management	3,491,298	4.2
Chemistry	2,976,055	3.6
Computer Science	4,424,489	5.3
Economics	3,038,286	3.6
Education	3,235,660	3.9
Engineering	2,724,588	3.3
Finance	3,003,076	3.6
Geography, Earth and Environmental Studies	2,434,735	2.9
History	4,477,520	5.4
Law	3,571,773	4.3
Linguistics	3,907,993	4.7
Mathematics	2,582,188	3.1
Medicine and Health Sciences	2,310,598	2.8
Music	2,856,329	3.4
Philosophy	4,054,361	4.9
Physics	2,280,228	2.7
Politics, Government & International Relations	2,731,239	3.3
Psychology	2,490,115	3.0
Social Sciences	2,211,225	2.6
Sports	1,673,741	2.0
Theology and Religion	3,220,531	3.9
Veterinary Science	2,496,334	3.0
TOTAL	83,554,346	100

3.2.1.1.4 Texts

The corpus contains 13,116 written texts from 2,158 different academic journals. All the texts were downloaded from online journals. Many journals (but not all) in the corpus are available in both electronic and paper format. While the most common text type is an article,

the corpus also contains reviews, reports, and progress reports, i.e. text types (or genres) that are prevalent only in specific domain categories, especially scientific ones. Texts from special issues were avoided, due to the fact that special issues focus on a specific topic; using several articles about a specific topic may have skewed the data. In addition, special issues sometimes focus on topics that are not normally found in the discipline.

As shown in Table 100 in Appendix 4, the majority of texts (97%) were published in the period of 2004-2008, making CAJA a synchronic and up-to-date corpus; the year 2006 dominates in all the domain categories, representing over 76% of texts in total. The rest of the texts were published between 1993 and 2003. Only two categories contain texts published before 2000 (Anthropology (13 texts) and Economics (3 texts)).

35.96% of texts are single-authored, 21.44% have two authors, and 42.6% have three or more authors. Table 101 in Appendix 4 shows that the distribution of the number of authors per text varies considerably across the domain categories; Humanities disciplines, such as History, tend to have more single-authored texts than Science disciplines, such as Chemistry. In fact, multi-authored texts in science disciplines often have ten or more different authors.

3.2.1.1.5 Download procedure

The download procedure consisted of three steps; first, the list of journals for each domain category was made. The journals were ranked according to one of the three criteria described below. The second step included checking whether the journals are available in electronic form, and accessible through Aston University. In the last step, the texts from each journal were downloaded and documented. The information about the texts was exported to EndNote, for all cases where the option was available on the journal website.

The lists of journals were made by using one of the following three sources (Table 10):

- the European Reference Index for the Humanities (ERIH, 2007) initial lists;
- Journal League Tables by the Aston Business School (2007);
- Journal Citation Reports (Science/Social Sciences editions, 2006-7)[33].

The ERIH initial lists were made by 15 expert panels after consultation with national scientific communities, subject associations and special research centres or advisors (ERIH, 2008). This method represented the opinions of the academic community, and was thus considered to be highly suitable for selecting and ranking journals for CAJA.

[33] Available online at the ISI Web of Knowledge - www.isiknowledge.com (accessed via Aston e-library).

Table 10. CAJA: The distribution of domain categories by the source from which the corpus journals were selected.

ERIH lists	Journal Citation Reports	Journal League Tables
Anthropology	Biochemistry*	Business and Management
Archaeology	Biology*	Economics
Arts and Art History	Chemistry*	Finance
Education	Computer Science**	
History	Engineering**	
Linguistics	Geography, Earth and Env. Studies*****	
Music	Law***	
Philosophy	Mathematics**	
Psychology	Medicine and Health Sciences**	
Theology and Religion	Physics**	
	Politics, Government & Int. Relations***	
	Social Sciences****	
	Sports*	
	Veterinary Science*	

* Science edition, 2006
** Science edition, 2007
*** Social Sciences edition, 2006
**** Social Sciences edition, 2007
***** Science and Social Sciences editions, 2007

The ERIH initial lists divide academic journals into categories A, B, and C. According to the ERIH guidelines (2007), category A includes "high-ranking international publications with a very strong reputation among researchers of the field in different countries, regularly cited all over the world", category B "standard international publications with a good reputation among researchers of the field in different countries", and category C "research journals with an important local / regional significance in Europe, occasionally cited outside the publishing country though their main target group is the domestic academic community". It was decided to include only journals from categories A and B, with the target ratio of journals from these categories being set as 2 : 1 (i.e. ⅔ A, ⅓ B). Journals from category C were not included, as they had limited target audiences. In addition, the language of the majority of journals in category C was not English.

However, the ERIH lists were only designed for Humanities subjects. This meant that alternative sources were needed for domain categories not covered by the ERIH lists. Initially, alternative sources were sought within Aston University. The Journal League Tables published annually by the Aston Business School were used to make journal lists for the domain categories of Business and Management, Economics, and Finance. The journals listed are the ones in which the Aston Business School staff are most likely to publish. The journals are ranked on the basis of statistical data (impact factor from Journal Citation Reports – see

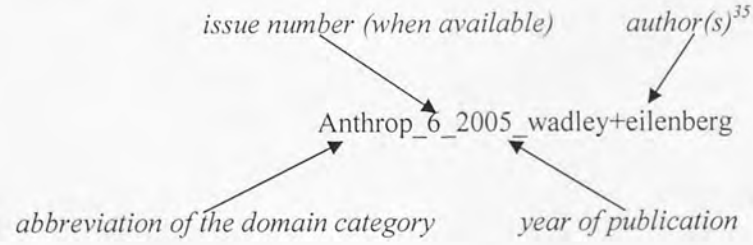
footnote 34) and "the evaluations of senior academic staff in Aston Business School and other international business schools" (Aston Business School, 2007). Considering that the Aston Business School is one of the leading business schools in the UK, it is believed that the Journal League Tables offer a good representation of the most widely used journals in Business and related subjects. The Journal League Tables use a five-category ranking system (0-4), namely World leading (4), Internationally excellent (3), Recognised internationally (2), Recognised nationally (1), and Unclassified (0). Only journals with a rank of 2 and above were selected for the corpus.

Journal Citation Reports (ISI Web of Knowledge), a website which ranks journals according to the statistical methods based on citations, was used for all remaining domain categories. The website is available in Science and Social Sciences editions. Journal Citation Reports offers different options for ranking journals. For this corpus, the journals were ranked by the impact factor[34], one of the measures used for determining the importance of a journal in its discipline. In most cases, the top 100 journals were selected for the corpus.

Architecture proved to be the only problematic category, as it did not feature in any of the three sources used. The journal list for this category was constructed by visiting various online links with information about available electronic journals for Architecture (e.g. websites of universities that offer courses in Architecture).

3.2.1.1.6 File naming, conversion and cleanup

The system used for naming the files was as follows:



This naming method made it easy to identify any duplicates, and allowed the grouping of files by publisher (rather than by domain) during cleanup, without facing difficulties in grouping

[34] "The journal impact factor is the average number of times articles from the journal published in the past two years have been cited in the JCR year." (Source: ISI Web of Knowledge, Journal Citation Reports – Help, http://admin-apps.isiknowledge.com/JCR/help/h_impfact.htm)

[35] Author information was provided as follows: author's surname for a single author, authors' surnames linked by "+" for two authors, and the first author followed by "etal" for three or more authors (e.g. smithetal).

them back into domain categories afterwards. Also, as the name of the file (or text) offers a great deal of information about the text already, there was no need to put any metadata within each file (which would be a very time-consuming task), which would then have to be ignored for language analyses.

The texts were downloaded in both HTML and PDF format, if available, with HTML being preferred[36]. The files were converted into plain text format (TXT) using a number of different tools (see Figure 9); the HTML files were converted directly into TXT format, whereas the PDF files were converted into HTML or DOC format first, and then into TXT[37]. In order to be able to use Sketch Engine, the corpus files needed to be in the same encoding. UTF-8 encoding was selected based on the recommendation of the Sketch Engine team (Sketch Engine is presented in more detail in 3.3.1).

Figure 9. CAJA: Tools and processes used for conversion and cleanup.

FILE TYPE	*PROCESS*	*SOFTWARE*
PDF		
	Conversion	1. PDF Ripper 2.01 2. ABC Amber PDF Converter
HTML	1. identifying file encoding 2. if encoding other than UTF-8, changing to UTF-8 3. changing HTML entities to single characters	1. NthGrep Pro 2.3 2. Unifier 4.0 3. InfoRapid Search and Replace
	Conversion	Detagger 2.4.0.12
TXT	Cleanup Extra cleanup + removal of line breaks and double spaces	1. Custom-made program written in Java for each publisher (written by a contracted programmer according to instructions established during the test conversion of HTML files) 2. InfoRapid Search and Replace

[36] Although websites allow HTML files to be saved in TXT format (the required format for the corpus files), this option was not used, as saving HTML file as TXT results in TXT file containing undesired text such as HTML tags and hyperlinks.

[37] The two-step conversion procedure for PDF files was used because the results were much better than the results from direct conversion of PDF into TXT.

Additional cleanup was needed to remove any remaining information that was not part of the text (e.g. menus, headers and footers), and 'unwanted' parts of the text (e.g. Reference sections, footnotes, author's biographical information). The unwanted parts were omitted because it was believed they could skew the frequency counts and other statistical information. Moreover, some contained personal information about the author, for example an email address or even a phone number.

3.2.1.1.7 Annotation

Lemmatisation and part-of-speech (POS) tagging, or grammatical tagging, and are essential for any lexicographic analysis. This enables advanced searches (e.g. searches by lemma, searches using corpus query language), and, in Sketch Engine in particular, the use of functions such as Word Sketch (see 3.3.1.2.2) which have been a key part of data analysis in this research.

Therefore, before uploading into the corpus data into Sketch Engine, the corpus was lemmatised and POS-tagged with TreeTagger, a tool developed by Helmut Schmid at the University of Stuttgart. The tagset used by TreeTagger (see Table 102 in Appendix 4)[38] is a slightly modified version of the Penn Treebank Tagset.

3.2.2 Secondary data

So far, the primary resource used in this research has been described. Additional resources, presented in this section, were consulted when building the sample entries. These resources include existing corpora and dictionaries, and the Pattern Dictionary of English Verbs (PDEV). Existing corpora provided the information on the registers and genres of the academic English not represented in CAJA (e.g. academic speech). Existing dictionaries were a source of ideas for the presentation of information in DOAE, helped (along with corpora) in the final stages of compiling dictionary entries to compensate for my lack of experience, and were used for the evaluation of the sample DOAE entries.

3.2.2.1 Corpora

[38] Also available at http://www.sketchengine.co.uk/tagsets/penn.html.

3.2.2.1.1 British Academic Spoken English (BASE) corpus

The BASE corpus was compiled between 1998 and 2005 at the University of Warwick and the University of Reading and contains 1,644,942 words or just over 196 hours of recorded speech. It consists of 160 lectures and 39 seminars, equally divided into four broad disciplines: Arts and Humanities, Life and Medical Sciences, Physical Sciences, and Social Studies and Sciences. There are many subject areas represented (see Table 11), however due to poor representativeness (and in many cases limited contents) these cannot be used as lower-level categories.

Table 11. Subject areas in the BASE corpus.

Biological Sciences	History of Art
Business	Law
Chemistry	Mathematics
Classics	Medicine
ELT and Linguistics	Meteorology
Economics	Philosophy
Engineering	Politics
English Literature	Psychology
Film and TV	Statistics
History	

The speakers are non-students (presumably staff, but no detailed information is given) and students while the audience ranges from pre-sessional students and undergraduates (years 1 to 4) to postgraduates and staff.

It should be noted that the BASE corpus within Sketch Engine (see 3.3.1), the main software used for data analysis in this thesis, contains lectures only. The BASE lectures make up 1,212,251 words. In Sketch Engine, the size of the corpus is slightly higher, namely 1,252,256 words, because the annotations that appear within square brackets (e.g. [[laughter]]) are included in the word count[39].

The BASE corpus has been used to compare the behaviour (meanings, frequency, phraseology, etc.) of a sample of words in spoken academic discourse with the behaviour of words in written academic discourse (CAJA).

3.2.2.1.2 Michigan Corpus of Academic Spoken English (MICASE)

The MICASE corpus was developed from 1997 to 2002 at the University of Michigan. It contains 152 transcripts of over 190 hours of speech or 1,848,364 words. MICASE is a corpus

[39] Paul Thompson (one of the directors of the BASE project), personal communication.

of academic speech at an American university, and can thus be considered an American equivalent of the BASE corpus. However, while the BASE corpus contains only lectures and seminars, the selection of speech events in MICASE is much more diverse (see Table 12).

Table 12. Speech events in MICASE.

Advising Sessions	Colloquia (public lectures)
Discussion Sections	Lectures (Small and Large)
Meetings	Office Hours
Seminars	Dissertation Defenses
Interviews	Lab Sections
Service Encounters	Student Presentations
Study Groups	Tours (Campus/Museum)

As described in the MICASE Manual (2003:5), "speech event attributes include the type of event, the subject area of the event, the extent to which an event is monologic or interactive, as well as the academic role or level of the majority of participants (e.g., whether the class was a graduate or an under-graduate class, or whether a meeting was primarily of senior faculty members)". More detailed information about participants is also available, such as gender, age group, whether NS or NNS, and native language (if native language is other than North American English).

Table 13. Academic divisions and disciplines in MICASE.

ACADEMIC DIVISION	DISCIPLINES
BIOLOGICAL AND HEALTH SCIENCES (BS)	Includes Biology, Biochemistry, Dentistry, Genetics, Immunology, Natural Resources, Neuroscience, Nursing, Pathology, Pharmacy, Physiology, Public Health
PHYSICAL SCIENCES AND ENGINEERING (PS)	Includes Astronomy, Chemistry, Computer Science, Engineering (all), Geology, Mathematics, Physics, Statistics, Technical Communication
SOCIAL SCIENCES AND EDUCATION (SS)	Includes Anthropology, Business Administration, Communication, Economics, Education, History, Public Policy, Political Science, Psychology, Social Work, Sociology, Urban and Regional Planning
HUMANITIES AND ARTS (HA)	Includes Area Studies (all), Architecture, Classics, Comparative Literature, English, Fine Arts (all), Foreign Languages, History of Art, Information and Library Science, Linguistics, Philosophy, Women's Studies

The speech events are divided into four broad categories (called divisions in the MICASE Manual): Biological and Health Sciences, Physical Sciences and Engineering, Social Sciences and Education, and Humanities and Arts. These categories are very similar to the ones used by the BASE and BAWE corpora. It is noteworthy that academic events in professional schools such as Medicine and Law were excluded. Each category contains speech events from

many different disciplines, 48 to be exact, (Table 13 above), however it should be noted that this means that on average, each discipline contains only three speech events (some contain only one). Consequently, there is not enough data to conduct lexicographic analysis on individual disciplines.

MICASE was consulted for a similar reason as BASE, namely to compare spoken academic discourse with written academic discourse.

3.2.2.1.3 British Academic Written English (BAWE) corpus

The BAWE corpus is a collection of assessed student writing, containing 6,506,995 words. In Sketch Engine, the size of the corpus is 8,336,262 words (see 3.3.1 for a detailed explanation). The corpus was collected in the period 2004-2007 at three UK universities: the University of Warwick, the University of Reading, and Oxford Brookes University. The corpus contains 2,896 texts (or 2,761 assignments[40]) from 35 disciplines that are divided into four broad categories. The categories are the same as those used in BASE. More information on BAWE is available in the BAWE corpus manual (Heuboeck et al., 2008).

The contributors were 1,039 undergraduate (Year 1-3) and Master's students (Thompson, 2008). Only assignments with a mark of 60 and above were collected, which means that the corpus is essentially a collection of higher-grade academic student writing. Out of 2,761 assignments, 1,953 were produced by native speakers of English, and 808 by non-native speakers.

The BAWE corpus has been used to compare student writing, or non-expert academic writing, with expert writing (i.e. CAJA).

3.2.2.1.4 British National Corpus (BNC)

The BNC is a 100-million word corpus of general British English, compiled between 1991 and 1994. It consists of 90% written texts, and 10% speech. The written texts range from newspaper articles and popular fiction to academic essays, while the spoken part of the corpus consists of transcribed informal conversations, business meetings, radio shows and phone-ins.

The BNC has already been used as a resource for many dictionaries, for example the Oxford Advanced Learner's Dictionary and the Longman Dictionary of Contemporary English.

[40] As stated in the BAWE Corpus Manual (Heuboeck et al., 2008), "the difference in the numbers of assignments and texts is due to the fact that some assignments ... consist of more than one text".

And although the corpus is almost 15 years old and contains texts from the 1960s onwards, it is still considered by many to be a good representation of recent English. The BNC was used in this research in addition to CAJA to check the wording of the definitions.

3.2.2.2 Existing dictionaries

Existing dictionaries present an important source of reference for publishers and lexicographers. Publishers, i.e. their marketing departments, compare their own dictionaries with competitors, focussing on the information that can be put on the cover, such as the number of headwords and examples. Lexicographers use existing dictionaries to compare the treatment of words and to identify any useful features that can be added to their own dictionary.

Also, for this researcher, consulting other dictionaries is necessary to maintain some consistency between the dictionaries that students are likely to use before starting their studies, and the Model being created here. This should reduce the list of new skills the students will need to learn in order to use DOAE effectively.

In this research, existing dictionaries were used in three ways. Initially, existing dictionaries served as a source of ideas for the presentation of information in DOAE. Of course, attention has been paid to both more and less effective practices; whereas good practices can provide the inspiration to find even better solutions, bad practices can avert us from mistakes made by others. Then, after the draft DOAE sample entries had been completed, existing dictionaries were used to identify meanings of the words that may have been missed during the analysis of the CAJA data, perhaps due to my lack of experience. The decision as to whether these meanings need to be included in DOAE has been made by checking their occurrence in academic corpora (CAJA, BAWE, BASE and MICASE). Finally, the completed sample entries for the proposed DOAE were compared with the whole of the entries in existing dictionaries, for evaluation purposes.

The types of dictionaries consulted were general NS dictionaries, advanced learners' dictionaries, and dictionaries for university students. Online Etymology Dictionary was consulted (in addition to CED CD-ROM and LED CD-ROM) only to obtain etymological information on the sample entries. The dictionaries used for each of these types are presented in Table 14 below. As the proposed format for DOAE is electronic (see Chapter 5), the electronic formats of the dictionaries were used.

Table 14. Model for DOAE: Dictionaries consulted during the analysis.

Type of dictionary	Full name	format	Abbreviation used
native-speaker	Collins English Dictionary (Desktop Edition), 1st edition, 2004	CD-ROM	CED CD-ROM
native-speaker	New Oxford Dictionary of English, 1998	CD-ROM (iFinger)	NODE CD-ROM
native-speaker	Dictionary.com Unabridged (based on Random House Dictionary, 2010)	online	Dictionary.com
advanced learner's	Oxford Advanced Learner's Dictionary, 7th edition, 2005	online	e-OALD
advanced learner's	Collins Cobuild Dictionary for Advanced Learners, 3rd edition, 2001	CD-ROM	COBUILD CD-ROM
advanced learner's	Cambridge Advanced Learner's Dictionary, 3rd edition, 2008	online	e-CALD
advanced learner's	Macmillan English Dictionary, 2nd edition, 2007	online	e-MED
advanced learner's	Longman Dictionary of Contemporary English, 4th edition, 2006	online	e-LDOCE
college	Merriam-Webster Collegiate Dictionary, 11th edition, 2003	CD-ROM	MWCD CD-ROM
student	Longman Exams Dictionary, 1st edition, 2006	CD-ROM	LED CD-ROM
etymological	Online Etymology Dictionary (http://www.etymonline.com/)	online	-

Two general NS dictionaries have been used, the Collins English Dictionary (2004; CED CD-ROM), and the New Oxford Dictionary of English (1998; NODE CD-ROM). CED is a comprehensive NS dictionary that follows British lexicographic tradition, but has adopted some conventions of American publishers, such as offering more encyclopaedic information. NODE, arguably the most corpus-driven NS dictionary so far, can be regarded as a successful modern hybrid of the NS dictionary and the learner's dictionary. The dictionary takes a revolutionary approach to the analysis of meaning, dividing core senses from subsenses. Moreover, the central and most current meanings are given first. Valuable syntactic information is provided by stressing common phrases (in bold text) in examples. More information on the advantages and limitations of NODE can be found in Landau (1999).

There are five widely-known advanced learners' dictionaries; the Oxford Advanced Learner's Dictionary (e-OALD, 2005), the Collins Cobuild Advanced Learner's Dictionary (COBUILD CD-ROM, 2001), Longman Dictionary of Contemporary English (e-LDOCE,

2006), the Cambridge Advanced Learner's Dictionary (e-CALD, 2008), and Macmillan English Dictionary (e-MED, 2007). This type of dictionary was the first to introduce certain new features, such as a corpus-based approach, which have been later adopted by some general NS dictionaries. Because each of these five dictionaries has its own specific features and ways of presenting information, all five dictionaries were consulted.

Dictionaries for university students were presented in more detail in section 2.2.1. MWCD CD-ROM (2003), as a dictionary for NSs, and LED CD-ROM (2006), as a dictionary for NNSs, were consulted.

Another dictionary that has been added to the list of dictionary resources on the basis of the results of the survey of student dictionary use was Dictionary.com. Dictionary.com is an online resource that is a collection of several dictionaries, from American general-purpose dictionaries to technical dictionaries and dictionaries of idioms. Sample searches revealed that the main source of entries is Dictionary.com Unabridged (based on Random House Dictionary, 2010), which was then selected to represent Dictionary.com in this research.

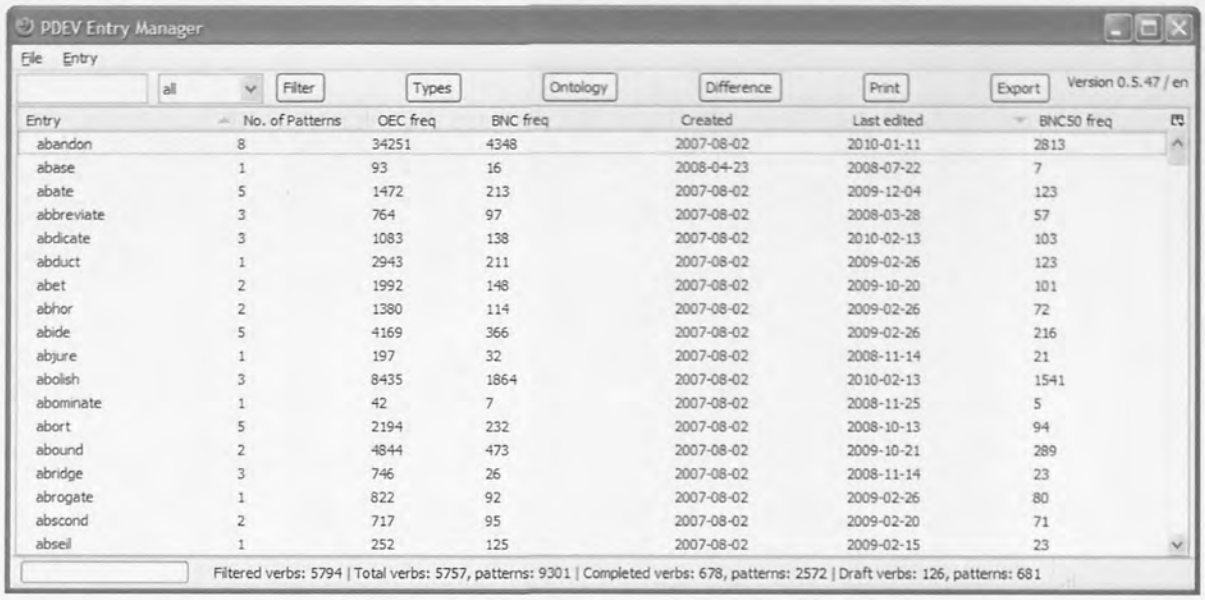
3.2.2.3 Pattern Dictionary of English Verbs (PDEV)

PDEV aims "to provide an inventory of all the normal patterns of use of all the normal verbs in English" (Hanks & Ježek, 2008:391). PDEV uses Corpus Pattern Analysis (CPA), "a new technique for mapping meaning onto words in text" (http://nlp.fi.muni.cz/projekty/cpa/). CPA is based on the Theory of Norms and Exploitations (Hanks, 2004; Hanks, forthcoming) which identifies word patterns in a large corpus first and only then tries to map them onto meanings.

Access to the database of PDEV is available in the form of a Firefox extension[41]. Figure 10 shows part of the PDEV Entry Manager, and Figure 11 shows the identified patterns of the verb *abate*. In addition, the user can also access various corpora (via a customized interface of Sketch Engine – for a more detailed description of the standard Sketch Engine interface, see 3.3.1) that have been used for CPA. Particularly noteworthy are BNC50 (the 50-million-word version of the BNC) where the user can look at the concordances of the verb, with the verb in each concordance line being assigned the pattern number from the PDEV entry (Figure 12), and OEC which is not available otherwise.

[41] Instructions are available at http://nlp.fi.muni.cz/projekty/cpa/.

Figure 10. PDEV Entry Manager.



Entry	No. of Patterns	OEC freq	BNC freq	Created	Last edited	BNC50 freq
abandon	8	34251	4348	2007-08-02	2010-01-11	2813
abase	1	93	16	2008-04-23	2008-07-22	7
abate	5	1472	213	2007-08-02	2009-12-04	123
abbreviate	3	764	97	2007-08-02	2008-03-28	57
abdicate	3	1083	138	2007-08-02	2010-02-13	103
abduct	1	2943	211	2007-08-02	2009-02-26	123
abet	2	1992	148	2007-08-02	2009-10-20	101
abhor	2	1380	114	2007-08-02	2009-02-26	72
abide	5	4169	366	2007-08-02	2009-02-26	216
abjure	1	197	32	2007-08-02	2008-11-14	21
abolish	3	8435	1864	2007-08-02	2010-02-13	1541
abominate	1	42	7	2007-08-02	2008-11-25	5
abort	5	2194	232	2007-08-02	2008-10-13	94
abound	2	4844	473	2007-08-02	2009-10-21	289
abridge	3	746	26	2007-08-02	2008-11-14	23
abrogate	1	822	92	2007-08-02	2009-02-26	80
abscond	2	717	95	2007-08-02	2009-02-20	71
abseil	1	252	125	2007-08-02	2009-02-15	23

Filtered verbs: 5794 | Total verbs: 5757, patterns: 9301 | Completed verbs: 678, patterns: 2572 | Draft verbs: 126, patterns: 681

The PDEV entries contain the percentage information for each pattern of the verb (see Figure 11), so it was interesting to see how much the pattern percentages in general English differ from those in academic English. Therefore, the PDEV preliminary results were compared with the sample (verb) entries to see to what degree the PDEV patterns map onto word meanings in academic English.

Figure 11. PDEV: Patterns for the verb *abate*.

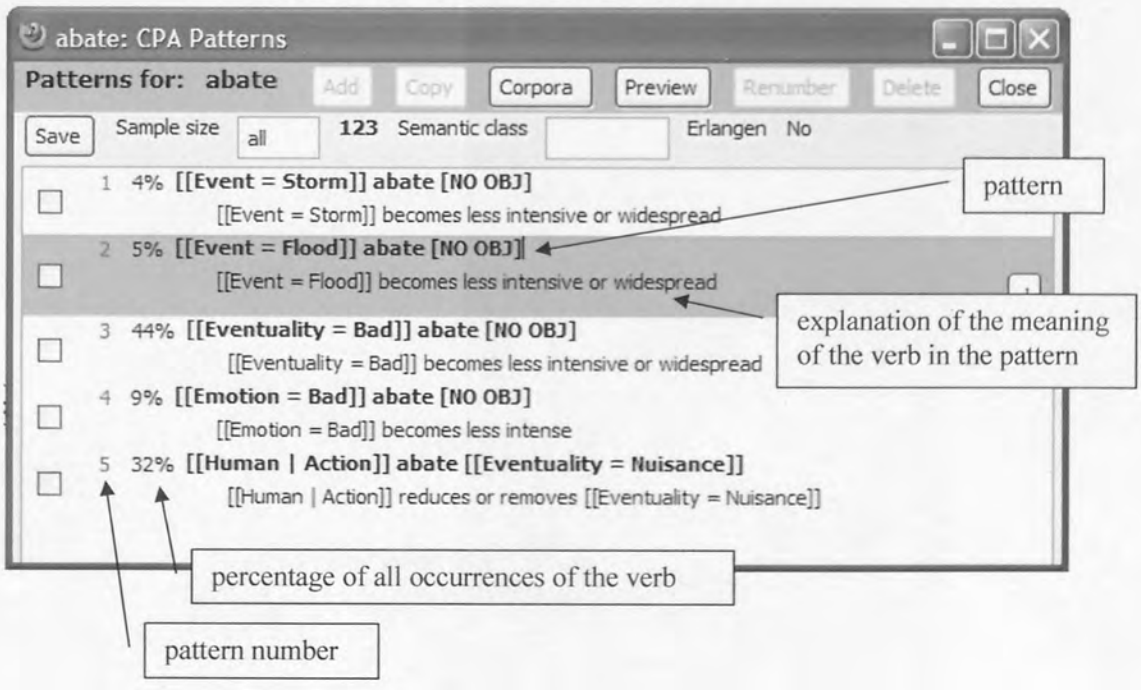
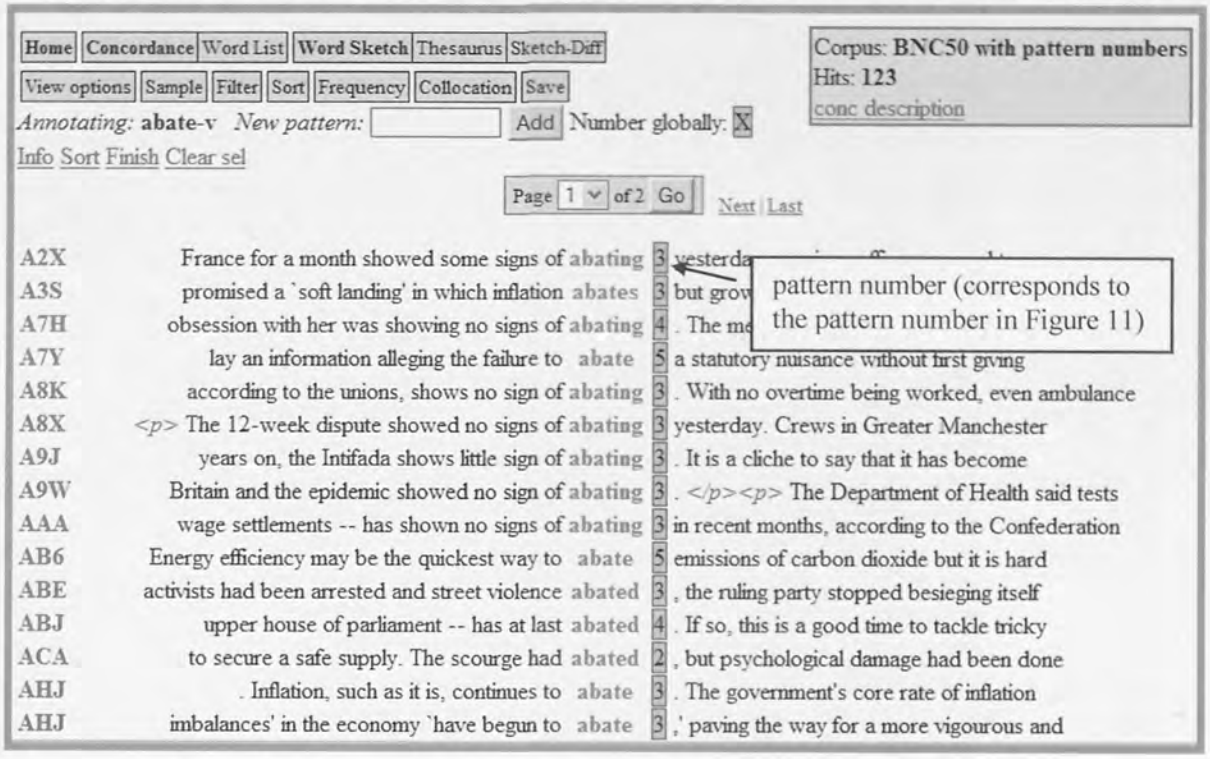


Figure 12. PDEV: Concordance lines of *abate* with assigned pattern numbers (BNC50 corpus).



3.2.2.3.1 CPA ontology and the Brandeis Semantic Ontology

The CPA technique, used by PDEV, introduces a new approach to the use of ontology in lexicography. The CPA ontology differs from traditional conceptual ontologies such as WordNet by grouping words according to their syntagmatic behaviour. According to Hanks and Ježek (2008:395), the CPA ontology is "a statistically based structure of collocational preferences", which they call shimmering lexical sets. Lexical sets shimmer because their members are not fixed; different words often have the same lexical sets which contain different members. Hanks and Ježek (2008:399) offer the example of verbs *wash* and *amputate*, which typically have [Body Part] as their direct object; while some members of the lexical set are shared by both verbs (e.g. *leg*, *arm*), others are not (e.g. *face* and *hair* can be *washed* but not *amputated*).

The CPA approach has already had an impact on one of the traditional ontologies, namely the Brandeis Semantic Ontology (BSO[42]). The authors of BSO have decided to use the

[42] The Brandeis Semantic Ontology is available online at http://eurydice.cs.brandeis.edu/BSOonline/newBSO/BSObrowser.py.

CPA results when developing BSO (Rumshisky et al., 2006). Both CPA and BSO are used in this thesis, in the following ways:

- The CPA method of identifying patterns first and then meanings has been tested on our sample entries. This includes listing members for each lexical set in the pattern. Because CPA often provides a superordinate term or a frequent collocate for semantic types in a pattern, certain patterns tend to have quite long and complex descriptions, like this pattern for *admit*:

  [[Human 1 | {Institution = Hospital}]] admit [[Human 2 = Patient]] {to | into [[Institution = Hospital]] | {care}} {for [[Activity = Treatment]]}

  The pattern consists of six sub-patterns:

  1. **[[Human 1]] admit [[Human 2 = Patient]]**
  2. **[[Human 1]] admit [[Human 2 = Patient]] to | into [[Institution = Hospital | {care}]]**
  3. **[[Human 1]] admit [[Human 2 = Patient]] to | into [[Institution = Hospital | {care}]] for [[Activity = Treatment]]**
  4. **[[Institution = Hospital]] admit [[Human 2 = Patient]]**
  5. **[[Institution = Hospital]] admit [[Human 2 = Patient]] to | into [[Institution = Hospital | {care}]]**
  6. **[[Institution = Hospital]] admit [[Human 2 = Patient]] to | into [[Institution = Hospital | {care}]] for [[Activity = Treatment]]**

  For dictionary purposes, it is better to keep sub-patterns separate because it is useful to retain a wider range of examples. Also, identical semantic types in different sub-patterns may have different lexical sets.

  In this thesis, lexical sets have been reduced to general superordinate items only, with members listed separately. Single square brackets are used instead of double ones, and specific words/collocates are offered in bold in DOAE patterns. For example, sub-pattern 1 above would be written like this[43]:

  [Human 1] **admit** [Human 2]

  [Human 1]=*nurse*...
  [Human 2]=*patient, child, children, man, people, person*...

[43] Collocates listed for lexical sets [Human 1] and [Human 2] have been obtained by creating the Word Sketch of the verb *admit* in the BNC, and analysing concordances of grammatical relations object and subject.

Also, a distinction is made between the passive and active forms, so for example if *admit* was found to occur in one of the sub-patterns above only in the passive, it would be written **be admitted**[44].

- BSO ontology has been used to identify names (i.e. superordinates) for lexical sets[45] in meaning patterns. Superordinates have then been used when writing definitions. Their role must therefore not be equated with the role they have in an ontology. Thus, BSO was not followed blindly – if a more suitable superordinate than the one offered by BSO (e.g. a specific collocate) was deemed more appropriate, it has been used instead.

3.3 Software

Tools in this thesis can be divided into corpus tools and lexicographic tools. Corpus tools have been used for the analysis of primary corpus data (CAJA), as well as secondary corpus data (e.g. BASE). To ensure the comparability of corpus results, corpora have been analysed with the same corpus tool (Sketch Engine) whenever possible. MICASE, however, is not available in Sketch Engine, so its own respective software system has been used. The lexicographic tool used in this thesis is called TshwaneLex, and has been used to store information obtained during corpus analysis, and to compile the Model for DOAE.

3.3.1 Sketch Engine[46]

Sketch Engine (Kilgarriff et al., 2004) is an online corpus software system which gives the users access to not only most of the basic functions (e.g. frequency lists, concordances, collocations), but also to advanced functions (e.g. Word Sketch, Thesaurus), which are of particular use to lexicographers.

Sketch Engine gives the user access to 24 corpora[47], including BASE, BAWE, and the BNC. These three corpora have been used when designing sample entries so it has been very

[44] In this case, *be* is a lemma. If the use of the passive is limited to a particular form (e.g. *is admitted*), then that specific form is provided in the pattern.
[45] CPA ontology cannot be used for this purpose as it does not offer the option of bottom up searches; i.e. searches of semantic types associated with individual lexical items.
[46] Since March 2010, Sketch Engine is offered through a new interface, named the Corpus Architect (http://ca.sketchengine.co.uk/).
[47] Many corpora in Sketch Engine are corpora of languages other than English, such as French, Chinese, Russian, and Slovene.

useful to have them all available in the same corpus software system. In addition, Sketch Engine gives the users the option of uploading their own corpora. CAJA was initially uploaded into the beta version of Sketch Engine with additional features, used especially by lexicographers. In September 2009, the expanded functionality of the beta version was added to the normal version. At this point, it is important to mention that Sketch Engine uses a different tokenisation[48] procedure to many other corpus tools and programs. The main difference lies in the fact that Sketch Engine counts each punctuation mark as a separate token (an item surrounded by spaces). This discrepancy in tokenisation results in word counts of corpora in Sketch Engine being higher than the official word counts found in the documentation (Table 15).

Table 15. Sketch Engine: Comparison of corpora sizes (official documentation vs. Sketch Engine).

	word count in documentation	word count in Sketch Engine
CAJA	83,554,346[49]	94,352,272
BASE	1,212,251	1,252,256
BAWE	6,506,995	8,336,262
BNC	100,000,000	112,181,850

Since most of the analysis has been performed in Sketch Engine, the external difference in word counts is not that problematic. In addition, the majority of corpora that have been used for comparison are available in Sketch Engine, which makes any statistical data comparable. Nevertheless, there were corpora consulted that are not available in Sketch Engine (e.g. MICASE), all using tokenisation that is different from the one used by Sketch Engine. Therefore, to make any comparison between the corpora not available in Sketch Engine and corpora available in Sketch Engine (CAJA, BASE, BAWE) valid, the sizes reported in the official corpus documentation needed to be used for the latter corpora.

The next section presents the functions available within Sketch Engine, starting with basic functions and then moving to more advanced functions, which are particularly useful for lexicographers.

[48] For explanations of the terms 'tokenisation' and 'token', see Glossary (page 27).
[49] The word count was obtained using a program written in the Java programming language.

3.3.1.1 Basic functions[50]

3.3.1.1.1 Concordance

After selecting a corpus in Sketch Engine, the user is taken to the Concordance window (Figure 13). This is a good starting point for basic queries; a word or a phrase can be searched. The basic query tool searches by lemma, so a query for *IDEA* produces a concordance for *idea* and *ideas* (Figure 14).

Figure 13. Sketch Engine: basic query window (lemma *IDEA*).

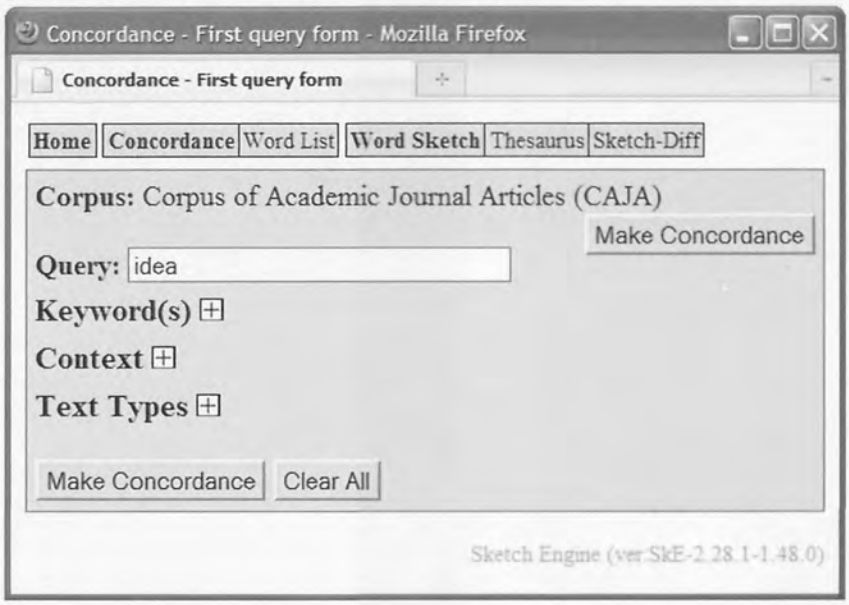
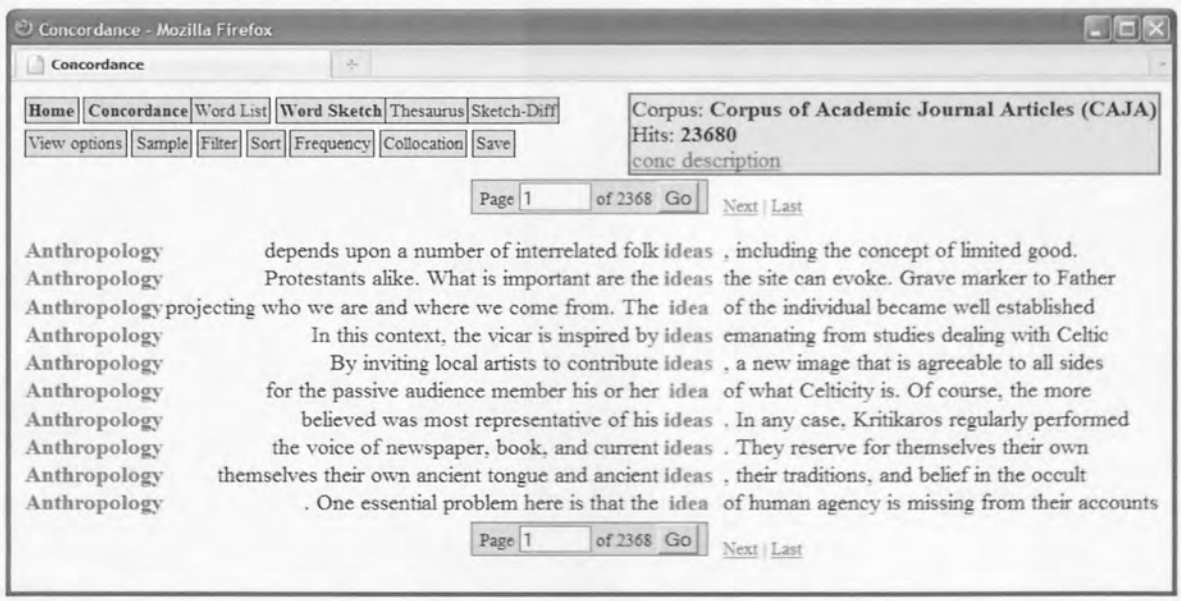


Figure 14. Sketch Engine: concordance window of a basic query (lemma *IDEA*).



[50] Explanations of many terms (e.g. concordance, lemma, node) introduced in this section can be found in the Glossary (pages 25-27).

The default setting for displaying a concordance is Key Word In Context (KWIC). Clicking on the domain label to the left of the concordance line (in blue) provides information about the source text (Figure 15), while clicking on the node (searched item, displayed in red) provides extra context (Figure 16) which can be expanded further if needed, by selecting 'expand left' or 'expand right' link.

Figure 15. Sketch Engine: source text information (for concordance line 1).

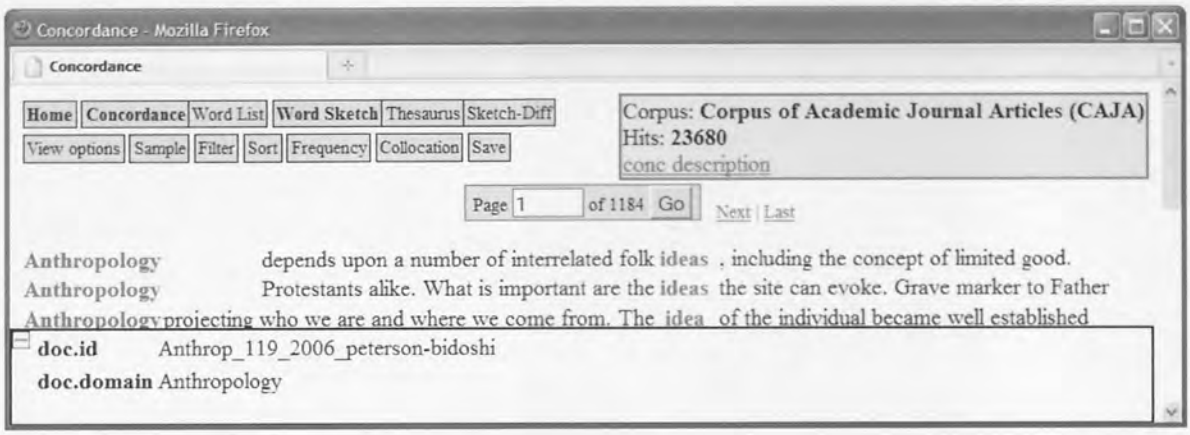
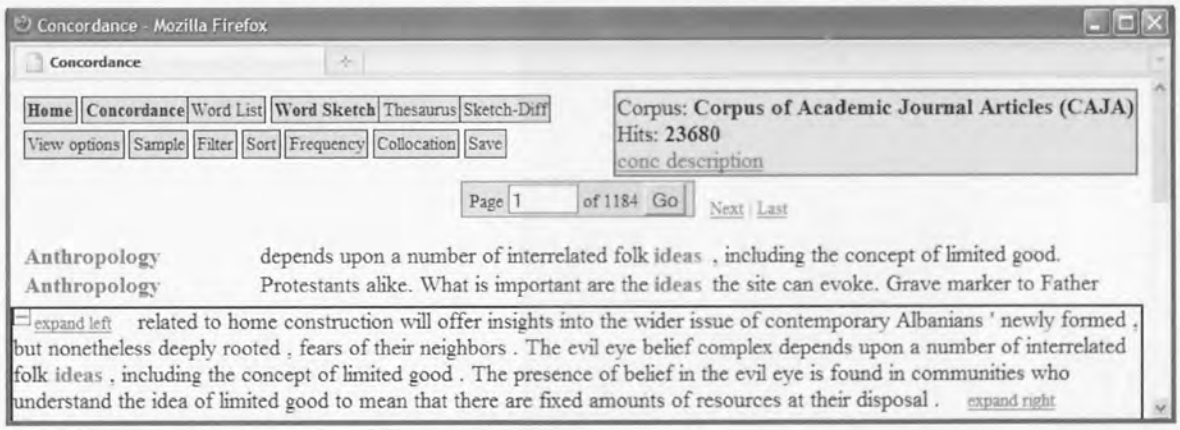


Figure 16. Sketch Engine: extra context (for concordance line 1).



More advanced query functions are available, but need to be activated by clicking a + sign next to them (Figure 13 above). The Keyword(s) function allows multi-word searches (Phrase), searches by lemma, searches by word form (allows specifying part of speech), or searches using Corpus Query Language (CQL), i.e. complex searches using regular or attribute expressions (e.g. "[tag="NNS.*"] [tag="IN.*"]" for searching a sequence of plural noun followed by a preposition). The Context function enables the user to limit the search further by specifying the word or phrase on the right and/or the left of the node. Figure 17 gives an example of an advanced query, where the noun *table* is searched up to five words to the left of

the verb *show*. Figure 18 then shows the results of the query (page 16 of 100 of the concordance).

Figure 17. Sketch Engine: advanced query.

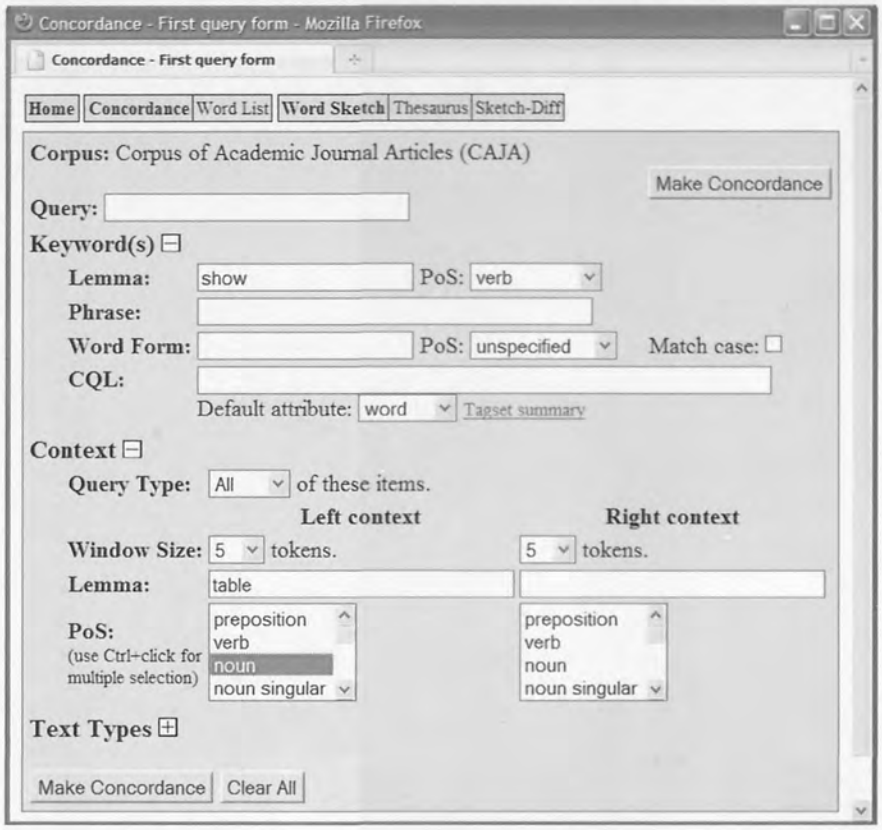
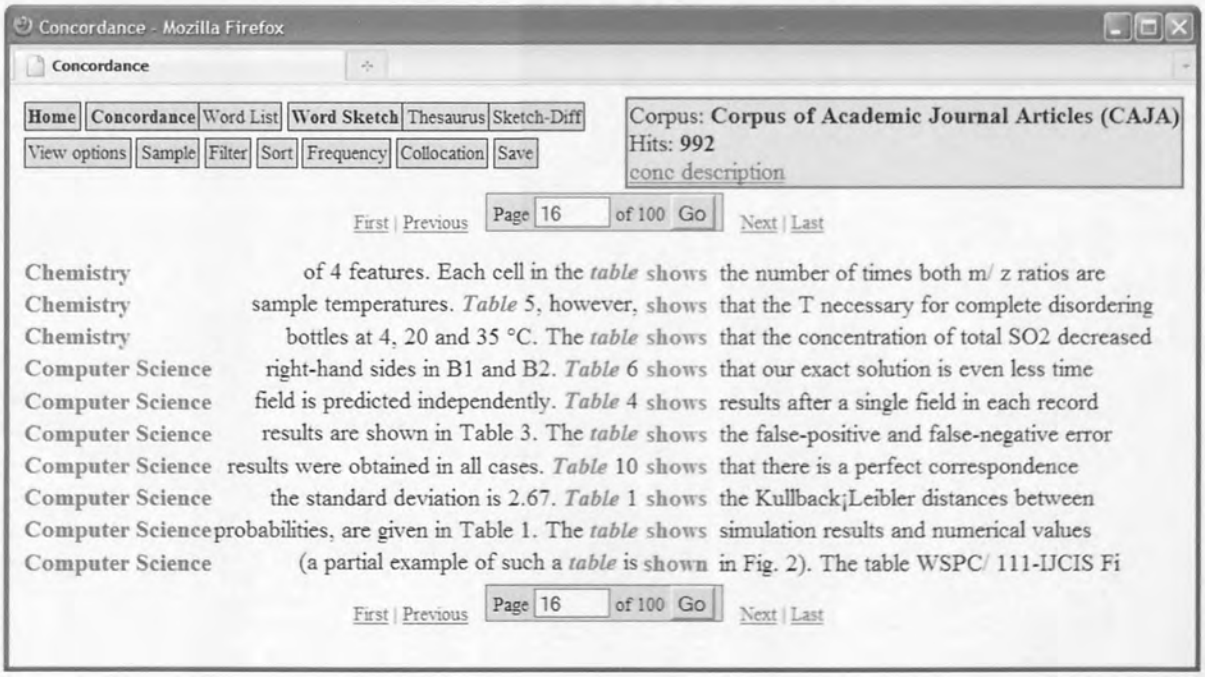
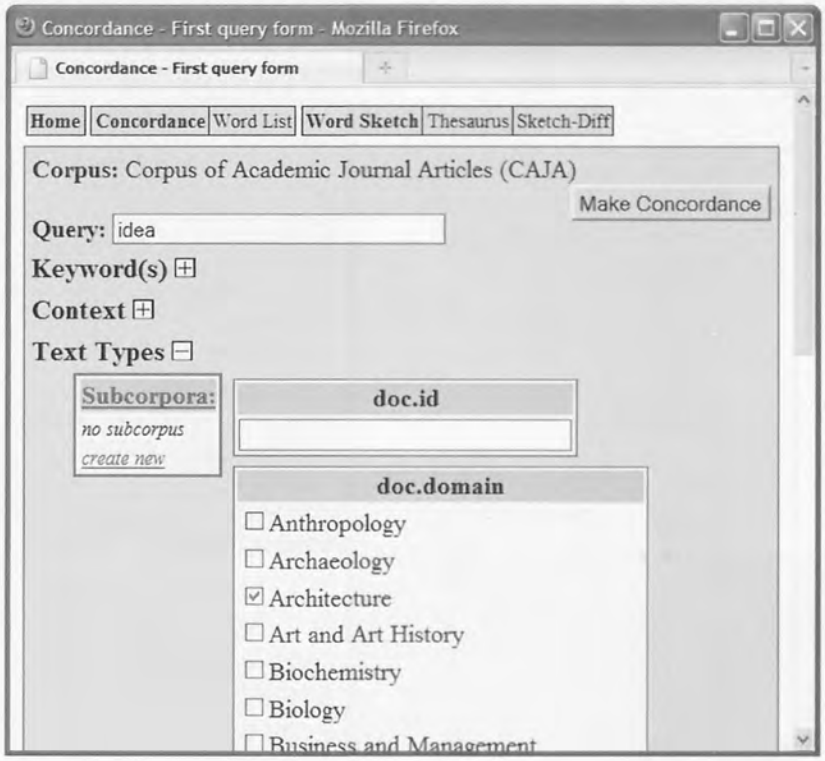


Figure 18. Sketch Engine: concordance lines of an advanced query (lemma *show*, the noun *table* up to 5 words to the left).



The Text Types function (Figure 19 below) gives the user an option to limit the search to parts of the corpus, either to specific type of document(s) (by text name – doc.id) or to user-created subcorpora. In the case of CAJA, domain subcorpora were automatically made available in Sketch Engine (listed under doc.domain), because the texts were divided into domain subcorpora before the upload.

Figure 19. Sketch Engine: Limiting search by Text Type.



Once a concordance is created, the user can manipulate it by using the buttons in the second row in the top left-hand corner of the screen (see Figure 20). Additional help with using the buttons is provided in the form of pop-up windows (not available for all the buttons) that contain a brief description of the function (Figure 20). The following paragraphs give a short description of the button functionality.

View Options gives access to settings, where the user can change the way the concordances are displayed (Figure 21 below). For example, the user can set the number of concordance lines displayed on a page (default setting is 20), set the context size on each side of the node (default setting is 40 characters), display tags indicating sentence or document breaks in the concordance lines (*Structures* box), and change the information about the source text that is displayed on the left side of the concordance line (*References* box).

Figure 20. Sketch Engine: manipulating the concordance.

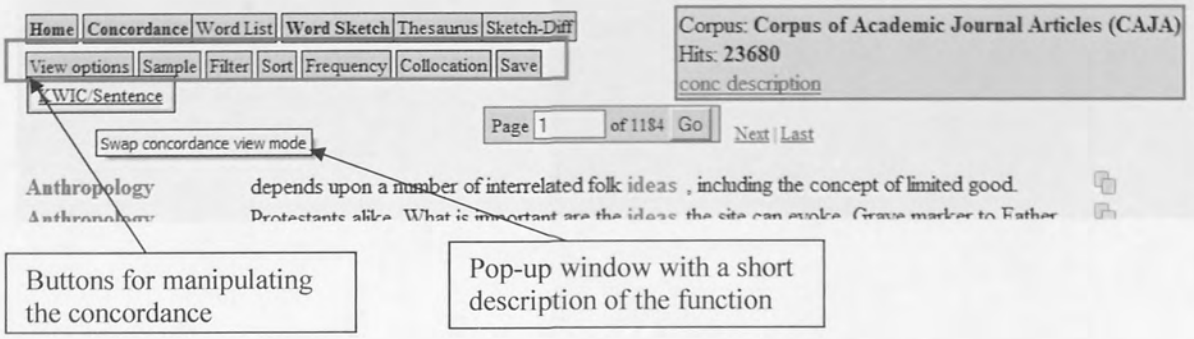


Figure 21. Sketch Engine: View Options screen.

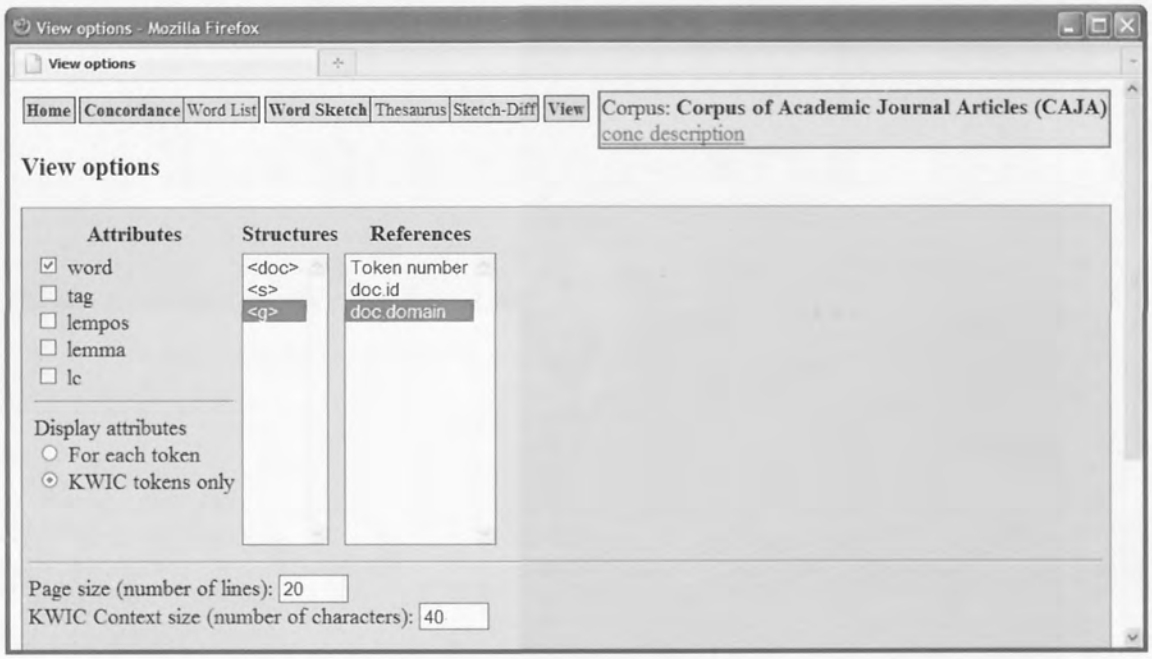
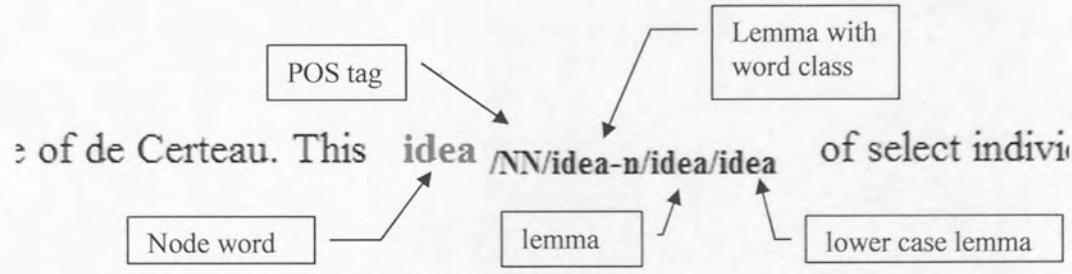


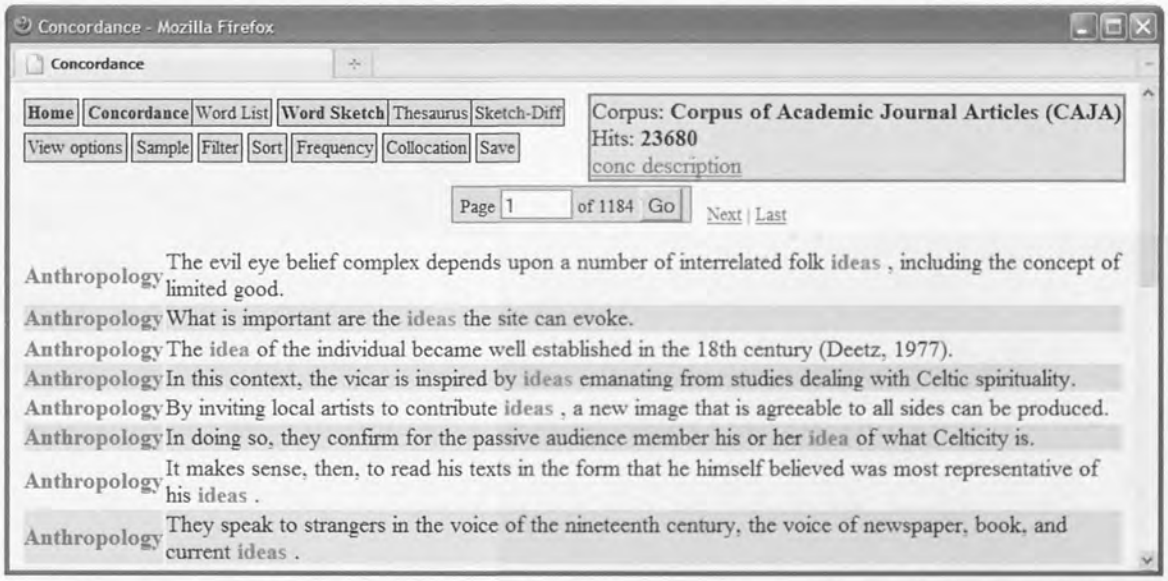
Figure 22. Sketch Engine: concordance line with all the Attributes displayed.



The user can also change the display to see more information about the node or for every word in the concordance; these can be set in the Attributes section of the screen (Figure 21 above). *Word* is selected by default, and other attributes can be added, including *tag* (POS-tags are displayed), *lempos* (lemma with word class suffix is displayed), *lemma* (lemma of the node is displayed), and *lc* (lower case lemma is displayed). Figure 22 above shows a concordance line with all the attributes of the node displayed. The attributes are also found in other functions, however they occur mostly in drop-down menus so only one of them can be selected at a time.

Moving the mouse cursor over the *View Options* button in the concordance window opens a submenu with the *KWIC/Sentence* button; clicking this button changes the concordance view mode (see Figure 23 below for an example of the *Sentence* mode).

Figure 23. Sketch Engine: Concordance lines displayed in the Sentence view mode.



The *Sample* button gives the user an option to limit the number of concordance lines by creating a random sample (the default is 250 lines). This is very useful when dealing with very frequent items, as the sample can make the analysis more manageable.

Filter (Figure 24 below) can be used to search for (positive filter) or exclude (negative filter) concordance lines containing the specified words (by lemma or word form) or multi-word items (i.e. phrases).

Figure 24. Sketch Engine: the Filter screen.

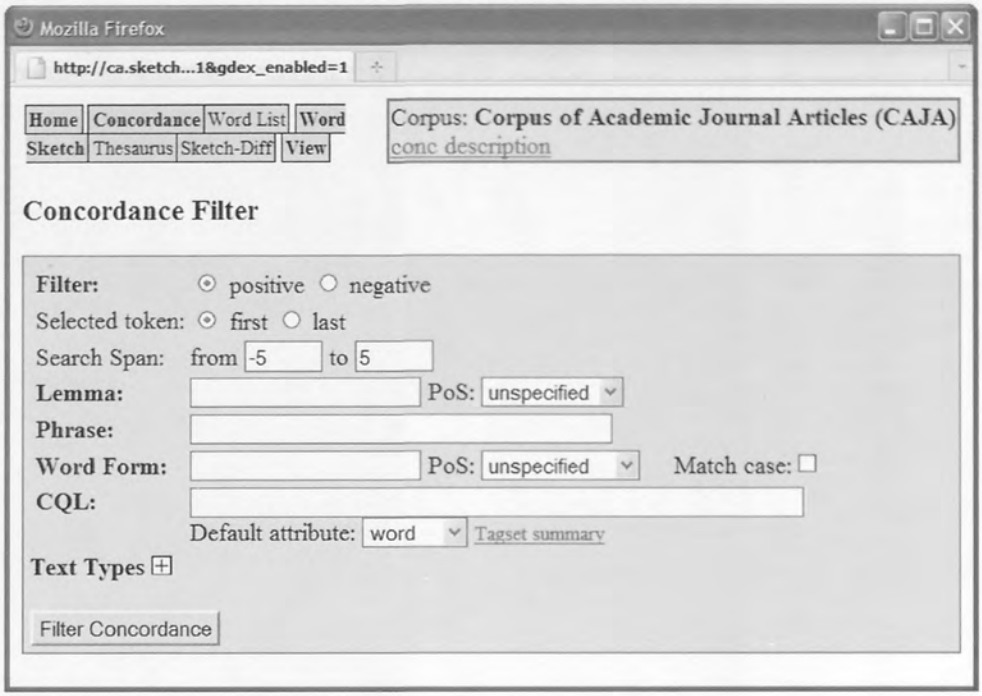
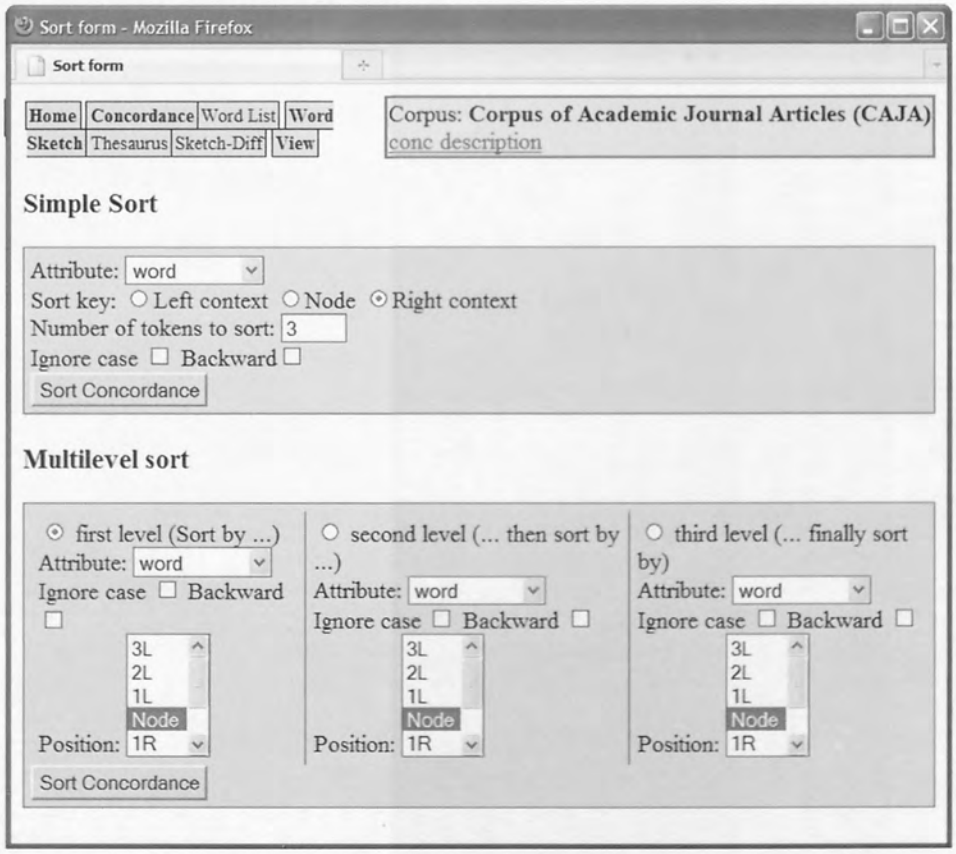
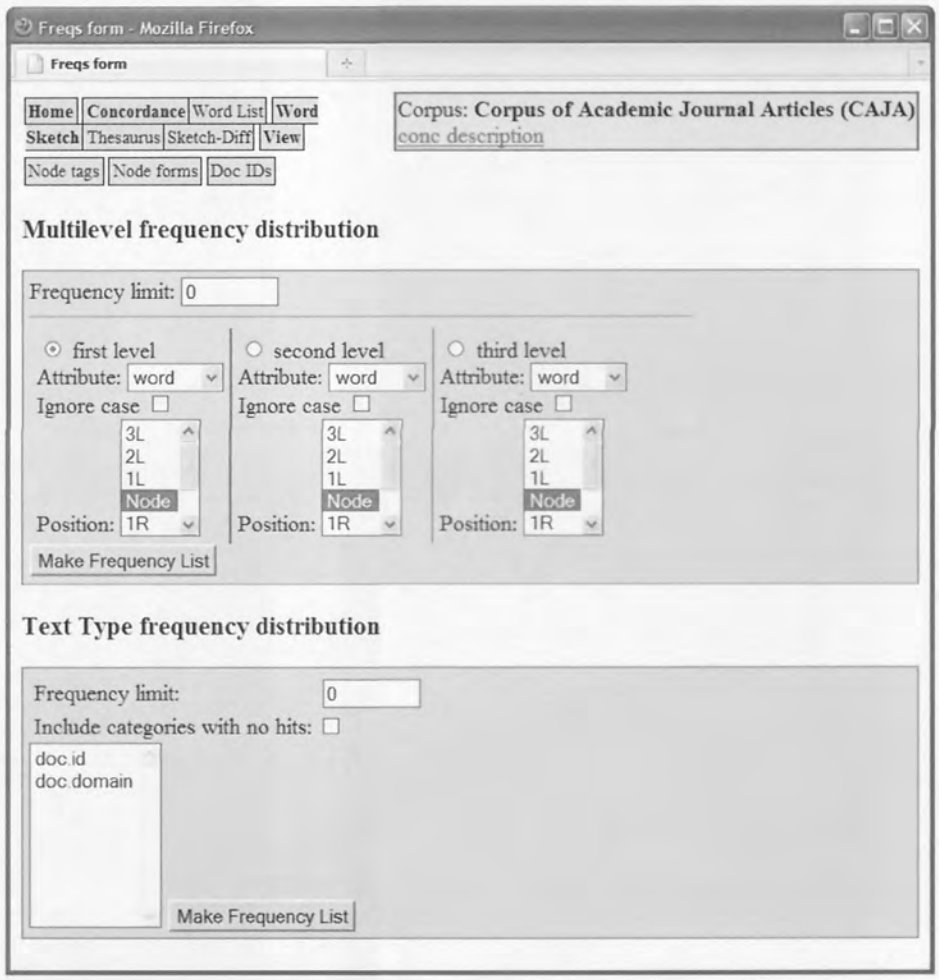


Figure 25. Sketch Engine: Complex sort options.



The *Sort* button offers two ways of sorting concordance lines: basic and complex sort. A basic sort is performed by moving a mouse cursor over the button, revealing a submenu with three options (sort by first word on the right, sort by node, and sort by first word on the left[51]). To access the complex sort options (Figure 25 above), the user needs to click on the *Sort* button, and then use either Simple Sort or Multilevel Sort (up to three levels)[52]. Concordance lines are sorted alphabetically (or in reverse alphabetical order if the 'Backward' option is selected) by the word in the sort position.

Figure 26. Sketch Engine: Frequency options.



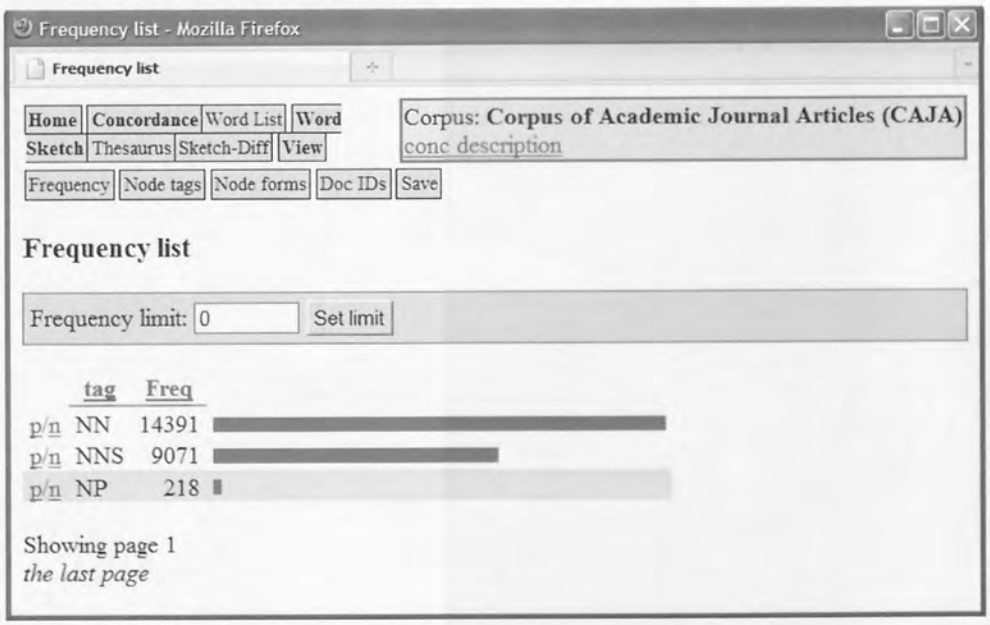
[51] The latest version of Sketch Engine also offers sorting by Text Types (by the name of the corpus document) – this function was not available at the time of the analysis.

[52] Multilevel Sort is performed by selecting the highest level of sorting one wants to use (e.g. third level for a three-level sort) – the settings need to be set for all levels included in the sort. A similar procedure is required when producing Multilevel frequency distribution (see Figure 26)

The *Frequency* button (Figure 18) can be used to produce frequency information about the searched item (node) and/or surrounding words (up to 3 words to the left and to the right), using the Multilevel frequency distribution function (Figure 26 above). Alternatively, the user can examine the distribution of the searched item across the subcorpora (Text Type frequency distribution).

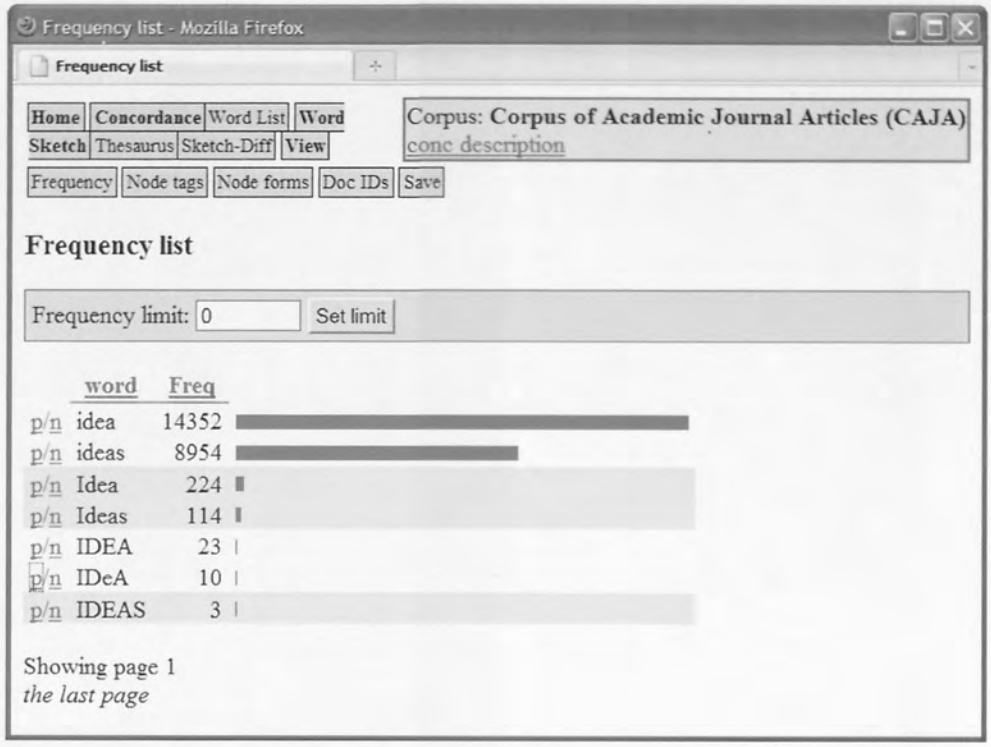
The second row of buttons in the left-hand corner contains three shortcuts, which are also available in the form of a submenu in the concordance window (by moving a mouse cursor over the *Frequency* button): *Node tags*, *Node forms* and *Doc IDs*. *Node tags* produces frequency information for the tags assigned to the node (Figure 27), *Node forms* produces frequency information for all the forms of the node (Figure 28) – which means that the function is most useful for lemma searches – and *Doc IDs* shows the distribution of the node across the subcorpora. Unless all the subcorpora are of exactly the same size, the domains need to be sorted by relative frequency of the node (normalized frequency) rather than raw frequency, as that ensures the comparability of subcorpora.

Figure 27. Sketch Engine: Frequency list by Node tags for *IDEA*.



Explanation of tags:
NN – noun, singular or mass
NNS – noun, plural
NP – proper noun, singular

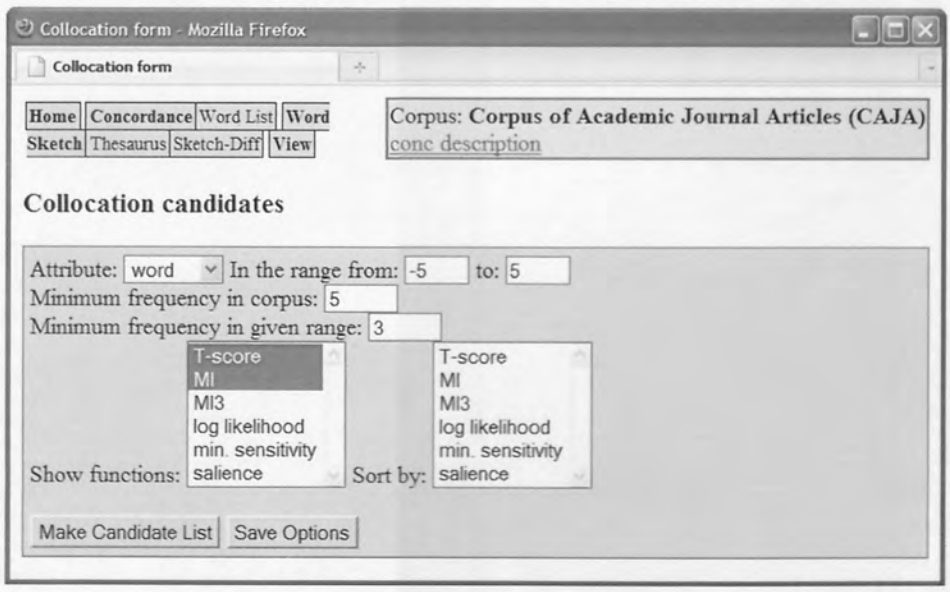
Figure 28. Sketch Engine: Frequency list by Node forms for *IDEA*.



Note: *IDeA* is an acronym for *Improvement and Development Agency*.

The *Collocation* button enables the user to generate lists of collocation candidates. The user can specify the range, minimum frequency of the item in the corpus, minimum frequency in the range, and the statistical method used for calculating collocation (Figure 29).

Figure 29. Sketch Engine: Collocation settings (default).



The statistical measures of collocation used in this research are T-score, Mutual Information (MI) score, and MI^3 score. These measures were selected because of my previous experience with them, and to have a range of different measures at my disposal when analysing collocation rather than relying on a single measure. Calculation formulae for the three measures are[53]:

$$\text{T-Score} \quad \frac{f_{AB} - \frac{f_A f_B}{N}}{\sqrt{f_{AB}}}$$

$$\text{MI-Score} \quad \log_2 \frac{f_{AB} N}{f_A f_B}$$

$$\text{MI}^3\text{-Score} \quad \log_2 \frac{f_{AB}^3 N}{f_A f_B}$$

N – total number of words in the corpus,
f_A – number of occurrences of the node in the corpus,
f_B – number of occurrences of the collocate in the corpus,
f_{AB} – number of co-occurrences of the collocate and the node (within a given span).

T-score measures "the confidence with which we can claim that there is some association" (McEnery et al., 2006:57), and MI score measures the strength of collocation. Collocations with high T-scores tend to include words with high frequency of occurrence in the corpus. On the other hand, collocations with high MI scores tend to be less frequent words. MI^3 score (Oakes, 1998:171-172), which uses a modified MI-score formula by cubing the frequency of co-occurrences of the collocate and the node, was developed to give more weight to frequent words.

The *Save* button[54] allows results to be saved in a TXT or XML file. This function is available for any output in Sketch Engine. The option to save the results in XML is especially useful for lexicographers, because most modern dictionary writing systems use this format (Atkins & Rundell, 2008).

[53] These forms of formulae have been copied from the document on the statistics used in Sketch Engine (Lexical Computing Ltd., 2007).
[54] This button should not be confused with the *Save Options* button which is used to save any changes made to the settings of functions (e.g. see Figure 29).

3.3.1.1.2 Word List

The *Word List* function allows the user to create word lists (for the entire corpus or a subcorpus). Word lists can be produced by attribute, for example by word or lemma (see Figure 22), and the minimum frequency can be set. Setting the minimum frequency is a very useful option, because it allows the omission of rare items.

The other two functions offered by *Word List*, but not used in this research, are keyword lists (only available when subcorpora are created), and lists of words that are most X (*Find X*). The *Find X* option can be used by lexicographers to add useful information to the entries, such as a *usually plural* note to the nouns which occur mainly in plural form, as exemplified by Kilgarriff and Rychly (2008). The feature, while useful, has not been used in this research because it was expected that the other analyses for sample entries would identify such trends in their use anyway.

3.3.1.2 Advanced functions

This section focuses on three more advanced features of Sketch Engine that are particularly useful for lexicographers: Good Dictionary Examples, Word Sketch, and Thesaurus (with the Sketch Difference function).

3.3.1.2.1 Good Dictionary Examples (GDEX)

GDEX is one of the latest additions to Sketch Engine, and is described in Kilgarriff et al. (2008). It can be used by lexicographers to obtain "good candidate dictionary examples" (Kilgarriff et al., 2008:431) from the corpus. The GDEX option is part of the *View Options* settings in the Concordance function of Sketch Engine. GDEX identifies the user-specified number of good examples (default is 100) which are then given priority in the concordance, i.e. they are displayed first. The GDEX heuristics, summarized from Kilgarriff et al. (2008), are shown in Table 16 below.

GDEX has already been successfully used in the preparation of two commercial dictionaries, which testifies to its potential. It was reported that a lot of time was saved in selecting example sentences (Kilgarriff et al., 2008). GDEX is a feature that lexicographers of the future are likely to use, so it is only appropriate to test it for the purposes of DOAE.

Table 16. Sketch Engine: GDEX heuristics.

- preferred sentence length: 10-25 words;
- sentences containing words which are not among the commonest 17,000 words are penalized, and additionally penalized for rare words;
- sentences containing pronouns and anaphors like *this, that, it,* or *one* are penalized (the argument is that these words often need further context to make sense);
- sentences with the target collocation in the main clause are preferred;
- whole sentences (beginning with a capital letter and ending with a punctuation – full stop, exclamation mark, or question mark) are preferred;
- sentences with third collocates are preferred (third collocates occur frequently with the node word and the primary collocate);
- sentences with the target collocation towards the end are preferred.

3.3.1.2.2 Word Sketch

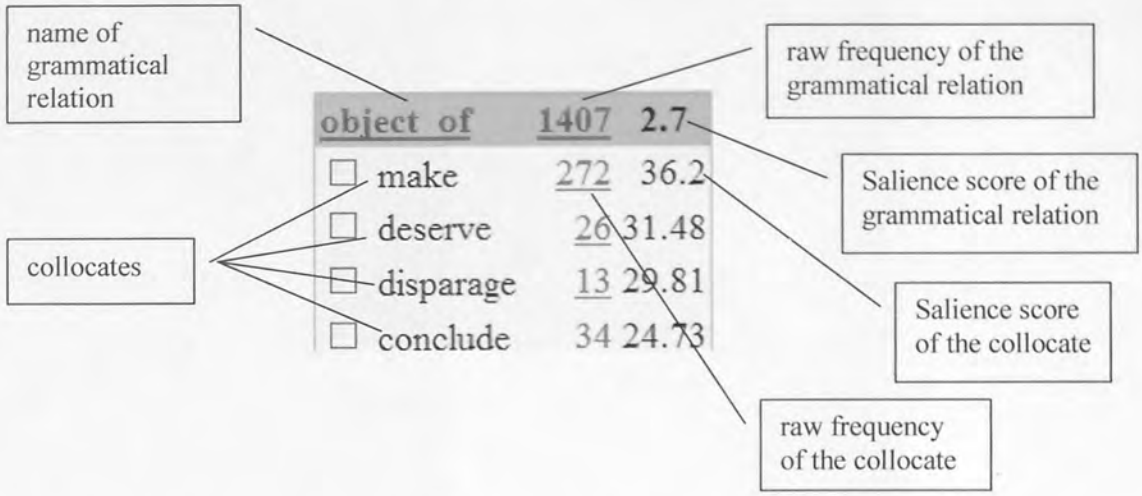
Kilgarriff et al. (2004:105) define word sketches as "one-page automatic, corpus-based summaries of a word's grammatical and collocational behaviour". Atkins and Rundell (2008) see word sketches as a type of lexical profiling, which has become the preferred starting point of lexicographers when analysing complex headwords.

Figure 99 and Figure 101 in Appendix 9 show examples of word sketches for two sample entries built for this Model, and it can be observed how collocates are grouped according to their grammatical relation. Each grammatical relation is defined in the Word Sketch definition file, which can be accessed by clicking on any relation name in the Word Sketch output (the user is taken to the part of the definition file where the selected relation is defined).

Word Sketch offers direct access to concordances of the collocates; the user needs to click on the number in the first column on the right of the collocate (e.g. 272 at *MAKE* in Figure 30) and that opens a new window (or tab, depending on the browser settings), displaying all the concordance lines where the node occurs with that particular collocate.

Word sketches are created for lemmas, so all forms of the headword are included when searching for its co-occurrence with the collocates. Similarly, collocates in a word sketch are listed as lemmas.

Figure 30. Sketch Engine: Word Sketch - explanation of the items in a grammatical relation of the noun *comment*



Collocates in a word sketch can be ranked by raw frequency or by salience. The information on raw frequency can be found in the first column on the right of the collocate, and salience score in the second column (see Figure 30). The statistical measure used for measuring salience is logDice (Lexical Computing Ltd., 2007:1-2):

$$14 + \log_2 \frac{2 \cdot ||w_1, R, w_2||}{||w_1, R, *|| + ||*, *, w_2||}$$

$||w_1,R,w_2||$ – number of occurrences of the triple (first word, grammatical relation, second word),
$||w_1,R,*||$ – number of occurrences of the first word in the grammatical relation with any second word
$||*,*,w_2||$ – number of occurrences of the second word in any grammatical relation with any first word

In the advanced settings of Word Sketch, the user can select the option of clustering collocates by level of similarity (see Figure 31 below). Collocates that occur in similar patterns are grouped together. Minimum level of similarity can be set (the default setting is 0.15). The formula for similarity score is the same as the one used by the Thesaurus function, which is described in more detail in 3.3.1.2.3.

The latest addition to Word Sketch is a feature called TickBox Lexicography, which gives the user the option to select collocates, select examples of the collocates[55], and copy them to the clipboard with XML tags (Steps 1-3 in Figure 100). This is very convenient for exporting examples directly into a dictionary writing system, in this case TshwaneLex (see 3.3.2).

[55] The default number of examples per collocate is 6.

Figure 31. Sketch Engine: clustered view of the collocates (the noun *comment*).

object_of	1407	2.7
make 272	632	36.2
follow 87 receive 35 offer 27 provide 52 require 20 include 22 give 25 need 9 consider 11 contain 9 take 13 present 8 find 8 do 9 have 16 use 9		
☐ deserve	26	31.48
disparage 13	20	29.81
solicit 7		
☐ conclude	34	24.73
post 13	24	24.54
submit 11		
hear 17	83	20.24
write 19 record 11 pass 9 read 8 publish 7 collect 6 draw 6		
add 27	54	18.65
relate 21 result 6		
echo 9	31	18.54
quote 10 appreciate 6 acknowledge 6		
attract 11	23	16.17
invite 7 direct 5		
elicit 8	14	14.23
stimulate 6		
see 20	25	10.08
think 5		
☐ code	6	9.8
		>>

3.3.1.2.3 Thesaurus and Sketch Difference

The Thesaurus function in Sketch Engine provides a list of "nearest neighbours" (Kilgarriff et al., 2004:113) for the word; the feature currently allows the building of lists for nouns, verbs, adjectives, adverbs, and prepositions. The similarity score between words is calculated with the following steps (Lexical Computing Ltd., 2007:2):

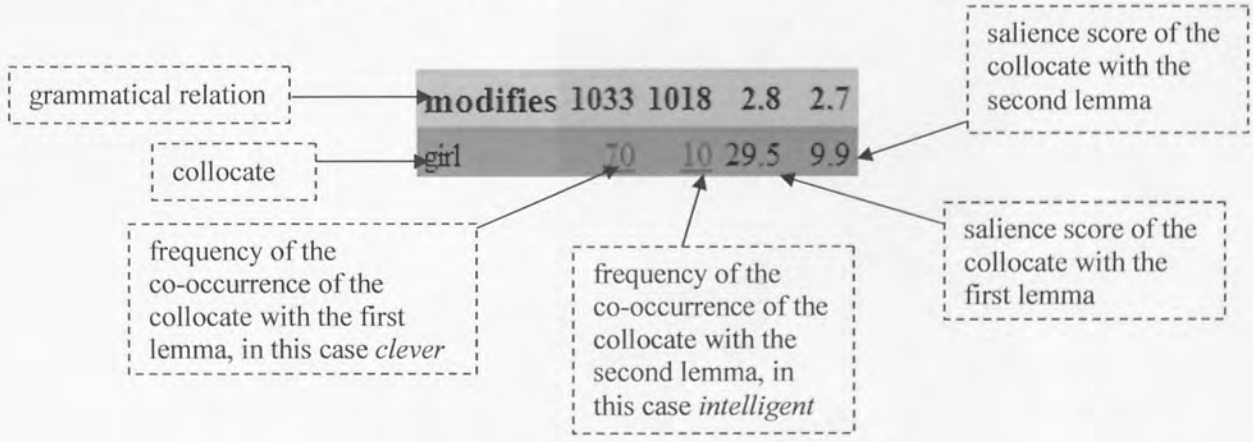
- compare *w1* (first lemma) and *w2*'s (second lemma) word sketches
- ignore contexts that supply no useful information (e. g. Association score (AS) < 0)
- find all the overlaps, e.g. where *w1* and *w2* "share a triple", so they share the same collocate in the same grammatical relation, as in *beer* and *wine* "sharing" (*drink*, OBJECT, *beer/wine*).

The equation[56] is:

$$Dist(w_1, w_2) = \frac{\sum_{(tup_i, tup_j) \in \{tup_{w_1} \cap tups_{w_2}\}} AS_i + AS_j - (AS_i - AS_j)^2/50}{\sum_{tup_i \in \{tup_{w_1} \cup tup_{w_2}\}} AS_i}$$

Thesaurus presents lexicographers with a list of potential (near-)synonyms (and, in some cases, antonyms) of the word. But this still leaves lexicographers with the task of describing the differences or similarities between the words. The Sketch Difference function offers help with this as "it shows those patterns and combinations that the two items have in common, and also those patterns and combinations that are more typical of, or unique to, one word rather than the other" (http://trac.sketchengine.co.uk/wiki/SkE/GettingStarted). Sketch Difference compares word sketches (see 3.3.1.2.2) for the two items, so it compares salience of the collocates and grammatical relations of the lemmas. Figure 33 shows the sketch difference for adjectives *clever* and *intelligent*, the first pair of near-synonyms explored with the Sketch Difference function (Kilgarriff et al., 2004). Figure 32 explains what the numbers next to each of the collocates mean.

Figure 32. Sketch Engine: explanation of information in the Sketch Difference output.



[56] The following note for the equation is provided: "The term (ASi−ASj)2/50 is subtracted in order to give less weight to shared triples, where the triple is far more salient with *w1* than *w2* or vice versa. We find that this contributes to more readily interpretable results, where words of similar frequency are more often identified as near neighbours of each other" (Lexical Computing Ltd., 2007:2).

Figure 33. Sketch Engine: the Sketch Difference for *clever* and *intelligent*.

clever/intelligent **preloaded/bnc freq = 2237/1844**

Common patterns

clever	6.0	4.0	2.0	0	-2.0	-4.0	-6.0	intelligent

adj_subject	128	119	8.5	7.9
man	3	3	4.6	4.7

adj_comp_of	37	20	6.3	3.4
look	6	4	13.4	11.6
become	3	3	8.2	9.4

and/or	396	742	1.9	3.6
witty	5	10	19.2	25.9
clever	10	4	23.5	12.2
charming	3	10	11.7	22.2
beautiful	7	17	14.7	21.6
bright	4	13	10.7	19.6
amusing	5	4	18.1	14.3
young	11	18	14.3	16.4
ambitious	5	4	16.1	12.5
quick	7	6	16.0	12.8
sexy	3	3	13.3	12.1
funny	4	3	12.4	8.9
interesting	3	4	8.0	8.5
strong	4	3	8.4	5.4
serious	3	4	7.4	7.8
good	5	8	6.3	7.3
new	4	5	3.8	3.2
other	4	3	3.5	1.1

modifies	1033	1018	2.8	2.7
girl	70	10	29.5	9.9
boy	62	7	29.1	8.1
man	55	66	18.1	20.1
people	25	58	9.9	17.5
woman	19	33	11.4	16.4
person	4	16	3.8	12.5
use	16	15	11.6	11.1
system	6	24	2.8	11.2
eye	5	14	3.7	9.6
thing	16	5	9.5	3.0
face	3	12	2.3	9.4
one	7	3	9.3	4.5
software	4	5	6.2	7.5
dog	4	5	6.1	7.4
way	15	11	7.3	5.5
child	9	6	5.9	3.9
player	4	3	5.3	4.0
head	5	3	3.7	1.9
control	3	3	2.7	2.8

modifier	666	474	1.5	1.0
very	272	74	46.6	29.8
too	76	20	33.9	19.4
more	21	90	14.0	32.2
so	69	16	29.5	15.1
as	30	12	22.8	14.9
less	4	21	7.2	22.7
extremely	14	5	21.0	12.1
really	18	8	18.1	12.0
quite	13	15	15.2	17.9
incredibly	5	3	17.1	12.9
also	3	3	2.7	3.4

pp_in-p	12	8	0.3	0.2
way	3	3	9.3	10.2

"clever" only patterns

adj_subject	128	8.5
boxing	3	13.7
George	3	9.1
bit	3	8.3

adj_comp_of	37	6.3
box	3	19.2
get	13	17.8
feel	3	8.9
seem	3	8.8

np_adj_comp_of	14	6.1
think	3	9.6

modifies	1033	2.8
dick	12	28.8
trick	18	25.3
clog	8	24.9
folly	10	20.9
chap	11	19.8
ploy	7	19.8
idea	30	18.3
pun	4	16.0
swine	4	15.5
wheeze	3	14.6
concealment	3	14.1
lass	4	14.1

and/or	396	1.9
cunning	5	19.9
little	19	19.3
brave	6	18.0
subtle	6	17.5
devious	3	14.6
intelligent	4	13.4
bloody	5	13.1
cool	4	12.4
efficient	4	11.7
dangerous	4	11.1
rich	4	10.6
pretty	3	10.5

modifier	666	1.5
fiendishly	5	25.3
that	11	21.5
extraordinarily	6	19.7
rather	13	17.9
pretty	8	16.8
particularly	11	16.5
fucking	5	16.5
awfully	4	16.3
jolly	3	13.6
ever so	3	13.2
terribly	3	11.2
bloody	3	10.8

pp_for-p	12	0.6
good	9	33.6

"intelligent" only patterns

adj_subject	119	7.9
human	3	12.1
child	5	8.7
woman	4	7.6

np_adj_comp_of	12	5.2
make	5	11.2

and/or	742	3.6
sensitive	27	30.4
articulate	11	28.9
thoughtful	8	22.1
lively	10	21.6
educated	7	19.9
well-educated	4	18.7
adaptive	5	18.3
attractive	9	16.3
good-looking	4	15.7
honest	6	15.6
human	13	15.3
informative	4	14.8

modifies	1018	2.7
being	33	27.9
hub	9	22.8
robotic	4	18.7
guess	7	18.5
conversation	11	15.4
read	3	14.0
life	26	12.7
robot	4	12.6
controller	5	12.4
subsystem	3	11.7
creature	6	11.2
electronics	4	10.8

modifier	474	1.0
highly	65	41.6
most	40	25.4
reasonably	6	16.6
obviously	7	16.5
fairly	5	12.5
seemingly	3	12.2
otherwise	3	9.9
equally	3	9.0

As shown in Figure 33 above, Sketch Difference makes effective use of colours to indicate similarities and differences in salience scores of collocates shared by the two lemmas. Light yellow colour indicates collocates with the same or similar salience score for both lemmas (in 'Common patterns), or collocates found with only one of the lemmas (in '...only patterns'). Green and red shades are used to label shared collocates that have a higher salience score when occurring with a particular lemma.

As Word Sketch, Sketch Difference offers the user direct access to concordance lines in which the collocate co-occurs with each of the two lemmas. Thesaurus and Sketch Difference are also directly linked – clicking on a lemma in the list in Thesaurus automatically opens Sketch Difference, which shows the collocational comparison between the lemma of the Thesaurus search and the lemma from the list.

3.3.2 TshwaneLex[57]

TshwaneLex is a dictionary-writing system developed by TshwaneDJe Human Language Technology. TshwaneLex is a user-friendly and highly customizable system that does not require a high level of computer literacy (Joffe & de Schryver, 2004). It can be used in the production of various types of dictionaries (e.g. monolingual, bilingual, thesauri), and various dictionary formats (paper, CD-ROM, web). It supports over 6000 languages. The list of publishers and institutions that have used TshwaneLex includes Oxford University Press, Pearson/Longman, Macmillan, Royal Spanish Academy, and the Czech Language Institute.

Some features of TshwaneLex are especially worth pointing out, as they were the deciding factor for choosing this particular software for compiling the Model for DOAE:

1. Customisable Document Type Definition (DTD[58]) via dictionary grammar editor (see Figure 34), without the need for high level programming skills.

This feature is highly valuable for the dictionary-making process because it provides lexicographers with the opportunity to influence the way that information is stored in the database and presented to the dictionary user, without requiring an extensive background in

[57] The information about TshwaneLex has been obtained from the TshwaneDJe company website (http://tshwanedje.com/), and articles by Joffe and de Schryver (2004), and Joffe, MacLeod and de Schryver (2008).
[58] Document Type Definition or DTD defines element types and attribute lists for an XML document. A dictionary DTD "defines the constituent elements of the dictionary and the allowable sequences in which they can occur" (Atkins & Rundell, 2008:116). One example of a definition in a DTD can be that examples may occur only within a sense, and must not precede the definition for that sense.

programming. The Tshwanelex dictionary grammar editor offers the user a much more user-friendly overview of the DTD and the relations between elements and attributes than the standard XML layout (see Figure 34 and Figure 35).

Figure 34. TshwaneLex: Dictionary grammar editor (DTD structure of element "Subentry").

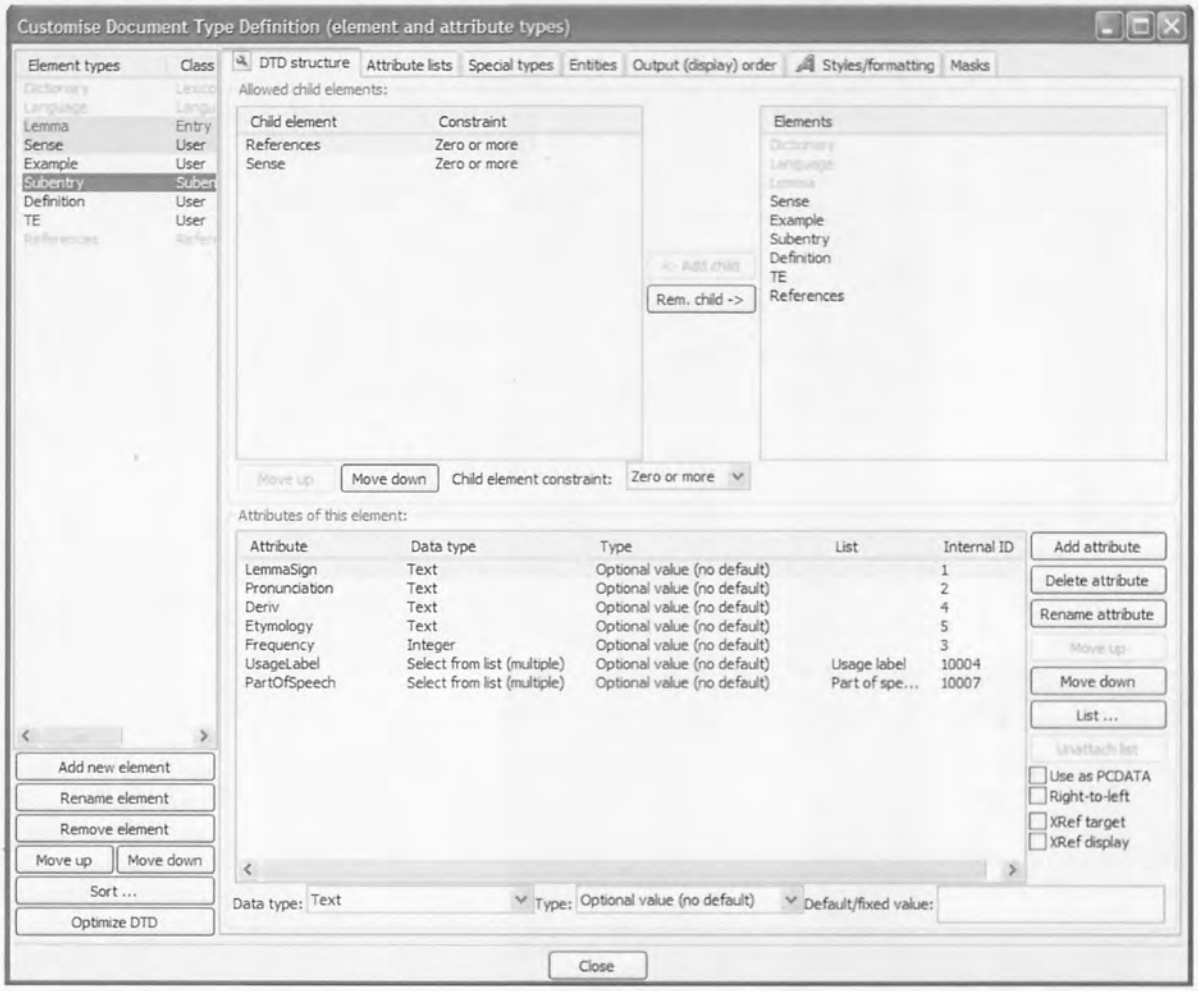


Figure 35. TshwaneLex: XML version of DTD (element "Subentry").

```
<order_element elementname="Subentry">
  <order_item type="attribute" name="LemmaSign" visible="1" />
  <order_item type="attribute" name="Pronunciation" visible="1" />
  <order_item type="attribute" name="Deriv" visible="1" />
  <order_item type="attribute" name="PartOfSpeech" visible="1" />
  <order_item type="attribute" name="UsageLabel" visible="1" />
  <order_item type="element" name="References" visible="1" />
  <order_item type="element" name="Sense" visible="1" />
  <order_item type="attribute" name="Etymology" visible="1" />
  <order_item type="attribute" name="Frequency" visible="1" />
</order_element>
```

Adding elements and attributes is also easy, as the dictionary grammar editor assists the user at almost every step – for example, in the list of elements in the top right window, the elements that cannot be added as child elements are greyed out (in the case of 'Subentry' in Figure 34, such elements are 'Dictionary', 'Language' and 'Lemma'). In addition, the users can use their own naming system for elements and attributes (but the names of elements and attributes are not allowed to have spaces in them).

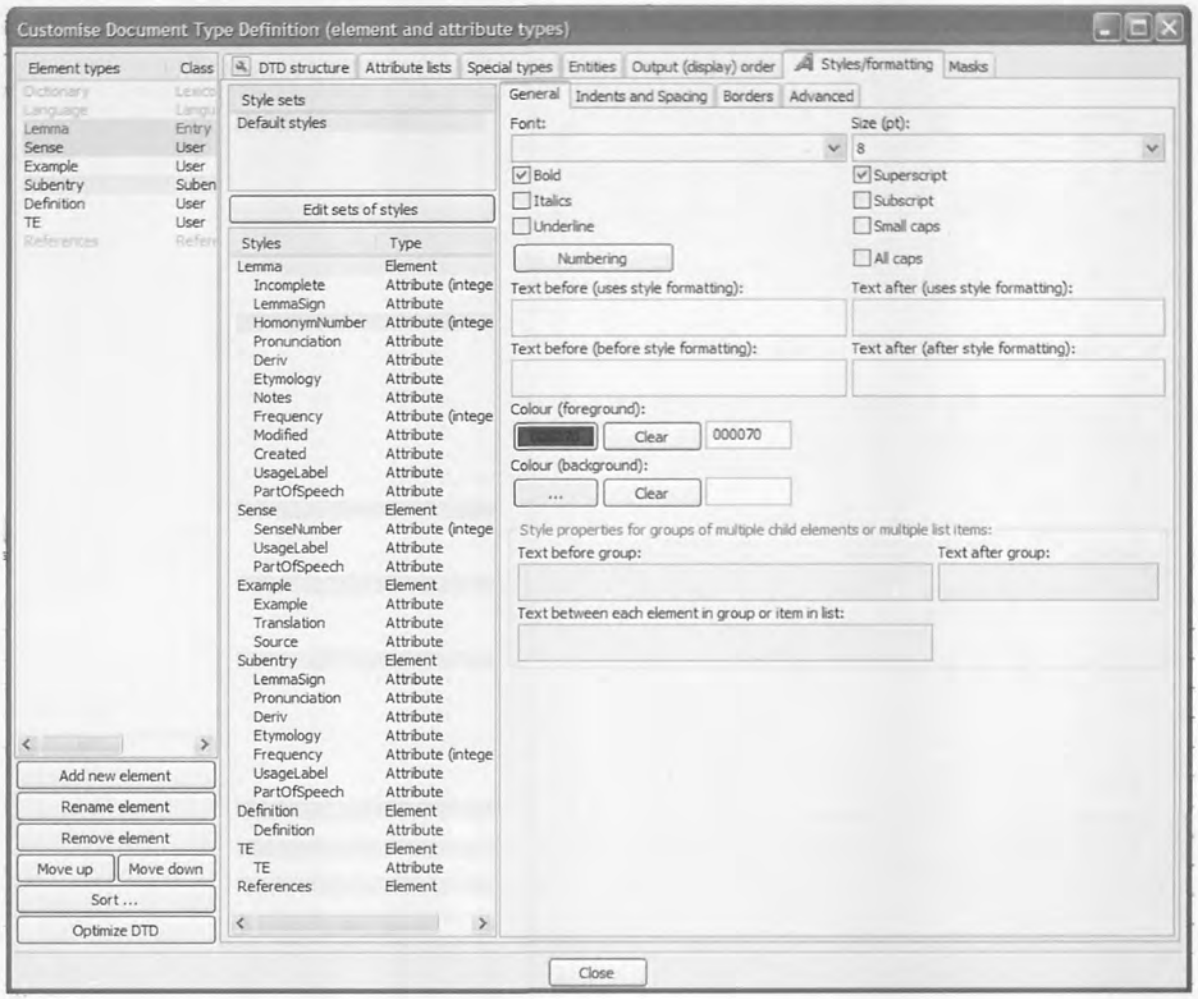
DTD can be modified while the dictionary is being compiled. Any changes to the DTD are immediately implemented in any entries that are already in the database. This can be extremely useful because lexicographers can identify a potential deficiency in the DTD while compiling dictionary entries, and suggest improvements which can be implemented immediately if considered appropriate. Because major changes to the DTD may result in the loss of data, it is wise to make a back-up of the database before making any changes to the DTD (TshwaneLex provides a pop-up window that advises the user to do this).

2. High level of automaticity.

Many processes in TshwaneLex are automatic, for example updating the database after changes to the DTD or styles have been made, numbering of senses, subsenses, and homonyms, updating cross-references, and checking for errors in the database. This greatly reduces the time for completing a dictionary project, and reduces errors in the dictionary.

In addition to having a customizable DTD, TshwaneLex also offers the option to customise styles (e.g. font, colour, layout, etc.) for every field in the dictionary (see Figure 36 below). Any changes made are immediately reflected in the entry preview. There is also the option to change the order in which parts of entry are displayed. Lexicographers can therefore experiment with different styles, formatting and layout to determine the best way of presenting dictionary entries.

Figure 36. TshwaneLex: Styles/Formatting tab.

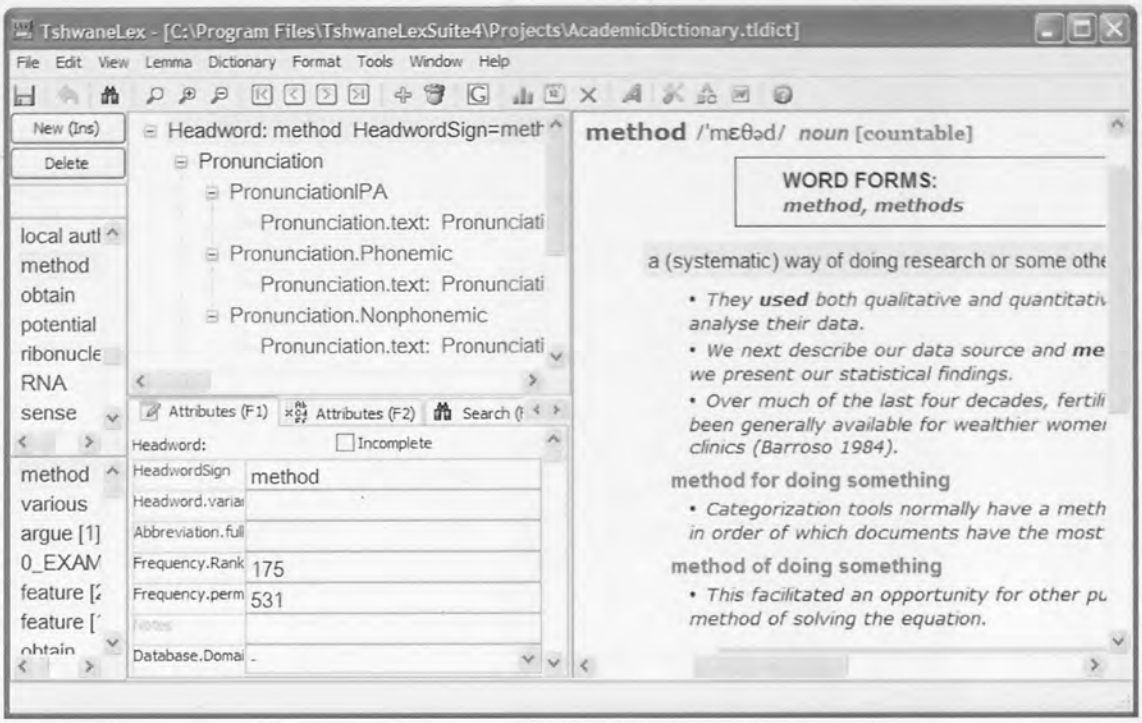


3. Entry preview.

TshwaneLex offers the lexicographer the option to automatically preview the dictionary entry that is being worked on (see Figure 37 below). In this way, the lexicographer sees how the dictionary user will see the entry, and can make any changes while working on the entry. There is also the option to make certain parts of the entry temporarily invisible, which is very useful if the lexicographer wants to focus on other parts.

Another benefit of this feature is that the dictionary entry can be directly compared to the corresponding entries in other dictionaries not only in terms of content, but also in terms of layout, style, and presentation. This was useful for comparing the sample entries for DOAE with entries in other dictionaries.

Figure 37. TshwaneLex: Entry preview (left window).



4. Multiple dictionaries from a single database.

Another very useful option in TshwaneLex is that several different sets of styles can be created for the same database, with each of the styles having a specific output (or preview). In this way, each style can represent a different dictionary format (e.g. paper or CD-ROM) or even a different type of dictionary user (e.g. beginner or advanced user).

5. Allows adding of sounds and images.

Sounds and images can be added to the database. This option enables the inclusion of features such as audio pronunciation and illustrations in the entries. It also offers the option to optimise the effectiveness of certain information – for example, the information on the frequency distribution of the headword across subcorpora is more effective if presented as, or accompanied by, an image.

6. Allows various data formats to be imported.

TshwaneLex allows the data to be imported as a wordlist, as Comma Separated Values, or in XML format. This enabled the import of collocates and their examples in XML format with from the TickBox Lexicography function in Sketch Engine (see 3.3.1.2.2). To make this

possible, an XML template had to be created for the TickBox Lexicography output which matches the DTD structure of DOAE in TshwaneLex (Figure 38).

Figure 38. Model for DOAE: XML template for the TickBox Lexicography output.

```
<Dictionary>
<Language>
<Headword HeadwordSign="">
<gramrel grname="">
<collocation collo="">
  <Example Example.number="" Database.DomainLabel="" Example=""
   Source="" />
</collocation>
</gramrel>
</Headword>
</Language>
</Dictionary>
```

7. Allows exporting of data into various formats.

The information recorded in the TshwaneLex database can be exported into formats such as DOC, RTF, HTML, and XML. In addition, TshwaneLex offers the user the option of exporting parts of an entry, as well as entire entries.

Other useful features of TshwaneLex include regular updates to the software system being provided on the website, compatibility with both Windows and Mac platforms, and the option to integrate the TshwaneLex electronic dictionary into Microsoft Word (i.e. the users can access the dictionary from Microsoft Word while writing texts). For publishers, the network and multi-user support are particularly valuable, as are the management tools that can be used to assign tasks to lexicographers, and monitor progress.

3.4 Data analysis – a corpus-driven approach

The Model for DOAE proposed in this study has drawn on information provided by corpus data, i.e. CAJA. Two distinctly different approaches to using corpus data, the corpus-based approach and the corpus driven-approach, have been identified by linguists. This section discusses which of the two approaches has been used in this research, and the rationale for this decision.

According to Tognini-Bonelli (2001), the difference between the corpus-based approach and the corpus-driven approach depends on the level of importance that is attributed to the corpus. In the corpus-based approach,

> "...corpus evidence is brought in as an extra bonus rather than as a determining factor with respect to **the analysis**, which **is still carried out according to pre-existing categories**..." (ibid.:66)

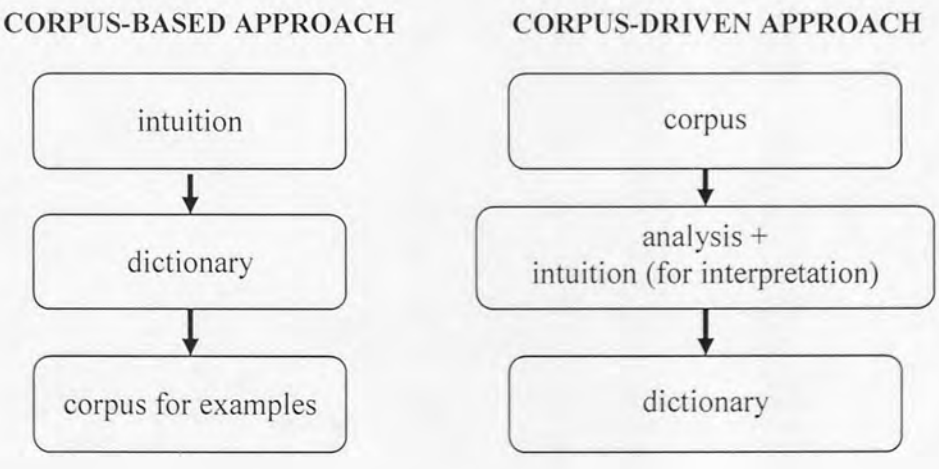
In the corpus-driven approach, on the other hand,

> "...the commitment of the linguist is to the integrity of the data as a whole, and descriptions aim to be comprehensive with respect to corpus evidence. The theoretical statements are fully consistent with, and reflect directly, the evidence provided by the corpus. Indeed, **many of the statements** are of a kind that **are not usually accessible by any other means than the inspection of corpus evidence**." (ibid.:84)

The text in bold (my emphasis) is particularly significant for selecting the approach most suitable for the description of academic English. In the definition of the corpus-based approach, 'pre-existing categories' can be interpreted as previously established word meanings and phrases. But such pre-existing categories are missing in academic English, as there is no comprehensive description of academic English available. Hence, the corpus-based approach is not suitable for DOAE.

In lexicography, the difference between the corpus-based approach and the corpus-driven approach is mainly in the different roles of corpus and intuition (Figure 39). In the corpus-based approach, the corpus is only used to support intuition, i.e. to provide examples. On the other hand, the corpus-driven approach entails putting the corpus at the forefront and using intuition only as an aid in interpreting the corpus data. Considering that it is questionable whether there is such a thing as a native speaker of academic English (see 2.1.4), it is even less likely that there is such a thing as native-speaker intuition in relation to academic English. Thus, DOAE simply must be driven by corpus data.

Figure 39. Corpus-based approach vs. corpus-driven approach in lexicography.



The corpus-driven approach was first used in the Cobuild project (see Sinclair, 1987 for a detailed account of the project) that produced the Collins Cobuild English Language Dictionary (1987). Since then, many types of dictionary, such as EFL dictionaries and some NS dictionaries (e.g. NODE), have relied extensively on corpus data; however, none of the dictionaries has matched the extent of 'trust' in corpus data demonstrated by the authors of the Cobuild dictionary.

So what does the corpus-driven approach entail for a dictionary-maker? As Krishnamurthy (2008) states,

> "A corpus-driven approach involves a bottom-up methodology, beginning by selecting unedited examples from the corpus, identifying their shared and individual features, and only then grouping them for the purpose of lexicographic presentation." (ibid.:231)
>
> "...the new entries, sense divisions, and definitions are fully consistent with, and reflect directly, the evidence of the corpus; examples are used verbatim; recurrent patterns form the basis for lexicographic categories; and the absence of an entry, or a pattern in an entry, is a meaningful lexical statement." (ibid.:240)

A corpus-driven dictionary therefore uses a corpus not only as a source of information for language description, but also as a resource for dictionary content.

Collocation is a notion that is closely connected with the corpus-driven approach, and pervades entries in a corpus-driven dictionary. Collocation underpins the idea that phraseology, or combinations of words, is the best starting point for the description of meaning (Sinclair, 2004b). Word meanings are restricted in terms of collocational preferences, and Sinclair (1991) describes this as the idiom principle[59]:

> "The principle of idiom is that a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments." (ibid.:110)

Not only do collocations help describe word meanings, they also play an important role in distinguishing between different meanings of the word. Hoey (2005:82) puts forward, and confirms, the following three hypotheses:

---

[59] Opposed to the idiom principle is the open-choice principle, or a 'slot-and-filler model (Sinclair, 1991:109), which is at work when a word has a fixed meaning. Sinclair (2004:29) argues that technical terms are used when applying the open-choice principle, in other words the meaning of technical terms is less dependent on their combinations with other words.

"1. Where it can be shown that a common sense[60] of a polysemous word is primed to favour certain collocations, semantic associations and/or colligations[61], the rarer sense of that word will be primed to avoid those collocations, semantic associations and colligations. The more common use of the word will make use of the collocations, semantic associations and colligations of the rarer word but, proportionally, less frequently.
2. Where two senses of a word are approximately as common as each other, they will both avoid each other's collocations, semantic associations and/or colligations.
3. Where either (1) or (2) do not apply, the effect will be humour, ambiguity (momentary or permanent), or a new meaning combining the two senses."

PDEV also uses a corpus-driven approach (see 3.2.2.3). The CPA approach used by PDEV combines the collocational preferences of the word into patterns, which are then mapped onto meanings.

Adopting the corpus-driven approach in designing the Model for DOAE has ensured that the dictionary represents the vocabulary that *is* used in academic English, as well as *how* it is used. More importantly, because the focus on collocation (driven by the corpus methodology) emphasizes combinations and patterning of words, the words are not explained in isolation, but in the context in which they are normally found. This increases the decoding and encoding value of the dictionary.

3.5 Final remarks on the methodology

This chapter has presented the data and software that have been used in this thesis. The main data are the results of the survey into students' dictionary use, and the corpus of academic journal articles (CAJA). Other data, such as other corpora and dictionaries, have been consulted mainly for improving and evaluating the dictionary. The Sketch Engine corpus system and TshwaneLex dictionary-writing software have been used for the analysis of the corpus data and the compilation of the sample dictionary entries respectively. Using the corpus-driven approach in compiling the sample dictionary entries affects not only how the data is analysed, but also how it is recorded in the dictionary-writing software and ultimately presented to the user.

[60] Hoey does not appear to distinguish between 'meaning' and 'sense' of the word. He does not use 'sense' as a lexicographic term, namely as a microstructural element of the dictionary entry that explains a meaning of the word.
[61] Hoey (2005:43) defines colligation as: "1. the grammatical company a word or word sequence keeps (or avoids keeping) either within its own group or at a higher rank; 2. the grammatical functions preferred or avoided by the group in which the word or word sequence participates; 3. the place in a sequence that a word or word sequence prefers (or avoids)."

4. DOAE: THE USER PROFILE

The first step in designing any dictionary is to create a user profile. A user profile provides dictionary-makers with important information about the potential users of the dictionary and their needs. The user profile is especially important if a completely new dictionary is being designed, as is the case in this thesis. But even if an existing dictionary is being updated, the makers of the dictionary should consider reviewing the last user profile and possibly updating it, bearing in mind that times change and so do users and their needs/skills.

4.1 Who are the target users of DOAE?

The typical users of DOAE are expected to be the users of academic English, namely university students. Students will be coming from different language backgrounds, and will be native or non-native speakers of English. They may study different subjects, or a combination of subjects, at different levels. Nonetheless, the dictionary is expected to be also useful to their teachers, university lecturers, researchers, pre-sessional tutors, and support staff.

One problem that has been repeatedly mentioned in this thesis is that very little is known about the dictionary use of university students, especially NSs. In addition, the research that has been conducted (e.g. Hartmann, 1999; Nesi & Haill, 2002) needs updating and/or lacks some important information, such as dictionary format preference. An online survey was conducted in an attempt to fill this gap.

4.1.1 Main survey

4.1.1.1 Students

The main survey was completed by 620 students[62] – 449 native speakers of English and 171 non-native speakers of English. The students were from 66 different countries of origin[63], with the majority of students (403) coming from the UK. Based on the country of origin information, it was assumed that they were NS and spoke British English.

[62] N=620 is used in all the tables and figures related to survey results, unless otherwise stated.

[63] The list of countries includes United Kingdom (207 students), but also England (185 students), Wales (6 students), Scotland (2 students), and Northern Ireland (3 students), which are of course part of the first one (the United Kingdom). This is because they were included in the list of countries offered in the BOS menu. All these students were considered to be of UK origin for the purposes of this thesis.

The NNS students came from 56 different language backgrounds (Table 103 in Appendix 5). The majority of students (77%) reported to have been learning English for more than 10 years (Table 17). The comparison of students' age, age when they started learning English, and reported period of learning English reveals that hardly any of the students have had gaps in learning English before starting their university studies.

Table 17. Main survey: NNS students by years of learning English (n=171).

Years of learning English	Percentage of students
1-5 years	7%
6-10 years	16%
11-15 years	36%
16-20 years	23%
more than 20 years	18%

The gender distribution was almost the same among NSs and NNSs (Table 18); nearly two thirds of the students were female, and one third were male. The majority of students (81%) were between 18 and 24 years old[64] (Figure 96 in Appendix 5).

Table 18. Main survey: Students by native-speaker status and gender.

	NS students	NNS students
Female	65%	65%
Male	34%	33%
No Answer	1%	2%

613 students were Aston University students, while 7 were studying at other UK universities[65]. The Aston students were from all four Aston Schools. As shown in Table 19, the highest percentages of students were from the School of Life and Health Sciences (LHS) and Aston Business School (ABS). The students on the Combined Honours courses were studying two subjects, either offered by the same School or two different Schools.

[64] One student, who put 9 years old in their questionnaire, was excluded from the age distribution figure as the information was probably a typo and would skew the data.

[65] The universities were Bath Spa University, The Arts Institute at Bournemouth, University of the West of England in Bristol, Coventry University, The University of Exeter, the University of Gloucester, and The Open University.

Table 19. Main survey: Distribution of Aston students by School.

	Students (n=613)
School of Life and Health Sciences	30%
Aston Business School	29%
School of Engineering and Applied Science	17%
School of Languages and Social Sciences	14%
Combined Honours	9%
no answer	1%

The distributions of NS students and NNS students by Aston Schools (Table 20) are very similar to the overall distribution of students by School. Only the percentages of NNS students from ABS (higher) and LHS (lower) are noticeably different from the corresponding percentages of the total students respectively.

Table 20. Main survey: Distribution of Aston students by School and native-speaker status.

	NS students (n=444)	NNS students (n=169)
School of Life and Health Sciences	34%	20%
Aston Business School	26%	37%
School of Engineering and Applied Science	17%	19%
School of Languages and Social Sciences	13%	16%
Combined Honours	9%	6%
no answer	1%	2%

The students were enrolled on a variety of courses and were studying a wide range of subjects. Table 21 shows the courses of study reported by the students, according to the School[66]. Courses that were attended by 10% or more of the students from the School are shown in bold.

[66] The courses with a single subject have not been included.

Table 21. Main survey: Students' courses of study by Aston School.

	course of study (number of students)
School of Life and Health Sciences	**Pharmacy (49), Human Psychology (43), Optometry (29), Psychology (23)*,** Biomedical Science (6), Audiology (5)**, Biology (4), Human Biology (4), Applied and Human Biology (3), Health Psychology (3)
Aston Business School	**Business and Management (39), Marketing (19), Business Administration (18),** Human Resource Management and Business (13), International Business and Management (12), Business (9), Accounting for Management (7), International Business (7), Logistics (7), Business Computing and IT (5), Human Resource Management (5)***, International Business and Economics (5), Management (4), Management and Strategy (4), Marketing Management (3), Work Psychology and Business (3), Finance (2)
School of Engineering and Applied Science	**Computing Science (21), Mechanical Engineering (10)**, Chemical Engineering (6), Engineering (5), Computing for Business (4), Electronic Engineering (4), Mathematics with Computing (4), Construction Management (3), Industrial Product Design (3), Pattern Analysis and Neural Networks (3), Product Design and Management (3), Biological Chemistry (2), Engineering Management (2), Geographic Information Systems (2), Technology and Enterprise Management (2), Telecommunications Technology (2), Transport Management (2)
School of Languages and Social Sciences	**Teaching English to Speakers of Other Languages - TESOL (18), Translation Studies (16),** European Studies and Modern Language (4), TESOL and Translation Studies (4), French (3), Sociology with Politics (3), Applied Linguistics (2), Education (2), French and Politics (3), French and Spanish (2), Politics and International Relations (2), Social Sciences (2)
Combined Honours	**International Business and Modern Languages (19), Business Administration and Psychology (6),** Biology and Psychology (4), Business Administration and International Relations (4), Psychology and Sociology (4), Business Administration with Public Policy and Management (3), English Language and Sociology (3), Business Administration and Sociology (2), English Language and Psychology (2), Human Psychology and Sociology (2), Mathematics with Economics (2)

* - includes a student of Coventry University
** - includes an Aston staff member, studying at the Open University
*** - includes an Aston staff member, studying at the University of Gloucester

The students ranged from undergraduates (year 1 to year 4) to Masters and PhD students (Table 22). The distribution of students across the levels of study was fairly even. There were considerably more undergraduate (Years 1 to 4) than postgraduate students (Masters and PhD), 75% and 25% respectively.

Table 22. Main survey: Students by native-speaker status and level of study.

	NS students (n=449)	NNS students (n=171)	all students (n=620)
undergraduate (year 1)	28%	22%	26%
undergraduate (year 2)	20%	13%	18%
undergraduate (year 3)	19%	13%	18%
undergraduate (year 4)	15%	8%	13%
postgraduate (MA)	10%	35%	17%
postgraduate (PhD)	8%	9%	8%

The students in the survey were representative of both the Aston University student population and the UK student population for the academic year 2007/08[67]. The students were very representative for age group (Table 104 in Appendix 5) and level of study (Table 107), two categories with an obvious correlation.

The student group was also quite representative in terms of country of domicile (Table 106), but was more representative of the Aston University student population than of the UK student population. This pattern of representativeness can be also claimed for the NNS students in the survey[68], assuming that a majority of non-UK students are also NNSs of English.

In terms of gender, the students in the survey are more representative of the UK student population than of the Aston University student population (Table 105). It is noteworthy that the gender distribution of the Aston University student population is not very representative of UK student population.

There is some similarity between the distribution across Aston Schools of both the students in the main survey and of the Aston University student population (Table 108), although the distribution of the students in the main survey is more even.

4.1.1.2 Dictionary format

Many students reported the preference of either online or paper dictionary format, whereas pocket electronic format and CD-ROM format were preferred by only 5% and 4% of the students respectively (see Table 23).

---

[67] Data for the academic year 2007/08 was used as the survey was conducted in that period. Information was obtained from Aston University Planning Office (2008a; 2008b; 2008c; 2008d) and HESA (2009).

[68] The actual percentage of NNSs in the survey was 28%, but as no comparable data was available for Aston University and UK Higher Education Institutions, the country of domicile data was used for comparison.

Table 23. Main survey: Preference of dictionary format.

online	44%
paper	40%
pocket electronic	5%
CD-ROM	4%
no preference	7%

The reasons that students gave for preferring individual dictionary formats are similar to the advantages of the formats described by de Schryver (2003; see also 2.3.1). The most frequently mentioned reasons for preferring individual format are summarized below:

a) Online format. Reported advantages include quick searches, ease of use, more content (e.g. terminology, examples), access to other resources (e.g. thesaurus), user-friendliness (copy and paste option, suggesting words when the word is misspelled), more up-to-date content, and free access. A large number of students mentioned convenience of using online dictionaries for academic work, much of which is done on a computer. Also, online format was often compared with paper format, the students pointing out that paper dictionaries are awkward to carry around; however, many believed that paper dictionaries are usually of better quality. Several students reported using the 'define:' command in Google. Two students pointed out that online dictionaries are "...better for the environment".

b) Paper format. Reported to be easy to use compared with other formats, mainly because no technical knowledge is required. Many students preferred the format because they were used to it, and they liked browsing through the book and learning other words than just the ones they were initially interested in. Furthermore, several students believed paper dictionaries to be of better quality ("...more likely to be academically sound..."), especially compared to online dictionaries. In fact, the advantages of paper format were often explained by comparing it with dictionaries on computers; for example, it was said that it is easier to read from a book than from a computer screen, and that unlike a computer, a paper dictionary cannot crash. Nevertheless, some students who preferred paper format admitted to using electronic dictionaries when a paper dictionary was not to hand.

c) Pocket electronic format. The students liked this format because of its ease of use (easy to carry around and can be used anywhere), quick searches, great number of entries, and because of offering access to more than one dictionary and to other functions, such as thesaurus and currency converter.

d) CD-ROM format. This format was preferred because it is quick, or quicker than some other formats, easy to use, has more content (e.g. more entries, recorded pronunciation, visual material), and is easy to access (mentioned by students who often carry their laptop around).

Many of the students that had no preference for any of the formats explained that they use both online format and paper format with equal frequency, but for different purposes; paper format is used when reading newspapers and books, whereas online format is used when doing academic work (academic writing in particular).

The results on the frequency of format use (Figure 97 in Appendix 5) were similar to the results on the format preference[69]; online and paper formats were used *all the time* or *often* by 65% and 47% of the students respectively. On the other hand, more than 94% and 93% of the students reported *rarely* or *almost never* using pocket electronic format and CD-ROM format respectively.

It is interesting to compare the results of format preference from the main survey with the results from the pilot survey. In the pilot survey, pocket electronic format was by far the most preferred (56% of the students), followed by paper format (29%), online format (8%), and CD-ROM format (7%). The considerable difference in format preference between the students in the pilot survey and the main survey could be explained by the differences between the two groups of students; the student group in the pilot survey was very homogenous (NNSs about to start Masters study, mostly from Asian countries), whereas the student group in the main survey was more diverse (NSs and NNSs from many different countries, studying different subjects at many different levels). Yet, if the format preference is correlated with the country of origin of the students in both studies, it becomes evident that the preference of pocket electronic format is by no means limited to NNS students from Asia; many NS students from the UK reported preference for this format. Also, the reasons for preference of a specific format and the frequency of use of each format were very similar to the ones reported by the students in the main survey.

4.1.1.3 Dictionaries used

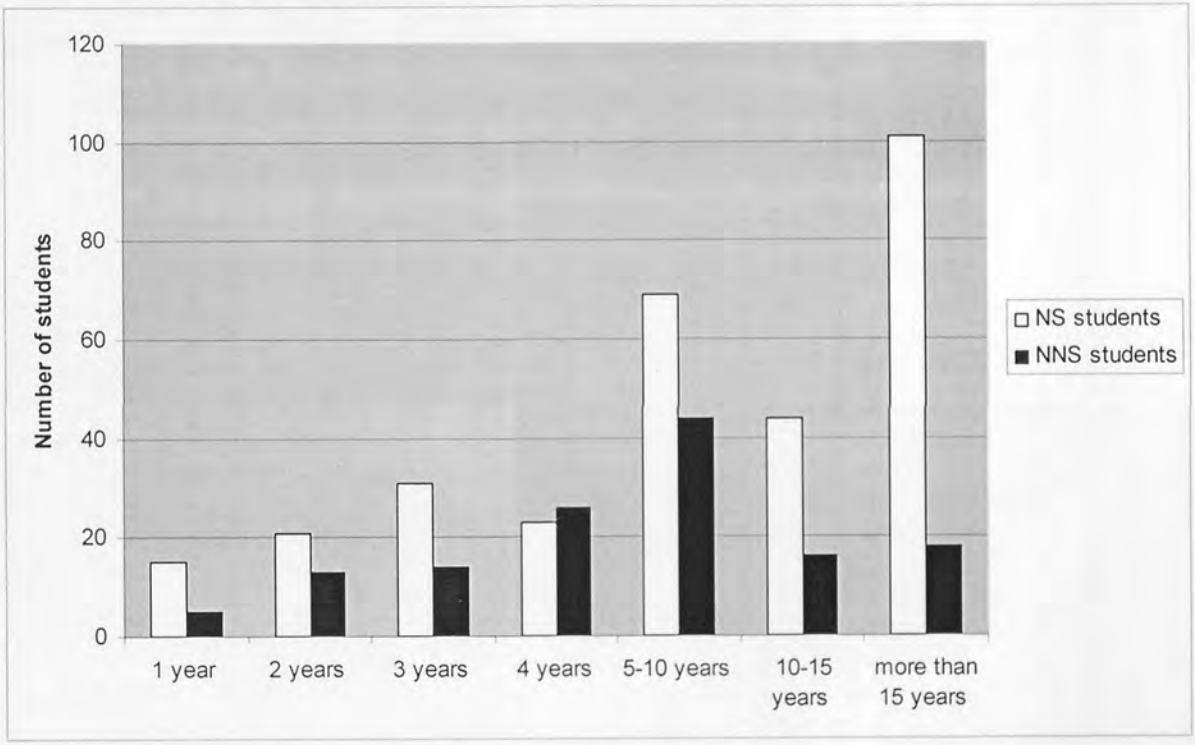
71% of the students reported using a monolingual English dictionary, and although the other 29% reported *not* using one, the answers they provided for some of the other questions

[69] The questions of format preference and frequency of use were conducted independently from one another (see Question 2 and Question 3 in Appendix 3).

(e.g. about the dictionary they use, and the frequency of use of each dictionary format) suggest that they may have not understood the question completely. Out of those 29%, most of the students were NSs; in fact, almost one third of the NSs said they do not use a monolingual dictionary.

The students who said they use a monolingual English dictionary were asked to estimate how long they have been using one (Figure 40). On average, the NS students were using their monolingual English dictionaries much longer than NNS students, with 23% using one for more than 15 years and nearly 50% reported using one for 5 years or more.

Figure 40. Main survey: Reported length of use of a monolingual English dictionary (n=440).



Students were also asked to provide information on the monolingual English dictionaries they were using[70]. Table 109 and Table 110 show lists of most frequently reported dictionaries by NS students and NNS students respectively. The predominant formats were online and paper,

[70] The analysis identified two problems connected with the answers to this question. Firstly, many students, especially NSs, reported using the print format of the Oxford English Dictionary. As it is highly unlikely that the students would own the 20 volumes of this historical dictionary, it is believed that the students probably meant the Oxford Dictionary of English or some other Oxford dictionary with a similar name. Secondly, some of the students who reported using the New Oxford Dictionary of English named Cambridge or Collins as the publisher of the dictionary, or reported using the dictionary online (the dictionary is only available in print and on CD-ROM). These inconsistencies might have been caused by the fact that the New Oxford Dictionary of English was offered as an example of the answer to Question 8. This may have influenced the students who did not remember the name of the dictionary they were using.

with online format being more or less limited to a few websites offering access to multiple dictionaries (e.g. Dictionary.com, Google, askoxford.com).

Although the NS list is dominated by dictionaries published by HarperCollins and Oxford University Press, by far the most popular dictionary (or rather an online collection of dictionaries) is Dictionary.com. Clearly, the internet has become an important reference source for university students. This is also supported by the fact that Google was named by 15 students (who reported using the 'define:' query to obtain a definition of a word).

The list of dictionaries most frequently used by NNS students (Table 110 in Appendix 5) contains many (NS) dictionaries that were also named by the NS students (these dictionaries are indicated in bold in Table 110). Interestingly, even advanced learner's dictionaries were used by a few NS students[71]; however, they were used more frequently by NNS students, the Oxford Advanced Learner's Dictionary being the most popular. Dictionary.com was also quite popular among NNSs, albeit not to the same degree as among the NSs. Dictionary.com is also the only significant difference between the dictionaries used by the NNS students in this survey, and the dictionaries used by the NNS students in the study by Nesi and Haill (2002), indicating a shift to online dictionary use.

More than 100 other dictionaries were mentioned by students, but many of them were reported only once or twice. Among them were various technical dictionaries, thesauri, and dictionaries for university students. The latter included MWCD (2 students), COEDUCS (1 student), and LED (1 student). Since dictionaries for university students have only recently appeared in the UK, these results are not all that surprising.

4.1.1.4 Knowledge and use of existing dictionaries for students

When asked about the three dictionaries for students (see 2.2.1), the students showed a low level of familiarity (Table 24). LED and MWCD are relatively unknown to students and are used by a very small percentage of students. Slightly more known to students and used by them is COEDUCS.

The relatively low level of student familiarity with the three dictionaries is not that surprising as LED and CODUCS are relatively new to the market, and MWCD is targeted at university and college students in the US.

---

[71] Other than Cambridge Advanced Learner's Dictionary (3 students), listed in Table 109, NSs also reported using OALD (2 students) and MEDAL (1 student).

Table 24. Main survey: Familiarity with the three dictionaries for students.

	LED	MWCD	COEDUCS
I've never heard of it before.	67%	73%	34%
I've heard of it, but haven't used it.	26%	20%	35%
I use it occasionally.	6%	5%	22%
I use it regularly.	1%	1%	8%
No answer	0%	0%	0%

NNS students were significantly more familiar with LED (Mann-Whitney U = 28284, N = 618, $p<0.01$), COEDUCS (Mann-Whitney U = 29385, N = 617, $p<0.01$), and MWCD (Mann-Whitney U = 32038, N = 617, $p<0.01$). This finding is particularly interesting in the case of COEDUCS and MWCD, as they target NS students.

4.1.1.5 Activities for which dictionaries are used

Students were asked to attribute a level of importance to using a dictionary for seven different activities. Five activities were related specifically to academic work, and two were less academic (*writing emails, letters & CVs*; *reading books & newspapers*). As shown in Table 25, the level of importance varies considerably across the activities. Dictionaries were considered very important for writing academic work, with 75% of the students considering them *important* or *very important*. Other activities where dictionary-use was ranked highly in terms of importance were reading academic work, presenting academic work, and writing emails, letters & CVs.

Table 25. Main survey: Importance of dictionaries for different activities.

	MEAN RANK
Writing academic work	3.07
Reading academic books, journals	2.61
Writing emails, letters, CVs	2.44
Presenting academic work	2.43
Reading books, newspapers	1.88
Listening to academic lectures	1.78
Speaking with lecturers	1.53

Ranks: 4=very important, 3=important, 2=not very important, 1=not important)

The activities where the use of dictionaries was not deemed important were reading books & newspapers, listening to academic lectures, and speaking with lecturers. The results for the last two activities are somewhat expected as the nature of listening and especially speaking hinders dictionary use.

Overall, NNS students considered dictionaries more important than NS students for all seven activities. Noticeable differences can be observed in presenting academic work (65% of NNSs considering it *important* or *very important* versus 45% of NSs), writing academic work (88% versus 71%), and reading books, newspapers (33% versus 16%).

The format of the dictionary also influences the importance attributed to dictionaries for specific activities. For example, a large number of the students in the pilot survey preferred pocket electronic dictionaries, which are more convenient to use during listening and speaking. Accordingly, the students in the pilot survey attributed much more importance to the use of dictionaries for listening and speaking activities than the students in the main survey.

4.1.1.6 Testing some general statements

In Question 7, the students were asked to express their degree of agreement with three statements about dictionary use. Disappointingly, half of the students partly or strongly agreed with the statement that they look at only the first sense, while only 21% strongly disagreed with it (Table 26). This supports the findings of Mitchell (1983), Tono (1984), Neubach and Cohen (1988), McCreary (2002), and Nesi and Haill (2002), and points to the importance of sense ordering.

Table 26. Main survey: Looking at only the first sense of the word.

strongly agree	9%
partly agree	41%
partly disagree	29%
strongly disagree	21%

More encouraging are the results of the second statement (Table 27), 39% of the students indicated that they to use more than one dictionary if the first does not give them information they are looking for. The level of agreement expressed by the NNS students is particularly high, with 54% strongly agreeing and 35% partially agreeing with the statement.

Table 27. Main survey: Using more than one dictionary.

strongly agree	39%
partly agree	34%
partly disagree	17%
strongly disagree	10%

The fact that 55% of students expressed some sort of agreement with the third statement (Table 28) is slightly worrying for dictionary-makers. The results support the belief that is often discussed among lexicographers, namely that certain publishers have an upper hand purely because of their name. So although a lesser-known publisher may produce an extremely good dictionary for students, many students will probably still choose to buy a dictionary by a more widely-known publisher.

Table 28. Main survey: Importance of a dictionary publisher's name.

When buying a dictionary, the name of the publisher is a very important factor.	
strongly agree	20%
partly agree	35%
partly disagree	24%
strongly disagree	21%

4.1.1.7 Dictionary-use (parts of entry, typical strategies)

In Question 6, the students were asked how frequently they consult different parts of a dictionary entry. Eight microstructural features were examined, ranging from definitions to collocations. Each category was accompanied with an explanation, as it was anticipated that some students may not be familiar with the meaning of words such as *synonym* or *collocate*.

The results in Table 29 show that definitions are by far the most frequently consulted microstructural feature. In fact, 60% of students reported consulting definitions almost always. On the other hand, there are four features that are very rarely looked for: collocates (over 50% of the students almost never consult them), pronunciation, frequent phrases, and usage and grammar.

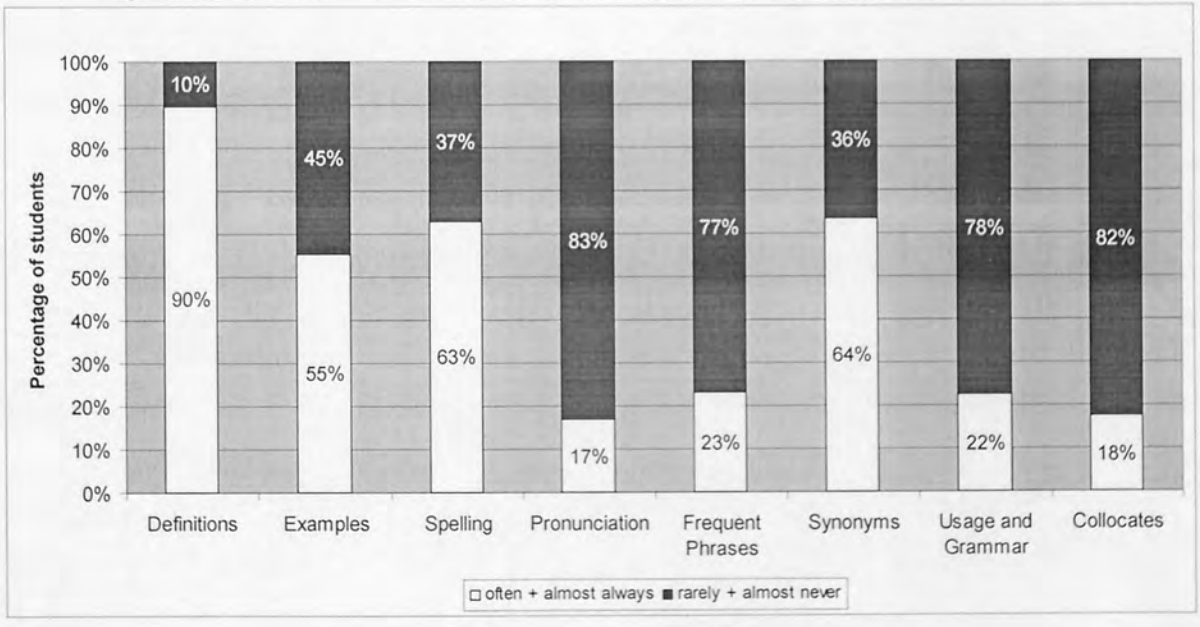
Table 29. Main survey: Frequency of using parts of dictionary entry.

	MEAN RANK
Definitions	3.47
Spelling	2.79
Synonyms	2.71
Examples	2.58
Usage and Grammar	1.84
Frequent phrases	1.83
Pronunciation	1.74
Collocates	1.67

Ranks; 4=almost always, 3=often, 2=rarely, 1=almost never

The results become even more significant when the answers *almost always* and *often*, and *rarely* and *almost never* are combined. This comparison of the frequent and infrequent use (or in some cases non-use) is shown in Figure 41. It becomes clear that the microstructural features can be divided into three groups: the first group contains the feature used frequently by a vast majority of students (definitions), the second group includes features used frequently by approximately two thirds of the students (synonyms, spelling, examples), and the third group includes four features that are (very) rarely consulted by approximately 80% of the students (frequent phrases, usage and grammar, collocates, pronunciation).

Figure 41. Main survey: Frequency of use of part of entry with conflated answers.



4.1.1.7.1 Dictionary use by native speakers and non-native speakers

NS students and NNS students do not differ significantly in the frequency of use of definitions (Mann-Whitney U = 34985, N = 620, p=0.05) and spelling (Mann-Whitney U = 36331, N = 620, p=0.28). The other six microstructural features are used significantly more frequently by NNS students (Mann-Whitney U = various, N = 620, p<0.01), which is also evident in Table 30. One possible explanation why the NNS students use some of the features (e.g. collocates, frequent phrases) much more frequently is that these features are limited to monolingual English dictionaries for foreign learners, or are given a more prominent role in them.

These results confirm most of the findings by other studies (Quirk, 1975; Tomaszczyk, 1979; Béjoint, 1981; Jackson, 1988; Battenburg, 1989; Harvey & Yuill, 1997; Hartmann, 1999), One noticeable difference is that the NNS students in this survey do not consult collocational and grammatical information as often as suggested by the studies of Béjoint (Béjoint, 1981) and Harvey & Yuill (1997), but also not as rarely as reported by Nesi (2000).

Table 30. Main survey: Use of microstructural features - NS and NSS students.

	NS students *(mean rank)*	NNS students *(mean rank)*
Definitions	3.44	3.56
Spelling	2.82	2.73
Synonyms	2.63	2.91
Examples	2.45	2.92
Usage and Grammar	1.72	2.16
Frequent phrases	1.66	2.27
Collocates	1.49	2.15
Pronunciation	1.60	2.10

4.1.1.7.2 Dictionary use by students from different Aston Schools

Students from different Schools at Aston University use dictionaries differently (Figure 42 and Table 31 below).

LHS students and Combined Honours students consult definitions more frequently than students of other Schools. LSS students use (the generally less consulted) five out of eight features more frequently than any other group of students. LHS students use three features (usage and grammar, frequent phrases, and collocates) less frequently than any other group of students.

Figure 42. Main survey: Use of microstructural features - by Aston Schools.

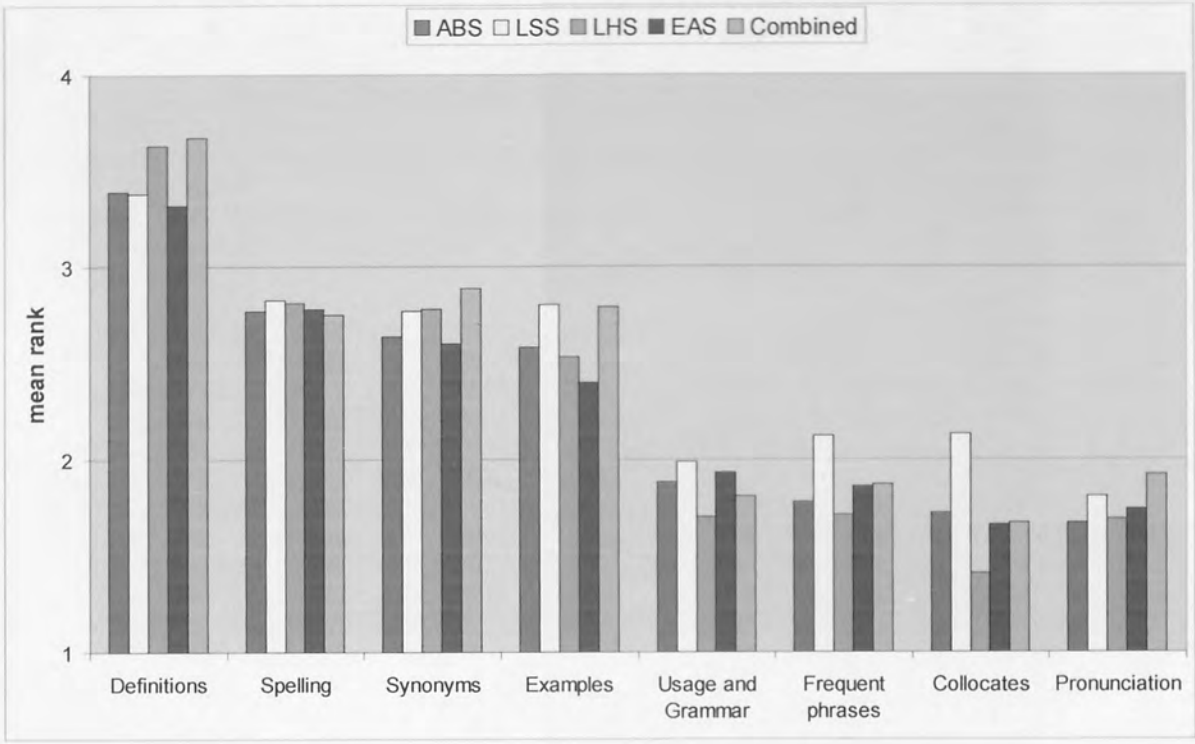


Table 31. Main Survey: Mean rank data for Figure 42.

	ABS	LSS	LHS	EAS	COMBINED
Definitions	3.39	3.38	3.63	3.32	3.67
Spelling	2.77	2.83	2.81	2.78	2.75
Synonyms	2.64	2.77	2.78	2.60	2.89
Examples	2.58	2.80	2.53	2.40	2.79
Usage and Grammar	1.88	1.99	1.70	1.93	1.81
Frequent phrases	1.78	2.12	1.71	1.86	1.87
Collocates	1.72	2.13	1.41	1.66	1.67
Pronunciation	1.67	1.81	1.69	1.74	1.92

Ranks; 4=almost always, 3=often, 2=rarely, 1=almost never

There are also some similarities between students from different Schools in the consultation of microstructural features. All groups of students exhibit the same pattern of feature consultation; definitions are by far most frequently consulted, followed by spelling, synonyms, and examples, whereas usage and grammar, frequent phrases, collocates, and pronunciation are consulted rarely or almost never. Also, students from all Aston Schools consult spelling with very similar frequency.

4.1.1.7.3 Dictionary use by students on different courses

There are also differences in consultation of dictionary microstructure features between students of individual subjects, as shown in Figure 43 and Table 32. Students of six subjects have been selected for comparison, each subject formed of one or more courses[72]. The findings of the comparison point to different needs of students of different subjects.

Figure 43. Main survey: Use of microstructural features by academic subjects.

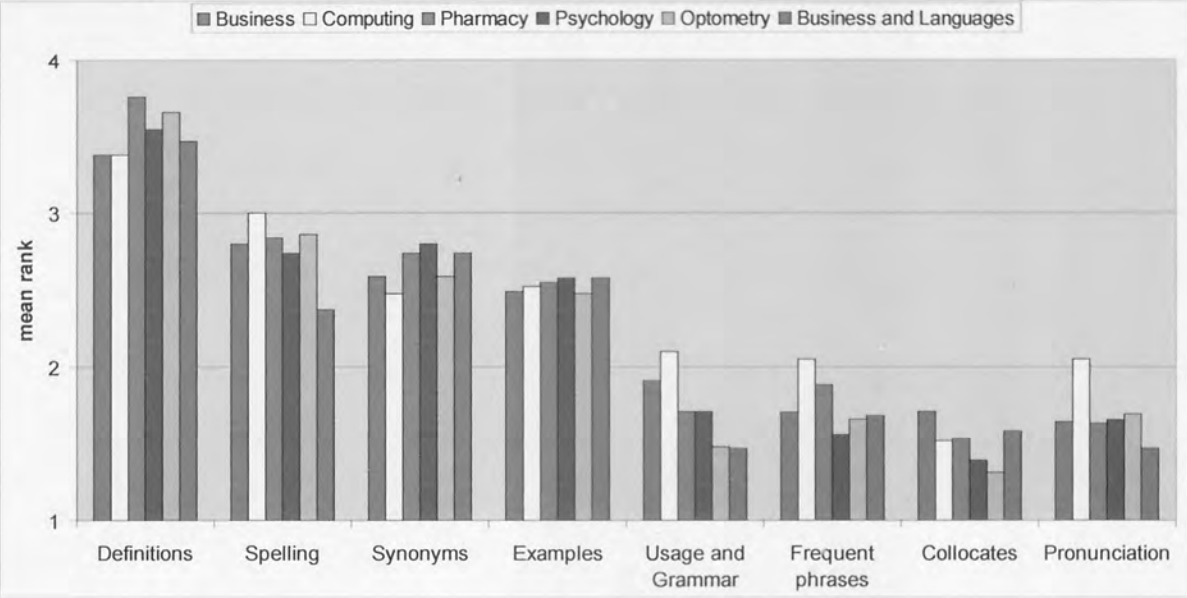


Table 32. Main Survey: Mean rank data for Figure 43.

	Business[a]	Computing[b]	Pharmacy[c]	Psychology[d]	Optometry[e]	Business + Languages[f]
Definitions	3.38	3.38	3.76	3.55	3.66	3.47
Spelling	2.80	3.00	2.84	2.74	2.86	2.37
Synonyms	2.59	2.48	2.74	2.80	2.59	2.74
Examples	2.49	2.52	2.55	2.58	2.48	2.58
Usage and Grammar	1.91	2.10	1.71	1.71	1.48	1.47
Frequent phrases	1.70	2.05	1.88	1.56	1.66	1.68
Collocates	1.71	1.52	1.53	1.39	1.31	1.58
Pronunciation	1.64	2.05	1.63	1.65	1.69	1.47

[a] - 66 students in total from three courses: Business and Management (39), Business Administration (18), and Business (9).

[b] - 21 Computing Science students.

[c] - 49 Pharmacy students.

[d] - 66 students in total from two courses: Human Psychology (43) and Psychology (23).

[e] - 29 Optometry students.

[f] - 19 students of International Business and Modern Languages.

[72] Only disciplines with more than 20 students have been included in the analysis.

Computing Science students use four features (spelling, usage and grammar, frequent phrases, and pronunciation) significantly more frequently than students of the other five subjects. Pharmacy students use definitions more frequently than the other groups. Definitions are used much less frequently by Business and Computing Science students than the other four groups of students. Students of International Business and Modern Languages use spelling much less frequently than the other five groups.

There are even differences between students of neighbouring subjects such as Pharmacy and Optometry (both courses are offered in LHS), which could be classified under the common category of medicine. On average, Pharmacy students use features such as definitions, synonyms, usage and grammar, frequent phrases, and collocates more frequently than Optometry students.

Some similarities between the six subjects are found in the ranking of features (see 4.1.1.7.2 for more), and in the frequency of consulting examples.

4.1.1.8 Reported language proficiency

The students were also asked to evaluate their own English language proficiency in listening, reading, speaking, and writing. As shown in Table 33, the majority of the students were very confident in their language proficiency[73]. The reported proficiency for receptive tasks (listening and reading) was higher than for productive tasks (speaking, writing).

250 students (213 NSs and 37 NNSs), representing 40% of the student group, reported to possess excellent proficiency for all four activities.

Table 33. Main survey: English proficiency for four language activities.

	MEAN RANK
Reading	5.41
Listening	5.40
Speaking	5.26
Writing	5.18

Ranks: 6=excellent, 5=very good, 4=good, 3=fair, 2=fair, 1=very poor

[73] The findings offered by this question are not very reliable because they elicit the students' perceptions and not factual information (Nesi, 2000; Hatherall, 1984). Moreover, lecturers at universities would probably argue that there is a big difference between the language proficiency perceived by the student and the actual proficiency. However, for researching dictionary use, perceived language proficiency is a valuable piece of information as the users' decision whether to consult a dictionary or not is more likely to rest on how confident they are in their language ability, rather than on the actual proficiency based on some formal assessment.

NS students were significantly more confident than NNS students in their English language proficiency in reading (Mann-Whitney U = 8.6200, N = 620, $p<0.01$), listening (Mann-Whitney U = 12.0292, N = 620, $p<0.01$), speaking (Mann-Whitney U = 8.7621, N = 620, $p<0.01$), and writing (Mann-Whitney U = 6.6245, N = 620, $p<0.01$). Nevertheless, mean ranks of NNS students for reported language proficiency are still quite high for all four activities (see Table 34).

Table 34. Main survey: English proficiency for four language activities - NS and NNS students

	Reading		Listening		Speaking		Writing	
	NS	NNS	NS	NNS	NS	NNS	NS	NNS
MEAN RANK	5.52	5.12	5.56	4.98	5.39	4.92	5.29	4.90

The correlation between the perceived English language proficiency and the frequency of dictionary use[74] (Table 35) reveals that the frequency of dictionary use in all four activities increases with lower language proficiency. But even though lower perceived proficiency seems to encourage more frequent dictionary use, the results show that all the students, regardless of their perceived English language proficiency, use dictionaries often or even more frequently (mean rank is always more than 3).

Table 35. Main survey: Comparison of English language proficiency and dictionary use.[75]

		excellent	very good	good	fair
LISTENING	*mean rank*	3.04	3.19	3.32	3.33

		excellent	very good	good	fair
READING	*mean rank*	3.02	3.16	3.39	3.59

		excellent	very good	good	fair
SPEAKING	*mean rank*	3.02	3.20	3.16	3.44

		excellent	very good	good	fair
WRITING	*mean rank*	3.05	3.10	3.21	3.41

Ranks: 4=all the time, 3=often, 2=rarely, 1=almost never

[74] This information was obtained by correlating the answers to Question 2 (Which dictionary format do you prefer?) and the relevant part of Question 3 (How often do you use these dictionary formats?) – for example, if the subject reported preference of paper format, the reported frequency of using paper format was used in correlation.

[75] Data for *poor* and *very poor* was not included due to very low representation (very few students, if any at all, selected one of these two answers).

Similar findings are found by correlating the reported proficiency data with the attributed importance of dictionary use for different activities (Figure 98 and Table 111 in Appendix 5). The lower the proficiency of the student, the greater the importance attributed to the use of dictionaries. For several activities (e.g. Presenting) there is a noticeable difference in importance attributed to dictionary use between the students with excellent language proficiency, and the students with very good language proficiency. But, regardless of their proficiency, the students do not consider the use of a dictionary to be important for reading books and newspapers, listening to academic lectures, and speaking to lecturers.

4.2 Profile of the potential users of DOAE

The users

The target users are thus university students, with different language backgrounds, and studying different subjects. Most students are between 18 and 24 years old. NNS students have been learning English for 10-15 years.

Students believe they possess very high English language proficiency in listening and reading, and slightly less high proficiency in speaking and writing. NNSs tend to be slightly less confident in their language skills than NSs.

Experience of using monolingual English dictionaries

NS students have more experience (around 5-10 years) than NNS students in using monolingual English dictionaries. It should be pointed out, though, that the purposes of using a dictionary at a university are likely to differ from the purposes of using a dictionary in a primary or a secondary school.

What monolingual English dictionaries do students currently use and how often?

Students, both NSs and NNSs of English, use large monolingual English dictionaries for NSs, mainly the ones produced by HarperCollins and Oxford University Press. Dictionaries for advanced learners are used by a large percentage of NNS students, and a very small percentage of NS students. Dictionaries available online are very popular among the students, especially Dictionary.com and Google. Dictionaries for university students are not known to many students, and are used by a very small percentage of students.

Students use dictionaries often, but only online and paper formats. Frequency of use is closely related to the nature of activity – students use dictionaries more frequently when doing

academic work. Reported language proficiency does not seem to play a major role in the frequency of dictionary use, although the frequency of dictionary use increases with lower proficiency.

Which dictionary formats do students prefer?

Students prefer online and paper dictionaries. Many students, while preferring one of the two formats, use both of them, depending on the circumstances. Online dictionaries are preferred for academic work.

The use of pocket electronic dictionaries seems to be a much localised phenomenon, as this dictionary format is used very frequently by students from Asia, and much less frequently by students from other parts of the world. Only a small percentage of students prefer dictionaries on CD-ROM, and use them frequently.

How do students choose and use dictionaries?

- When buying a dictionary, the name of the publisher is very important to the students. This finding is further supported by the fact that many students reported using dictionaries of well-known dictionary publishers, such as Oxford and HarperCollins. However, the students who use Dictionary.com, the first hit in Google if the word "dictionary" is searched for, obviously care much less about the name of the publisher.
- When consulting a dictionary, many students admit looking at only the first sense in the dictionary entry.
- Many students are prepared to consult more than one dictionary if they do not find a satisfactory answer in the first one.
- Students who differ in native language or subject of study often use dictionaries differently (i.e. for different activities, consult different microstructural features).

Which microstructural features do they consult?

Students frequently consult definitions, synonyms, spelling and examples. NNS students consult frequent phrases, usage and grammar information, pronunciation, and collocational information quite frequently. These four features are rarely consulted by NS students – perhaps because these features are not prominent or available in dictionaries for NSs, rather than because they are not interested in them.

What do students use dictionaries for?

Students use dictionaries for both encoding and decoding. Students consider dictionaries to be especially important for writing academic work (e.g. essays), reading academic books & journals, writing emails, letters, & CVs, and presenting academic work[76]. NNSs attribute more importance to the use of dictionaries than NSs, irrespective of the activity.

This profile, while based on a survey with UK students, is likely to characterize a student at any university where English is the language of instruction. After all, many students in the survey came from countries outside the UK. The only information which could be regarded as UK-student specific is the one about the monolingual English dictionaries used. Students in the US, for example, are likely to use dictionaries by American publishers (see McCreary, 2002). However, this may only be the case with some dictionaries, as dictionaries such as Dictionary.com are not publisher- or country-specific.

4.3 Main implications of the user profile for the Model for DOAE

The user profile has provided a great deal of information about students' dictionary use. All the information will be useful to dictionary-makers, but some of the information is particularly relevant for the Model for DOAE proposed in this thesis. The information in question and its implications for the dictionary design are presented here:

a) There has been a fall in popularity of the paper format in comparison with previous studies into student dictionary use (Hartmann, 1999; Nesi & Haill, 2002), and considerable increase in the popularity of the online format. The online format is also the preferred format for students consulting dictionaries while doing academic work. In addition, the online format can contain much more information than any other format. For these reasons, the online format will be used as a point of departure for the proposed dictionary Model.

b) Students mainly use dictionaries for writing academic work, doing presentations of academic work, and reading academic material. The finding that dictionaries are considered most important for writing academic work corresponds to the results obtained by Hartmann (1999). Considering that writing is an encoding activity, encoding information should play an important role in the entries of DOAE.

[76] Presenting academic work can be classified as a writing activity as students are more likely to use a dictionary when preparing their presentations rather than during the presentations.

c) NS students and NNS students use dictionaries in different ways, most pertinent differences being the ones related to the use of microstructural features. Similar differences are also observed between students of different subjects. The ability to tailor dictionary displays to the identified consultation habits of different types of students would add a valuable dimension to the user-friendliness of the Model.

d) Many students admitted using the 'choose the first definition' strategy which confirms the findings of previous studies (see 2.3.4.1). The proposed Model needs to bear this in mind, and offer the most relevant senses first, i.e. academic senses. However, different senses may have a different importance for different types of students (e.g. students of different subjects and/or from different language backgrounds). Hence, whether a feature can be offered that could re-order senses according to different types of students needs to be explored.

The user profile, especially its main implications discussed above, will therefore be used as a basis for the next stage in which the Model for DOAE will be designed.

5. MACROSTRUCTURE OF THE MODEL FOR DOAE

Now that the user profile has been established and the corpus has been built, the work on the actual dictionary can begin. The creation of a dictionary starts with important decisions about dictionary macrostructure, which mainly involves compiling a list of headwords. The procedure of headword selection for the proposed Model, and decisions on what constitutes a headword (i.e. criteria for which words or phrases can be headwords) are discussed in this chapter. A short discussion on accompanying material is also provided.

Another important macrostructural decision is the selection of dictionary format. Online format has been selected as the primary format for the proposed Model because it represents the most complete format in terms of the data it can contain, and in terms of customizability of dictionary information display. Even more importantly, online format is the preferred format among most university students, especially for academic work (see 4.1.1.2).

5.1 Headword list

The headword list must be prepared in advance in order for the work to be efficiently distributed among lexicographers and properly monitored. How extensive the headword list is depends on the size of the dictionary, which is dictated by the needs of the target users. Students mainly use dictionaries for writing and reading academic work, so this dictionary Model needs to offer comprehensive coverage of academic English.

The main criterion for selecting headwords is corpus frequency. A lemma has to be found in the corpus to be considered for headword status. Furthermore, as will be demonstrated next, it is essential to set a minimum frequency for headword status to exclude very rare items.

Setting frequency or any other selection criteria for headwords is complicated by the fact that headwords differ in terms of structure; most are single words (single-word headwords), while others consist of more than one word (multi-word headwords). Each type of headword requires its own set of instructions which stipulate how, and at which stage of the dictionary-making process, the headwords should be identified and selected.

5.1.1 Selecting single-word headwords

Atkins and Rundell (2008) divide single-word items into simple words, abbreviations and contractions, and partial words. Simple words consist of lexical words (nouns, verbs,

adjectives, adverbs, interjections) and grammatical words (prepositions, conjunctions, pronouns, auxiliary verbs, determiners). Abbreviations consist of alphabetisms (e.g. *CNN*), and acronyms (e.g. *CALL*). Both subgroups of simple words are normally given headword status in dictionaries, and DOAE does not depart from this practice.

Partial words consist of bound affixes (e.g. *im-* in *imperfect*), productive affixes (e.g. *ex-* in *ex-husband*), and combining forms (e.g. *snow-covered*). However, although partial words are found as headwords in many dictionaries, they are not given headword status in DOAE. One reason is that students are expected to be familiar with word formation, and the other is that they are unlikely to look for partial words when encountering/producing words and phrases. It is much better to give headword status to words containing partial words (especially combining forms and productive affixes), provided they meet the minimum corpus frequency criterion.

A lemma list (alphabetically ordered) has been selected as the most suitable type of wordlist for the selection of single-word headwords. Lemmatisation uses a lemma list that is based on existing data from dictionaries and corpora, and is better than a non-lemmatised wordlist as it does not include inflected forms of known lemmas. Word forms of unknown lemmas (lemmas not on the lemma list) are all listed as lemmas, but they are likely to be infrequent.

The problem of using a lemma list is that sometimes an inflected form of a lemma can be a headword candidate. Especially problematic, for both headword selection and lemmatisation, are past participles of verbs that are also adjectives. For example, both uses of *broken*, namely as the part participle form of the verb *break* and as an adjective, are treated by Sketch Engine as an inflected form of the lemma *break*. One could of course use a non-lemmatised wordlist where items like *broken* would be listed separately, and could be made headwords. However, such practice could affect analysis as concordance lines could show that the headword status is not justified (e.g. *broken* is never used as an adjective), but the lexicographer, influenced by the pre-determined headword list, may still compile an entry rather than accept its non-headword status. Thus, it is better to use a lemma list and leave the decision of the treatment of such problematic cases to lexicographers.

The selection of headwords is not a simple process of 'one lemma - one headword', partly because the list of lemmas and associated word forms used by Sketch Engine to produce lemma list does not include all the lemmas in the corpus. Especially problematic for identification of candidate headwords are errors caused by conversion, tokenisation and other

processes that are part of corpus creation. These issues become apparent after a quick examination of random parts of the lemma list. Table 112 and Table 113 in Appendix 6 show the first 50 lemmas in an alphabetically ordered lemma list, and lemmas beginning with *extent*, respectively. One thing that both extracts from the lemma list share is a high number of lemmas with frequency of 1. Furthermore, most of the single-occurrence lemmas appear to be the result of an error in tokenisation or conversion rather than actual words. This is also true of several lemmas with a frequency of more than 1.

The usefulness of the lemma list can therefore be significantly improved by introducing a cut-off point, i.e. a minimum frequency for a lemma. A minimum of 5 occurrences has been selected for DOAE. On the one hand, the introduction of this cut-off point aims at eliminating a vast majority of lemmas that are the result of various errors in corpus creation. On the other hand, lemmas with frequency of below 5 can be considered extremely rare, so they are less likely to be encountered by students, which means there is also less frequent need to look them up.

Introducing a minimum frequency of 5 makes the lemma list much more manageable for dictionary makers, reducing the number of lemmas from original 1,248,248 to 217,389. As shown in Table 36, the biggest drop in lemma count is caused by omitting lemmas with a frequency of 1 which represent slightly over 60% of lemmas in the complete lemma list.

Table 36. DOAE macrostructure: CAJA lemma count at different cut-off points.

CAJA lemma count	cut-off point
1,248,248	no minimum frequency
492,339	minimum frequency = 2
332,876	minimum frequency = 3
260,277	minimum frequency = 4
217,389	minimum frequency = 5

It is worth noting that the lemmatised list with a cut-off frequency of 5 is not free of lemmas produced by errors in corpus creation. For example, *extent'* from Table 113 will be on the lemma list because it has a frequency of more than 5; however, it is not a candidate for a headword. Moreover, none of the lemmas in Table 112 with a frequency of more than 5 will become a headword as most are punctuation marks, and the remaining one is an error (##). In fact, the only lemma from the two aforementioned tables that is a candidate headword is *extent*.

The selection of headwords is not always straightforward. It is always difficult to decide whether a lemma should be made a headword without looking at concordance lines. However,

at this stage, the examination of concordances should be avoided as it would make headword selection a very time-consuming process. It is thus better to leave a note under the headword to which the problematic lemma is (probably) related in meaning, leaving the decision to the lexicographer working on that headword.

The selection of headwords is exemplified by an extract from the lemma list (frequency ≥ 5), provided in Table 37 (the complete lemma list, as well as the lemma list ordered by frequency, are provided on the attached CD-ROM – Appendix 13). Firstly, lemmas that were clearly not candidates for headword (shown in italics in the table) were excluded. Then, lemmas that were candidates for headword status (e.g. *sense*) were identified (shown in bold in the table). The status of the remaining lemmas was difficult to determine (e.g. *sense-perception*), so they were listed as potential candidates in the form of a note under the entry *sense*.

Table 37. DOAE macrostructure: CAJA lemmas starting with *sense* (alphabetically ordered, frequency ≥ 5).

lemma	frequency
sensationalistic	5
sensationalize	19
sensationally	7
sense	26396
sense-bites'	9
sense-data	11
sense-individuation	6
sense-making	62
sense-of-direction	21
sense-perception	12
sense/	7
sensegiving	13
sensei	5
senseless	75
senselessness	13
sensels	7
sensemaking	153
senses'	5
sense—that	5
sense—the	6
sense'	38
sense"	9
sensi	7
sensibility	774
sensible	838

The procedure of headword selection described above will provide a comprehensive list of single-word headwords. But what about the lemmas that have been left out of the selection, namely lemmas with a frequency of less than 5? Should these lemmas be completely ignored?

First of all, not all 'lemmas' are actual lemmas. In fact, the lemmas can be divided into the following groups:

a) Rare lemmas (frequency < 5). These lemmas would be candidate headwords if it were not for their low frequency.
b) Error-lemmas, related to rare lemmas. These lemmas are rare lemmas, normally containing punctuation mark(s) and/or number(s) and/or letter(s). They are normally the result of errors in tokenisation, conversion, and/or a typo.
c) Error-lemmas, related to lemmas already identified as headwords. These lemmas have similar origins to lemmas under b).

One problem that arises from this grouping is that a rare lemma may become a candidate headword if its frequency, combined with the frequency of error-lemma(s) related to it, is 5 or more. An example of this is the lemma *attributor* (frequency = 2) which has two related error-lemmas, namely *'attributor* (frequency = 1) and *attributor's* (frequency = 2). These additional candidate headwords could be identified by a computer program that would search the list of lemmas with a frequency of less than 5, identify lemmas with a common sequence of characters, extract those lemmas, and compare them with the existing list of headwords (and their word-forms) to exclude lemmas from group c) above.

Another problem concerns error-lemmas under c). Although these lemmas are represented by existing headwords, their concordance lines may be excluded from the analysis. This could be considered problematic if the combined frequency of error-lemmas related to a particular headword is a large proportion of, or perhaps even greater than of frequency of the headword. Such problematic cases could be identified by the program written for identification of additional headwords (described in the previous paragraph), the only difference being that the program would list the excluded lemmas as well as their (combined) frequency. If a certain percentage of headword frequency was exceeded, the program would create a note for the lexicographer which would be entered into the database under the relevant headword.

Writing such programs, and testing them, is beyond the scope of this thesis, so it was decided to exemplify the search for additional headwords with a semi-automated approach, using the lemma list, and the Microsoft Excel program. ATTRIBUTE was selected as the sample headword. Because there was no headword list to use as a reference, the search also looked for derivatives of ATTRIBUTE that are candidate headwords. The search was conducted as follows:

1) The lemma list was searched for all lemmas containing the expression *attribut*[77]. 205 lemmas, including ATTRIBUTE, were found (see Table 114 in Appendix 6).
2) 87 error-lemmas of ATTRIBUTE with a combined frequency of 197 were identified (provided in italics in Table 114). Nine lemmas have a frequency of 5 or more, and a combined frequency of 101 (see Table 38). A high combined frequency of error-lemmas warrants a note in the database[78].

Table 38. DOAE macrostructure: CAJA error-lemmas of ATTRIBUTE (frequency ≥ 5).

Lemma	Frequency
Attributes	30
Attribute	24
attributes'	10
Attributed	8
ATTRIBUTES	8
ATTRIBUTE	6
attribute(s	5
attributedto	5
Attributing	5

3) The remaining 118 lemmas were ordered by frequency, and 26 lemmas with a frequency of 5 or more were focused on. Lemmas that are direct candidates (e.g. derivatives) for headword status were identified, and then their error-lemmas were searched for among the lemmas with a frequency of 4 or less (see results in Table 39). Hyphenated lemmas were considered separately as their headword status was often difficult to determine.

Table 39. DOAE macrostructure: List of derivatives of *attribute* that are candidate headwords.

lemma and error-lemmas (frequency)	combined frequency	headword
attribution (1602), *Attribution* (13), *Attributions* (5), *attribution-based* (2), *attribution-poor* (1), *attribution/* (1), '*attributions* (1), *"attribution-type* (1), *attribution-* (1), *attribution.79* (1), *attributions.*1 (1), *attribution'* (1), *attribution-which* (1)	1631	**ATTRIBUTION**
attributable	1329	**ATTRIBUTABLE**
attributional (122), *Attributional* (14)	136	**ATTRIBUTIONAL**
attributive (102), *Attributive* (2), '*attributive'* (1)	105	**ATTRIBUTIVE**
unattributed	15	**UNATTRIBUTED**
misattribute (11), *misattributes* (2), *misattributing* (2)	15	**MISATTRIBUTE**
misattribution (5), *misattributions* (4)	9	**MISATTRIBUTION**
attributively (5), '*attributively* (1)	6	**ATTRIBUTIVELY**

[77] *Attribut* was selected because it is the stem of all the inflected forms and derivatives of *ATTRIBUTE*.
[78] The analysis has later revealed that the most frequent error-lemma, *Attributes*, is in fact included in the lemma search of ATTRIBUTE in Sketch Engine.

4) The remaining 76 lemmas [79] (frequency < 5) were searched for variant (error-)lemmas of the same lemma whose combined frequency was 5 or more. Hyphenated lemmas were not included in the search. Two candidate headwords were identified: OVERATTRIBUTE and ATTRIBUTOR (Table 40).

Table 40. DOAE macrostructure: Two candidate headwords from lemmas of *attribute* (frequency ≤ 5).

lemma and variant (error-)lemmas (frequency)	combined frequency	headword candidate
over-attribute (4), *overattribute* (4), *over-attributes* (1)	9	**OVERATTRIBUTE**
attributor (2), *attributor's* (2), '*attributor* (1)	5	**ATTRIBUTOR**

5) Finally, hyphenated lemmas (the ones not already included in the earlier steps), and their variants (including error-lemmas) encountered during the analysis were examined in more detail (Table 41). These hyphenated lemmas were not candidate headwords, but some of them were variant forms of (more frequent) non-hyphenated forms (which were searched using the Concordance function in this case), which means that a note about the more frequent of these items (e.g. *attribute-level*) had to be made under the entry ATTRIBUTE for consultation during the meaning analysis.

Table 41. DOAE macrostructure: Hyphenated lemmas encountered during the analysis, and their variant spellings.

combining form (lemma variants)	variant form(s) (frequency)	total frequency
attribute-level (59), *Attribute-Level* (4), *Attribute-level* (4), *attributelevel* (1)	*attribute level, attribute levels* (65)	133
attribute-value (27), *attribute-values* (2), *Attribute-Value* (1), *Attribute-value* (1), *attribute-value/* (1), *attributevalue* (1)	*attribute value, attribute values* (54)	87
attribute-based (32)	/	32
multi-attribute (16), *multiattribute* (13), *Multiattribute* (3)	/	32
sub-attributes (20), *sub-attributes/* (1), '*sub-attributes* (1), '*sub-attributes'* (3)	/	25
brand-attribute (11)	*brand attribute, brand attributes* (13)	24
k-attribute (7)	*k attribute, k attributes (5)*	12
attribute-independence (4), *Attribute-independence* (1), *attributeindependence* (1)	*attribute independence* (5)	11
attribute-specific (7)	/	7
self-attribution	/	5

[79] There were 92 lemmas containing *attribut* with frequency of 4 or less, but 16 were already identified as error lemmas of candidate headwords in step 3 (see Table 39).

5.1.2 Selecting multi-word headwords

Different classifications of multi-word items have been developed by linguists (e.g. see Moon, 1992; Cowie, 1998; Moon, 1998a; Cowie, 1999 for discussion); with a clear sets of criteria for the subclasses as yet to emerge (Atkins & Rundell, 2008). But theory, while useful, is often difficult to apply in lexicography, where a more pragmatic and less comprehensive approach is needed when dealing with multi-word items (Moon, 1996). Atkins and Rundell (2008) list various types of multi-word items provided in dictionaries, however of course, not all of them are given headword status (see 5.1.4.1 for more).

One thing that all multi-word items have in common is that their headword status cannot be determined solely by corpus frequency. Other criteria need to be considered, such as transparency of meaning and saliency. Lists of N-grams (sequences of words) are therefore of limited use, as a closer analysis of meaning is required. Furthermore, lists of N-grams miss occurrences of certain multi-word items such as phrasal verbs where the verb and particle are separated by one word or more. Consequently, the identification of multi-word headword candidates should be made by the lexicographers at the analysis stage.

5.1.3 Variant forms[80]

Some items in the headword list will be different forms of the same headword, and often the only difference will be the use of a hyphen (e.g. *co-operative* and *cooperative)*. These items need to be conflated into one headword, with the most frequent form listed as the headword, and the other form(s) listed as variants. Conflation of variant forms does not include variant spellings, as all variant spellings are given headword status (see 5.1.4.4).

In addition, there will be items in the headword list that have multi-word variants which do not feature on the lemma list because they contain spaces (e.g. *well-educated* and *well educated)*. Multi-word variants should be identified during the preparation of the headword list to reduce the possibility of variant forms being overlooked during the analysis. A multi-word variant may also become the headword if other variants are less frequent.

Another good argument for identification of form-related headwords and multi-word variants when building the headword list is that the process can be automated, and can save a great deal of time (and money). But most importantly, accounting for all the variant forms of

[80] A discussion of how variant forms and variant spellings are dealt with in DOAE is found in 5.1.4.3.

the headword ensures that any information about the headword recorded during the analysis takes all variant forms into account.

5.1.4 Factors to consider when building the headword list

Previous sections have demonstrated that establishing whether an item should be given headword status is not always straightforward. There are many factors that dictionary-makers need to consider when assigning headword status to lexical items, and some have already been addressed (e.g. corpus frequency). This section addresses the remaining factors, thus setting the guidelines for editors and lexicographers of the dictionary.

5.1.4.1 Multi-word items

Multi-word items play a very important role in DOAE, representing the close relationship between the meaning of a word and its phraseologies. The idea that words rarely function in isolation is underpinned by the idiom principle (Sinclair, 1991), which argues that most of the language consists of semi-preconstructed phrases. This view has slowly been adopted by lexicographers who have shifted from a word-based to a phrase-based approach to corpus data (Rundell, 1998).

Since phraseology is a very important part of academic English (see 2.1.3), DOAE needs to provide extensive coverage of multi-word expressions. There are three types of multi-word items: phrases and idioms, compounds, and phrasal verbs. Phrases and idioms are not given headword status, so they are discussed in the chapter dealing with microstructure (6.3.3.3).

Compounds and phrasal verbs are treated as headwords, but cross-references also need to be provided under the relevant entries (e.g. *civil* and *servant* will need to include cross-references to the compound headword *civil servant*). Compounds and phrasal verbs may sometimes represent the most frequent uses of certain single-word headwords, in which case a compound or a phrasal verb (or one of its meanings) may perhaps also need to be given a sense status under that single-word headword.

Phrasal verbs in particular need to be hyperlinked with the main verb entry because, as Atkins and Rundell (2008:182) suggest, NSs often do not know what a phrasal verb is, and it is assumed that NNSs do. This difference in expected user knowledge is reflected in the treatment of phrasal verbs in existing dictionaries. Dictionaries for foreign learners normally provide them

under verb entries (e.g. *talk down* is found under *talk* in LED CD-ROM, e-LDOCE, and e-OALD – but is a headword in COBUILD CD-ROM), whereas dictionaries for NSs treat them as headwords (e.g. *talk down* is a headword in CED CD-ROM and MWCD CD-ROM – but is offered under *talk* in NODE CD-ROM).

5.1.4.2 Proper names

Proper names represent encyclopaedic rather than linguistic information, and can be found in (large) monolingual dictionaries, especially in US dictionaries. According to Atkins and Rundell (2008), the three main groups of proper names are place names, personal names, and other names (e.g. ceremonies, organizations). In English, the majority of proper names start with a capital letter, which makes them easier to identify, however there are some notable exceptions like *i-*, *e-*, etc. (e.g. *iPhone*). Another issue with identifying proper names is the fact that they can be single-word items or multi-word items. Single-word proper names can therefore be identified during the building of headword list, whereas multi-word proper names have to be identified by lexicographers during analysis.

The target users of DOAE (university students) will use the dictionary during their studies, a period when a person encounters (and needs to produce) a great deal of information, including encyclopaedic information[81]. This calls for inclusion of proper names from all three groups, with the exception of personal names and surnames.

5.1.4.3 Inflected forms

There are two types of inflected forms: regular and irregular. Regular inflected forms follow the regular language patterns, such as adding suffixes *–s*, *-ed,* and *–ing* to verbs. Regular inflected forms will be listed under their headword. Irregular inflections like irregular plurals of nouns (e.g. *oxen*), irregular comparatives and superlatives of adjectives (e.g. *worse* and *worst*), and irregular verb inflections (e.g. *brought*) will be given headword status.

5.1.4.4 Variant spellings

Variant spellings, namely British English spellings and American English spellings, will be given headword status. Variant spellings are found in CAJA as the texts in the corpus have

[81] Proper names present an interesting dilemma for DOAE because although a dictionary has a different purpose from an encyclopaedia, students may still expect to find encyclopaedic information in the dictionary.

been produced by authors from all over the world. The decision on which variant spelling will be the main one in the dictionary should be left to the user. Hence, the database entries of headwords with variant spellings need to have the information recorded in two ways: one with British English as the main spelling, and the other one with American English as the main spelling.

However, the selection of spelling will not apply solely to headwords with variant spellings. Variant spellings are likely to be encountered within dictionary entries as well (e.g. in examples). For this reason, a dictionary project needs to select one spelling as the main one for initial analysis, and once the entries are built, add the information related to the other spelling. For the proposed dictionary Model, British English has been selected as the main spelling for recording the information in the database (the American English version is discussed in 6.4), as the design of the Model is based on the user profile obtained from the survey of UK students.

Variant spellings of the headword should be analysed together, as a link between the variant spellings needs to be established in the database (or severed if the variant spellings have different meanings).

5.1.4.5 Derivatives

Derivatives are derived from other headwords, usually by adding a suffix. Derived forms will be considered for headword status according to the same criteria as other items (frequency ≥ 5; see 5.1.1); if a derived form does not meet the criteria for selection, it will not be included in DOAE. One option that was considered was including rare derivatives (frequency < 5) under the related headword (a practice found in many NS dictionaries), but this was ultimately rejected because it would mean that DOAE would provide its users with (many) undefined words.

5.1.4.6 Vocabulary coverage

Vocabulary covered by DOAE will be dictated by CAJA. As the corpus contains academic articles, some vocabulary types are more likely to feature than others. Some domain-specific vocabulary will be covered as the corpus contains domain-specific journal articles. But DOAE will not offer a comprehensive coverage of the technical terminology of each domain (that is the role of technical dictionaries). However, the coverage could be later improved by creating domain-specific corpora and monitoring user searches.

5.1.4.7 Homographs

Homograph headwords are used for words with the same form, but with a different meaning, etymology, and/or pronunciation, and/or belonging to a different word class. Homographs will not be given headword status in DOAE; all homograph variants will be treated under one entry. This decision is based on the fact that meaning (rather than etymology, pronunciation, or word class) is nearly always consulted by the majority of students, so it should be the primary source of navigation through entries. To assist the users (in locating the sense and/or word class), menus will be used, a feature that has been successfully implemented by dictionaries for advanced learners of English.

5.1.4.8 Headwords from a single text

Some headwords, single-word or multi-word, may come from a single corpus text only, i.e. they will not be found in a range of texts. The current version of Sketch Engine does not provide this information in wordlists, so the concordance lines of each headword need to be examined. Yet, even if such information had been available, it would have been difficult to make decisions based solely on it. Headwords from a single text could be nonce words, used by the author(s) only for that particular text. On the other hand, such headwords could be technical terms used in a very narrow field, and limited to a single text due to a relatively small size of domain subcorpora.

The decision on whether to omit these headwords, or to treat them under other headwords, should be left to lexicographers who will probably need to consult other resources, such as subject experts, other dictionaries and even the web. It would be useful, however, if the lexicographers had the information on single-text headwords available before the analysis.

5.1.5 Dealing with candidate headwords

The headword list is by no means fixed. Some headwords, especially abbreviations, acronyms, and multi-word items (see 5.1.2), are very difficult to identify before the lexicographic work takes place. It is important to establish a procedure for how new candidate headwords will be dealt with. It may for example occur that two or more lexicographers identify the same multi-word item to be candidate headword when working on the entries of its constituent parts. Thus, to avoid duplication of work, lexicographers should initially make a record of any candidate headwords rather than immediately creating new entries. The editorial

team can then decide on the status of the candidate headwords, and assign them to lexicographers.

Figure 44. DOAE database: Entry Candidates - *significant other* under *significant*.

Headword:	☑ Incomplete
HeadwordSign	significant
Entry.candidate:	
Candidate.name	significant other
Explanation	
Notes	mainly in plural
Variants	
Example: 1	
Database.Domai	Anthropology
Example	Thus, it could be expected that family members, couples, and close friends might display their cl
Original.example	
Source	Anthrop_35_2006_stivers+robinson
Example: 2	
Database.Domai	Business and Management
Example	The entrepreneurs' household and significant others should also be taken into consideration.
Original.example	
Source	Business_21_2006_kolvereid+isaksen

In this Model, candidate headwords were recorded in the database under the element Entry.Candidates. Additional information can be provided, such as an explanation of meaning, notes, variants, and examples (see Figure 44).

5.1.6 Adding headwords not found in the corpus

The procedures and guidelines described so far address the majority of issues associated with building the headword list. The question yet to be answered is whether any items not found in the corpus should be included. Such items include the items belonging to a lexical set, technical vocabulary not found in the corpus, and the items found in the entries but not in the corpus.

5.1.6.1 Items belonging to a lexical set

A lexical set "denotes a group of words similar in meaning that belong to the same wordclass" (Atkins & Rundell, 2008:139). Examples include colours, months of the year, animals, capital cities, etc. A decision to include all the members of lexical sets, regardless of their presence or absence in the corpus, is unlikely to increase the headword list significantly. It would also go against the corpus-driven approach of DOAE. Besides, an 83.5-million-word

corpus will almost definitely feature all the members of common lexical sets such as days of the week or months of the year (and indeed, it does). Nonetheless, the editorial team should draw up a list of lexical sets that should feature in the dictionary, and check whether all the members have already been made headwords from the corpus list.

5.1.6.2 Technical vocabulary not found in the corpus

The corpus may not provide a comprehensive account of the technical vocabulary of every domain. One option would be to consult experts in the field and check for any glaring omissions. The arguments against this option are the corpus-driven approach and the purpose of the dictionary. Both arguments are driven by the idea that the corpus and the dictionary will represent the academic language that the users of the dictionary are most likely to encounter. Another argument, which is basically valid for all the items discussed in this section, is that limitations are needed to prevent the dictionary project spiralling out of control.

5.1.6.3 Items found in the entries but not in the corpus

One of the main principles of dictionary-making is that every word in a dictionary should be defined (Landau, 2001). This principle applies mainly to comprehensive NS dictionaries, however even dictionaries for foreign learners of English have attempted to avoid breaking it (e.g. by using defining vocabulary).

Whereas definitions that include words that are not found in the corpus can be avoided by setting guidelines, it is sometimes more difficult to avoid using such words in examples. One example of this problem is infrequent technical words that are likely to be surrounded by other technical words; the lexicographer then faces two issues: scarcity of examples as well as the vocabulary in the examples.

Inflected forms can also present a potential source of items not found in the corpus. What if, for example, the continuous forms of a verb are not found in the corpus, or are very infrequent? Should the inflected forms still be provided? The answer is 'yes', but the absence or low frequency of the continuous forms needs to be highlighted in the entry (e.g. by offering no or few examples with the continuous forms and/or adding note "rarely used in continuous forms").

As a general rule, the words that are not found in the corpus should not be given headword status in the dictionary. The focus of lexicographers should be on the words and phrases that the

users of the dictionary are likely to encounter. Besides, there is always a possibility of monitoring dictionary use and adding unsuccessfully searched headwords in subsequent editions of the dictionary (see 7.6).

5.1.7 Organizing the headword list

The headword list is usually organized alphabetically, but there are still decisions to be made as to whether to alphabetize word by word or letter by letter. For example, in a word-by-word headword list, *bank account, bank draft*, and *Bank of England* precede *banker* (see Figure 45, whereas in a letter-by-letter list, *banker* appears after *bank account* and *bank draft* but before *Bank of England*[82].

Figure 45. CED CD-ROM: Results for *bank*.

Atkins and Rundell (2008:190) argue that the issue of alphabetization "relates only to print dictionaries", and does not present problems "for editors or users of electronic dictionaries". This is clearly not the case because electronic dictionaries have limits in terms of the amount of information they can display on the screen at the same time. For example, the search *bank* in the CED CD-ROM (Figure 45) lists entries from *bank* {1} to *Banka*, which do

[82] Example entries were taken from CED CD-ROM.

not include *banker*. To see the three homograph entries for *banker*, the user needs to go to the next page in Index or open one of the entries at the bottom of the list.

Any decision of alphabetization will have its own problems, so it is essential to select one type of alphabetization and use it consistently. All efforts should then be made to enable users to identify the relevant sense and/or entry without any problems. TshwaneLex, the dictionary-writing software used in this thesis, comes with the default setting of 'table-based (alphabetical) sorting' which uses letter-by-letter alphabetization (Figure 46). This setting has been considered appropriate for the proposed Model.

Figure 46. TshwaneLex: Alphabetical ordering of headwords.

5.2 Accompanying material

Dictionaries normally contain accompanying material which can include the editor's address, instructions on how to use a dictionary, and material that is supposed to help the users with some other language-related activities (e.g. a section on how to write an essay or a CV).

The presentation of accompanying material depends on dictionary format. Print dictionaries usually offer accompanying material at the beginning or the end of the dictionary, and sometimes in the middle. The users therefore need to refer to those parts of the dictionary every time they need them. Electronic dictionaries have the advantage of making some accompanying material (e.g. instructions on how to use the dictionary) directly available whenever it is needed, namely when consulting the entry (i.e. a permanent 'Help' icon on every display). Accompanying material in the proposed Model is discussed in more detail in 7.5.

Other accompanying material such as a section on how to write an essay is not discussed in this thesis, as this type of material is not lexicographic in nature. Moreover, users accessing an online dictionary proposed by this Model will have access to the web anyway, where such useful material is in abundance. Nonetheless, to ensure that the relevant material is consulted, and quickly found, DOAE could offer hyperlinks to approved external material of this nature, as opposed to developing its own.

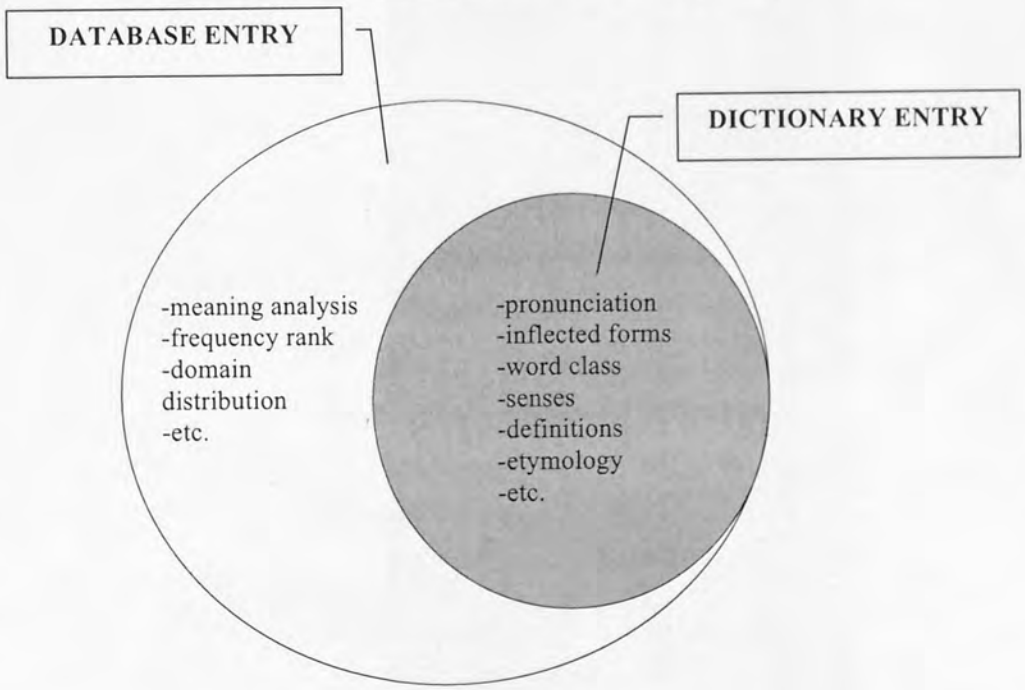
6. MICROSTRUCTURE OF THE MODEL FOR DOAE

This chapter addresses the dictionary microstructure, that is the contents of the dictionary entries. But before dictionary entries can be compiled, the analysis of corpus data needs to be conducted, and various types of information recorded. The chapter therefore also depicts the process of analysis in detail, and discusses which information is recorded, and how.

As shown by the literature on student dictionary use, as well as the survey conducted for the purposes of this thesis, students consult certain microstructural features more often than other features. Nonetheless, no feature is completely ignored by all the students. This means that the database of DOAE should contain all the different types of information, even the ones which are rarely consulted by students (e.g. etymology).

However, a great deal of information recorded during the analysis (e.g. candidate examples, and database labels) is not offered to the dictionary users, but is essential for lexicographers when compiling the entry, or modifying it for different dictionary formats. It is therefore important at this point to make a distinction between a database entry, which contains all the information about the headword recorded by lexicographers, and a dictionary entry, which is the parts of the database entry which are made available to users (Figure 47).

Figure 47. DOAE: Database entry vs. dictionary entry.



In addition to deciding which types of information should be recorded in the database, it is also important to consider how the information will be recorded in order to make it easy to manipulate (e.g. to ensure quick importing of corpus data into the database), and to make it as customisable as possible when it comes to dictionary output.

The part of the database that describes what information can be recorded, and how, is DTD (see 3.3.2). It is better to more or less finalize the DTD before assigning the work to lexicographers as any (major) changes to the DTD will affect the work of anyone working on the dictionary database, and may even result in the loss of data. In fact, a dictionary model like the one presented in this thesis is an appropriate opportunity to design the dictionary DTD.

The DTD of the Model for DOAE (provided on the attached CD-ROM, Appendix 13) is based on sample entries built for the Model. The design of the DTD was driven by two main principles; the first focused on the needs of lexicographers and aimed to make the recording of information in the database as quick as possible. The second attended to the needs of the users, making the presentation and customisability of dictionary entries as efficient and user-friendly as possible.

To ensure that the Model and its DTD included as many different types of information as possible, it was important to select a variety of sample entries, namely entries that differed in characteristics such as frequency, form, word class, etc.

6.1 DOAE sample entries

In total, 38 sample entries were built (see Table 42 below). The complete sample database entries are provided on the attached CD-ROM (Appendix 13). As mentioned, the selection of sample entries could not be completely random; it had to ensure that a variety of different types of entry were covered. As shown in Table 43 below, the selected entries thus differ in terms of:

- Frequency rank (this was the main criterion for the selection of sample entries). More than a third of sample entries are among the top 1000 most frequent words in CAJA. On the other hand, nearly a quarter of sample entries, mostly multi-word headwords, rank between 10,000-100,000 among the most frequent lemmas in CAJA.
- Form. There are 28 single-word headwords, and 10 multi-word headwords.
- Word class. There are 21 nouns, 6 verbs, 7 adjectives, 4 adverbs, 2 abbreviations, and 1 conjunction. 5 headwords contain two word classes.

Table 42. DOAE: Sample entries.

albeit	justify
analysis of variance	local authority
ANOVA	method
argue	obtain
assortment	potential
attribute	ribonucleic acid
authority	RNA
CEO	significant
chief executive	significant other
chief executive officer	state-of-the-art
electric potential	subsequent
et al.	subsequently
etc.	take*
et cetera	taken
FACT	therefore
fact	thus
feature	thusly
feature film	took
justified	various

* - the noun sub-entry for *take* was created in full, while the verb sub-entry was mainly used to demonstrate how to identify phrasal verbs

The sample entries also differ in terms of level of polysemy; for example, headwords such as *fact* and *argue* have many different meanings, whereas headwords such as *various* and *albeit* have only few meanings or only a single meaning. In addition, some of the headwords have or contain technical meanings. Level of polysemy and technical meanings of headwords, however, were established only after the analysis.

It is worth pointing out that the sample dictionary entries devised for this thesis are not complete – certain features such as cross-references and synonyms are missing because these features can only be added once all the dictionary entries are compiled.

The number of entries could be considered small by lexicographic standards, but the aim was to strike the balance between breadth and depth of the analysis. It should be noted that the process of compiling entries was cyclical – almost every entry introduced a change to the DTD, for example a new element or attribute, which meant that the new type of information was missing in all the previously compiled entries, and had to be added at that point. In addition, a considerable amount of time was devoted to designing the entry output, and making the entry information as customisable as possible.

Table 43. DOAE: Sample entries - corpus frequency, rank, and word class(es).

sample entry	raw frequency (CAJA)	frequency per million words	rank	word class
et al.	73,317	785	1-500	abbreviation
take[a]	70,165	752		verb
thus	65,488	702		adverb
method	50,067	536		noun
significant	47,839	512		adjective
therefore	43,255	463		adverb
obtain	42,160	452		verb
fact	40,722	436		noun
argue	27,635	296		verb
feature	24,979	268		noun
various	24,638	264		adjective
potential	16,766	180	500-1000	adjective
authority	13,219	142		noun
potential	12,026	129		noun
subsequent	9,779	105	1000-2000	adjective
attribute	6,952	75		verb
RNA	6,345	68		noun
attribute	6,053	65		noun
etc.	5,721	61		abbreviation
subsequently	5,331	57		adverb
justified	2,825	30	2000-3000	adjective
feature	2,690	29		verb
justify[a]	2,688	29		verb
ANOVA	2,641	28		noun
CEO	2,284	24	3000-4000	noun
albeit	2,060	22		conjunction
analysis of variance	1,098	12	4000-10,000	noun
local authority	1,091	12		noun
state-of-the-art[b]	363	4	10,000-50,000	adjective, noun
assortment	349	4		noun
chief executive	165	2		noun
significant other	133	1		noun
take	95	1		noun
chief executive officer	80	<1		noun
FACT	71	<1		noun
electric potential	69	<1		noun
feature film	62	<1		noun
et cetera	29	<1	50,000-100,000	/
ribonucleic acid	17	<1		noun
thusly	15	<1		adverb

[a] – includes occurrences of the headword *taken* (adjective)
[b] – excludes occurrences of the headword *justified*
[c] – includes 117 occurrences of the variant *state of the art*

6.2 DOAE database: Some key microstructural elements

Most microstructural elements will be discussed in detail at the stages at which they are added. Some elements however need to be addressed immediately as they are used in various stages of the analysis, and for different purposes.

6.2.1 Domain labels

Students, as well as being categorised by their native language, can also be categorised by their subject of study. Students of a particular subject may be more interested in meanings, phrases, collocates, etc. that are more relevant to their subject of study (i.e. domain).

Lexicographers use domain labels to help the users to distinguish between word senses, especially "when a word is used in two or more different disciplines with different meanings, or if it is used in one sense technically and in another popularly..." (Landau, 2001:226). This dictionary Model, uses domain labels not only to mark headwords, senses, subsenses, phrases and other information made available to the user, but also to mark information that is used by lexicographers during the analysis (e.g. domain distribution of word classes of the headword, domain distribution of collocates) – these labels for labelling domain distribution are called database domain labels. Furthermore, the same domain labels are used for designing user-friendly features such as domain-customisable sense ordering.

Domain labels used in this dictionary Model are based on the subcorpora in CAJA. In fact, the design of CAJA envisaged the potential of detailed subcorpus division for domain labelling (see 3.2.1.1.3). This approach follows the idea proposed by Landau (2001) who believed that subcorpora of the corpus that a dictionary is based on could provide a more accurate and reliable set of domain categories.

Because domain categories represented by subcorpora are rather narrow, they can be grouped into broader categories. Using a hierarchical structure for domain labels allows vocabulary items to be marked more accurately (Atkins & Rundell, 2008). For example, a sense or collocate may be shared by several (neighbouring) domains, and using a single broader label is more user-friendly than providing a label for each domain.

This Model adopts a three-level classification (see Table 44) – the 28 labels at the lowest level (Level 3) are based on the 28 domain subcorpora in CAJA[83], 6 labels at the medium level

[83] Level 3 labels are not exactly equivalent to the subcorpora in the CAJA corpus; seven labels have been shortened due to the length of the subcorpus name: subcorpus name 'Business and Management' has been shortened to the

(Level 2) are names for groups of closely related domains, and the 2 labels at the top level (Level 1) correspond to two broadest categories of domains.

Table 44. DOAE database: domain labels.

Level 3 label	Level 2 label	Level 1 label
Architecture	ARTS	ARTS AND HUMANITIES
Arts and Art History		
Linguistics		
Music		
Archaeology	HUMANITIES	
History		
Philosophy		
Religion		
Business	BUSINESS SCIENCES	
Economics		
Finance		
Law		
Anthropology	SOCIAL SCIENCES	
Education		
Politics		
Psychology		
Sociology		
Computing	APPLIED SCIENCES	SCIENCES
Engineering		
Mathematics		
Physics		
Biochemistry	LIFE SCIENCES	
Biology		
Chemistry		
Geography		
Medicine		
Sports		
Veterinary Science		

Level 2 categories were formed by consulting various resources: websites of nine UK universities, the Dewey Decimal system, and categories in some of the existing corpora of academic discourse (e.g. BASE, BAWE). The aim was to design between five and eight Level 2 categories. The categorization of certain subjects is problematic, but that is unavoidable since no two systems use the same classification.

---

label 'Business', 'Computer Science' to 'Computing', 'Geography, Earth and Environmental Studies' to 'Geography', 'Medicine and Health Sciences' to 'Medicine', 'Politics, Government & International Relations' to 'Politics' 'Social Sciences' to 'Sociology', and 'Theology and Religion' to 'Religion'.

More specific labels than Level 3 labels are sometimes necessary or more suitable (e.g. *Grammar* under *Linguistics*). These labels are not part of the corpus design, so they need be added to the entries (and the list of labels) during the dictionary-making process[84]. The editorial team must then ensure that the labels added during the analysis are used consistently throughout the dictionary.

6.2.2 Other labels

Many other types of label are likely to feature in DOAE, and while not all of them are found in this Model, it is important to discuss their use in the dictionary:

a) Regional labels tell the user that an item is predominantly used in a particular language variety. Regional labels are expected to occur frequently in DOAE, especially labels *American English* and *British English*, as one of these two language varieties is used by most publishers of journals found in CAJA (see also 3.2.1.1.2).

b) Frequency labels, such as *usually* and *often,* indicate a degree of use of the item, and are most often used in combination with other labels. Because of the abundance of frequency information made available by CAJA, frequency labels can be made even more specific or, alternatively, can be replaced, or supported, by frequency graphs (see 6.3.3.9.5).

c) Register labels show that the use of the item indicates a specific manner of speech or writing. Register labels include *informal*, *formal*, *spoken,* and *written.* This particular dictionary Model, if based solely on CAJA, would not include many register labels, as the corpus contains only written academic texts. But since consulting other academic corpora has been made part of the analysis, the register element (especially the distinction between spoken and written discourse) can be made at that stage.

d) Attitude labels such as *pejorative, derogatory, ironically*, *humorous* refer to a writer's attitude. According to Atkins and Rundell (2008), these labels occur in most learner's dictionaries, whereas in dictionaries for NSs this kind of information is often included in the definition text. Since DOAE is in some way a learner's dictionary, it should include attitude labels. But the use of labels such as *pejorative* and *derogatory* may not be user-friendly as the users may not be familiar with the meaning of those words. So in some cases including this information in the definition may be more user-friendly (e.g.

[84] These types of label have been listed under the element Subdomain labels in the database of this model.

to say something *in a critical way)*, especially given the fact that the definition is the most frequently consulted part of the entry.

Labels unlikely to occur in DOAE, mainly due to the nature of the text in CAJA, are:

a) Certain register labels, such as *slang, offensive, vulgar* and *taboo*[85]. Slang, offensive, vulgar, and taboo items are unlikely to be found in the CAJA texts (or in academic discourse in general).
b) Meaning labels are used to make a distinction between literal and figurative meanings of the item. Meaning labels are used where "the sense shift is not so well established as to constitute a new lexical unit" (Atkins & Rundell, 2008:230).
c) Style labels such as *poetic, journalism, literary*, which are used to indicate that the item has a specific use in specific functional variety of language.
d) Temporal labels such as *archaic*, *dated, old-fashioned*, which are predominantly used for words or word meanings that are no longer or rarely used.

The use of labels will therefore largely be determined by the specific uses found in the corpus. But despite the difference in how frequently they occur in the dictionary, all labels have the same importance; in fact, rarely used labels, such as *taboo*, have an even more important role because the users need to be made aware of these words or senses of words "to avoid the embarrassment of using them inadvertently" (Landau, 2001:230).

6.2.3 Grammar information

Grammar information is found throughout the database, and ranges from very basic word class information to complex grammatical relations of the headword. Some grammar information is recorded in the database purely to assist the lexicographers when compiling dictionary entries; thus, the database entries contain far more grammar information than the dictionary entries.

The decision about which grammar information to present in the dictionary should be based on the needs and language proficiency of students. In my main survey, students reported that they only consulted grammar information explicitly (see 4.1.1.7), confirming the findings of many past studies (Béjoint, 1981; Harvey & Yuill, 1997; Nesi, 2000; Bogaards & van der Kloot, 2001). Instead, the studies suggest that students often look in examples and definitions

[85] There is no unanimity in the categorization of *slang* and *taboo* labels in the literature. For example, Atkins and Rundell (2008) and Svensén (1993) treat them as subsets of register labels, whereas Landau (2001) considers them as completely separate categories.

for grammar information (Harvey & Yuill, 1997; Bogaards & van der Kloot, 2002; Dziemianko, 2006).

Though NNS students consult grammar information more frequently, in this dictionary Model the view is taken that neither NS or NNS students know a great deal about grammar, the view which is based partly on the evidence that students have difficulties with grammar (Nesi & Haill, 2002), and partly on personal experience (my own and informal reports from colleagues[86]). Consequently, the amount of grammar information offered in the dictionary is limited. Grammar information is not considered an essential part of the entry; this approach is reflected in macrostructural guidelines such as treating homographs under a single entry, and in (frequent) use of full-sentence definitions.

Grammar information in the dictionary entries can be divided into three layers: word class markers, grammar labels, and constructions[87]. Word class markers show the word class of the headword (e.g. noun, adjective), grammar labels provide sub-categorization within the word class (e.g. *intransitive* and *transitive* for verbs, *countable* and *uncountable* for nouns) and any additional grammar information (e.g. *only before noun*), and constructions present information about the frequent syntactic patterns of the word. Constructions do not contain grammar codes, or abbreviations (e.g. *sb* for *somebody*)

This three-layer division of grammar information is also used when recording the information in the database. This is done to allow flexibility as to when and how grammar information is presented to different groups of users (e.g. NNSs and NSs), or individual users.

6.2.4 Examples

Examples are found throughout the database, and their role is to provide illustrations of real language use. All the examples are taken from the corpus; invented examples are not used at any point because "(h)owever plausible an invented example might be, it cannot be offered as a genuine instance of language in use" (Sinclair, 1991:4).

The majority of examples are recorded during the meaning analysis; among those examples, a selection of examples is made for the database entry. Some database examples need to be slightly modified to make them suitable for the dictionary entry. The selection of examples for the dictionary entry and related issues are discussed in 6.3.3.4.

---

[86] Many university lecturers in the School of Language and Social Sciences at Aston University that teach linguistics modules mentioned that their students have difficulties even in identifying basic word classes.
[87] The term 'construction' was adopted from Atkins and Rundell (2008).

All examples are saved in full-sentence format. Sometimes, more context (sentences before and/or after) needs to be provided, for example if the meaning/function of the headword cannot be exemplified by a single sentence (e.g. *therefore*).

Examples are recorded either semi-automatically or manually. The semi-automatic approach utilizes the TickBox Lexicography function in Word Sketch, and involves selecting the examples, exporting them into XML form, and then importing them into the database (see Figure 100 in Appendix 9). Domain label, text of example, and source file id are saved automatically. The manual approach consists of copying the example and name of the source file in Concordance, and pasting the information into the Example element in the database. Then, the database domain label needs to be selected from the drop-down menu.

Each example is recorded in the database with the following type of information (see Figure 48 and Figure 49):

a) **Number**. Assigned automatically. Used in the database to inform lexicographers of a number of examples for a sense, subsense, meaning pattern, etc.

b) **Bullet point**. Provided automatically. It is a constituent part of example display in the dictionary entry.

c) **Database domain label**. Selected from the 28 L3 domain labels (drop-down menu). The database automatically adds L2 and L1 domain labels (e.g. *Arts and Humanities: Arts* is prefixed to *Architecture* – see Figure 49 below).

d) **Example**. Contains the text of the example. If an original text is modified in any way, then the modified text is saved here.

e) **Original example**. Optional information which is only used when the original text is modified. Modifications to examples are normally made when compiling the dictionary entry.

f) **Source**. Filename of the corpus text from the example was taken.

Figure 48. DOAE database: An example in the entry for *obtain*.

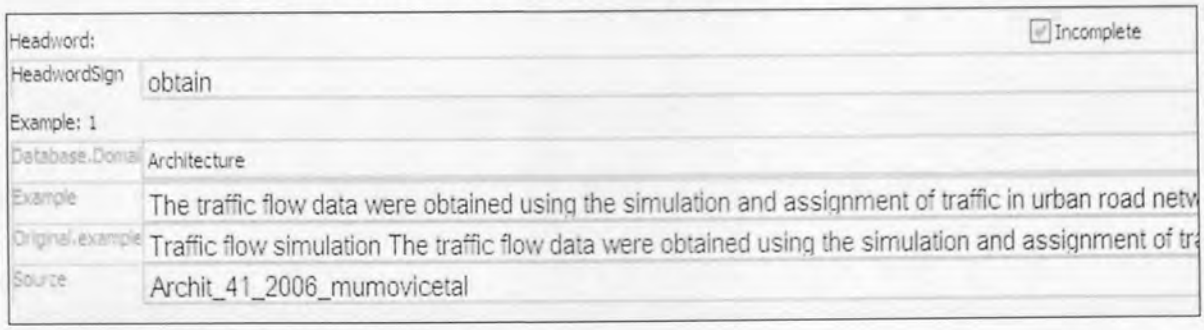


Figure 49. DOAE database: An example of the verb *obtain* saved in the database (XML format).

<Example>
<Example.Example.number>**1**</Example.Example.number>
<Example.Database.DomainLabel>**Arts and Humanities: Arts: Architecture**</Example.Database.DomainLabel>
<Example.Example>**The traffic flow data were obtained using the simulation and assignment of traffic in urban road networks (SATURN) traffic flow simulation software.**</Example.Example>
<Example.Original.example>**Traffic flow simulation The traffic flow data were obtained using the simulation and assignment of traffic in urban road networks (SATURN) traffic flow simulation software.**</Example.Original.example>
</Example>

6.3 DOAE: Database entry

The creation of a database entry (the sample database DOAE entries are provided on the CD-ROM; Appendix 13) consists of the following stages:

a) **Recording basic information**: frequency, pronunciation, inflected forms, and domain distribution of the headword. This is done as a separate stage because most of this process can be done semi-automatically, and not necessarily by lexicographers.

b) **Meaning analysis** focuses on identifying meanings or meaning patterns, recording related meaning explanations, grammatical relations, collocates, and examples.

c) **Compiling dictionary entry** involves writing definitions and selecting the appropriate information from the database (e.g. examples, labels) that will be included in the dictionary.

The order of the three stages also represents the order in which information is recorded during the analysis. Each stage is now discussed in turn.

6.3.1 Recording basic information

This stage involves recording the following types of information:

- word class
- frequency rank (rank on the lemma list)
- frequency (per million words)
- inflected forms
- pronunciation
- domain distribution and
- etymology.

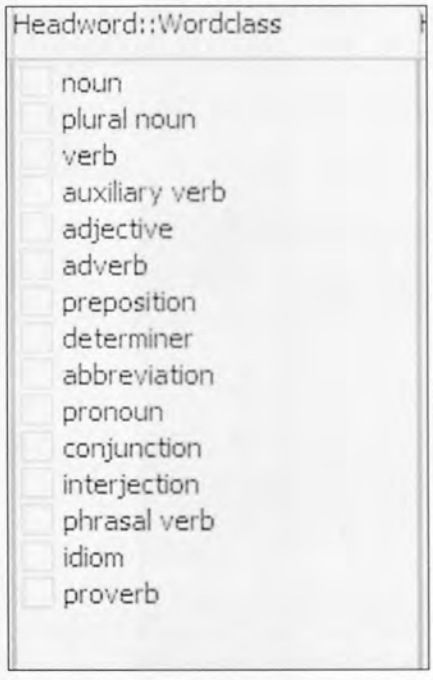
The information is obtained from the CAJA corpus in Sketch Engine, with the exception of pronunciation and etymology, which are obtained from existing dictionaries.

6.3.1.1 Word class

Word class information is an inherent part of a dictionary entry. Other than playing an informative role, it can also assist the user in navigating in the entry, or selecting the relevant homograph headword. Dictionaries are very similar in the way they present word class information, and in the categories they use.

A list of word classes, which can be modified (i.e. word class categories can be added or omitted), is provided in the database in the form of a menu (Figure 50). The list is available at headword level, sub-entry level, and even at (sub)sense level, which explains why it includes items such as 'idiom' and 'proverb' that traditionally have headword status.

Figure 50. DOAE database: List of available word classes.



In Sketch Engine, word class information about the headword is obtained by performing a concordance search, and selecting the Node tag option in the Frequency function[88]. This presents us with a list of POS-tags for the headword, and their frequency (see e.g. Figure 51).

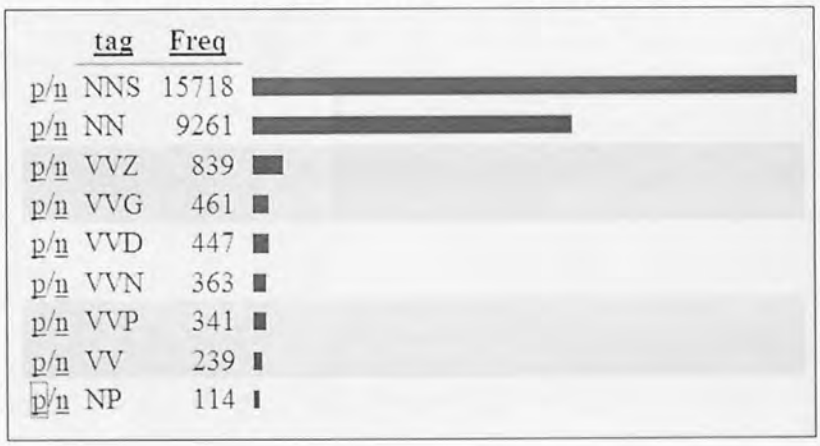
[88] An alternative would be to get word class information from lemmapos wordlist (lemmatized wordlist where lemmas are divided by part of speech), however using the Frequency function is more efficient as it provides the information on both word classes and inflected forms. Also, the analysis of lemmapos wordlist requires searching for lemmas among several error-lemmas, so the lexicographers may decide to rely on their intuition and potentially miss some word classes by not looking for them.

POS-tags specify a sub-category of word class rather than just word class (e.g. nouns are represented by tags for singular or mass noun, plural noun, singular proper noun, and plural proper noun), so sub-categories of the same word class need to be grouped together when determining word classes of the headword.

Headwords used as a single word class, especially grammatical words, do not have a long list of tags to analyse. For example, all occurrences of THUS were tagged with RB (see Table 102 in Appendix 4 for full list of tags and their descriptions), the tag for adverb. So, in the entry for *thus*, the box 'adverb' in the list of options for headword word class was ticked.

When a headword has more than one word class, sub-entries need to be created in the database. For example, the list of tags for FEATURE, shown in Figure 51, indicated that FEATURE was used as a noun and a verb. Thus, two sub-entries, one for each word class, were created. The information about the frequency of the tags for this particular headword was very useful (especially the graphic presentation) because it helped determine the order of sub-entries (e.g. noun uses of FEATURE were much more frequent).

Figure 51. DOAE: Recording basic information – POS-tags for lemma FEATURE.



The order of sub-entries sometimes needs to be determined on the basis of domain distribution of each word class, as well as its frequency. This is particularly relevant for headwords that display a more even frequency distribution of word classes. An example of such a headword is ATTRIBUTE, which was only slightly more frequently used as a verb (53% of all occurrences) than as a noun (47% of all occurrences) (see Table 45 below). Based purely on frequency, the verb sub-entry would be offered first, which is a user-friendly solution for students of most subjects (Table 46 below), but not for students of Business and Management, Economics, Computer Science, Psychology, Architecture, Finance, or Mathematics (in bold italics in Table 46), who are (much) more likely to encounter ATTRIBUTE as a noun.

Table 45. DOAE: Recording basic information – Node tags (by word class) for lemma ATTRIBUTE.

NOUN	
tag	**Frequency**
NNS	3830
NN	2224
NP	19
JJ*	10
TOTAL	**6083**

* - tagging errors

VERB	
tag	**Frequency**
VVN	3561
VVD	1243
VVZ	717
VVP	578
VV	512
VVG	341
TOTAL	**6952**

Table 46. DOAE: Recording basic information - Verb uses and noun uses of ATTRIBUTE in the 28 domain subcorpora.

	verb uses of ATTRIBUTE (%)	noun uses of ATTRIBUTE (%)
Business and Management	***20***	***80***
Economics	***24***	***76***
Computer Science	***12***	***88***
Philosophy	59	41
Psychology	***48***	***52***
Theology and Religion	65	35
Archaeology	66	34
Linguistics	78	22
Geography	65	35
Arts and Art History	60	40
Architecture	***28***	***72***
Anthropology	67	33
Finance	***49***	***51***
Music	71	29
History	78	22
Sociology	59	41
Education	55	45
Chemistry	94	6
Engineering	63	37
Law	73	27
Politics	70	30
Physics	98	2
Sports	78	22
Mathematics	***27***	***73***
Biology	70	30
Medicine	74	26
Biochemistry	83	17
Veterinary Science	84	16

The solution proposed for this Model is to record the frequency distribution of word classes across domains for each headword in the database, and design as many different variants of sub-entry order as the distribution indicates. Each variant of sub-entry order would then be labelled with all the relevant domain labels (level 2 and level 1 domain categories could be used as well), so that it would be made available to students of those subjects.

On the lists of tags for FEATURE and ATTRIBUTE, there were some tags that seem to point to unexpected uses of the two headwords. The tags included NP tag (proper noun in singular) at both FEATURE and ATTRIBUTE, and JJ tag (adjective) at ATTRIBUTE only. A closer inspection of NP concordance lines for FEATURE (see Table 116 in Appendix 7 for a random sample of 20 concordance lines) showed that in all of them FEATURE was a singular noun written with a capital letter (e.g. *...table 5, Feature E...*), and often used in compounds (e.g. *...Web Feature Service...*). Similarly, in all the concordance lines of NP and JJ, ATTRIBUTE was actually used as a noun (see Table 117 and Table 118). These were thus examples of incorrect tagging, so no sub-entries were created for these word classes.

6.3.1.2 Frequency rank

Information on the rank of the headword in the lemma list (ordered by frequency) is valuable information to record in the database. The information allows us to label the headwords according to their occurrence among the top 1000, 2000, or n-thousand words, if we wish. Moreover, frequency rank can be used as a deciding criterion when reducing the headword list, for example for a compact version of the dictionary.

Multi-word headwords are not provided with a frequency rank as they do not feature in the lemma list. It is therefore useful to record an additional, different type of frequency information for the headwords, one which can be provided for every single headword. This type of frequency information is discussed next.

6.3.1.3 Frequency

The frequency of a headword is provided as the average number of occurrences per million words. The CAJA word count in Sketch Engine (93.5 million words), as opposed to the word count obtained by using the program written in Java (83.5), was used when making calculations. This was necessary because frequency ranks and other frequency information

about the headwords and related items (e.g. collocates) were already based on the frequency information produced by Sketch Engine.

The attribute in TshwaneLex that contains frequency information accepts only integer values, so the frequency cannot be expressed with decimal numbers (any decimal numbers are rounded to the lower whole number, e.g. 3.45 to 3). This is problematic for the less frequent headwords. For example, headwords or sub-entries with a frequency of less than 1 per million words were all assigned the frequency of 0 per million words, yet, for these headwords, which are likely to be candidates for omission in smaller versions of the dictionary, it would be particularly useful to have more specific frequency information. The problem has been addressed by using a minus sign instead of a decimal point; for example, the frequency of *thusly* (0.16 per million words) was recorded as -16 in the database.

If the headword had more than one word class/sub-entry, the frequency is provided for each sub-entry, as well as for the headword. This can assist the lexicographer when preparing a compact version of the dictionary, so that not only entries, but also sub-entries can be considered for omission.

6.3.1.4 Inflected forms

Regular and irregular inflected forms of the headword were recorded in the database DOAE entries. Variant inflected forms (e.g. American spellings) were also recorded, and regional labels were provided.

Inflected forms of the headword were identified by conducting a lemma search for each identified word class of the headword, and selecting the Node forms function in the Frequency button in Concordance. The results for the verb and noun *attribute* are shown in Table 47 and Table 48 respectively. It was evident that all the inflected forms of both verb and noun occur, so they were recorded in the relevant sub-entries.

Table 47. DOAE: Recording basic information – Node forms of *attribute* (verb).

Word	Frequency	%
attributed	4801	69
attribute	1089	15.7
attributes	716	10.3
attributing	330	4.8
Attributing	11	<1
Attributed	3	<1
Attributes	1	<1
Attribute	1	<1
Total	6952	100

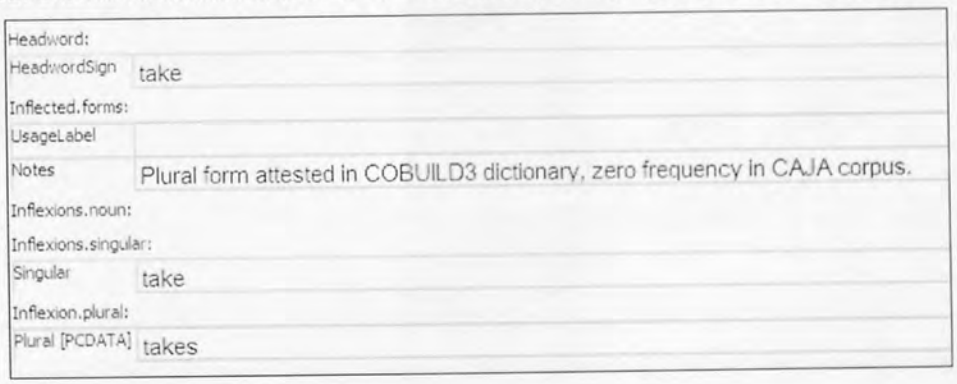
Table 48. DOAE: Recording basic information – Node forms of *attribute* (noun).

Word	Frequency	%
attributes	3781	62.5
attribute	2196	36.8
Attributes	47	<1
Attribute	28	<1
ATTRIBUTES	1	<1
Total	6053	100

Any noticeable disparities in the frequency distribution of inflected forms need to be recorded in the database, as such information is of importance for the processes of definition writing and example selection. The verb *attribute* was in need of such a note because the frequency distribution of inflected forms showed that *attributed* accounts for 69% of all verb uses.

In some cases, certain inflected forms may be very rare or may not be found in the corpus. Such cases require checking another (general) corpus and/or dictionary. The missing inflected forms are still recorded in the database, along with a note. Figure 52 shows a note for the plural form of the noun *take*, which was not found in CAJA.

Figure 52. DOAE database: Inflected forms of *take* (noun), and a related note in the database.



If any irregular inflected forms are identified, one additional step is needed, namely checking whether those inflected forms have been given headword status.

6.3.1.5 Pronunciation

Pronunciation can be presented in a dictionary in two different forms, written (transcribed) and spoken (audio). Both forms are recorded in the database of DOAE. For practical reasons, existing dictionaries were used as sources of written pronunciation. The IPA

pronunciation has been copied from CED CD-ROM, phonemic respelling from MWCD CD-ROM, and non-phonemic respelling from Dictionary.com.

The analysis of the dictionaries used by or targeted at students has revealed that the dictionaries use one of three spelling systems[89]: phonetic transcription (normally using the International Phonetic Alphabet (IPA) symbols), phonemic respelling, or non-phonemic respelling[90] (see Table 49 for an example). In dictionaries consulted for the purposes of this thesis, phonetic transcription is used by CED CD-ROM, most dictionaries for foreign learners (e-CALD, e-MED, e-OALD), and the only dictionary for NNS students (LED CD-ROM). The phonemic respelling system is used by American dictionaries (i.e. MWCD CD-ROM). Non-phonemic spelling system is found only in Dictionary.com (named 'spelled pronunciation' in the dictionary).

Table 49. Three spelling systems: the word *jongleur*.

jongleur (French) /ʒɔ̃glœr/	IPA – CED CD-ROM
Main Entry: **jon·gleur** Pronunciation: zhō^{n}-'glər	Phonemic respelling – MWCD CD-ROM
jon-gleur [jong-gler;	Non-phonemic respelling – Dictionary.com

Each of these three spelling systems has its drawbacks. The IPA and phonemic respelling system introduce many new symbols that users need to familiarize themselves with in order to use the system effectively. Dictionaries using these systems must therefore provide a pronunciation guide for the users. A non-phonemic respelling system, on the other hand, "relies on the familiarity of the spelling-to-sound rules rather than on their systematicity", but "due to the unsystematic nature of English spelling, the English spelling system is simply not up to the task of representing the pronunciation of words unambiguously" (Fraser, 1997:184).

University students are a very heterogeneous group so all three written pronunciation systems are recorded in the database. Then, the choice of the pronunciation system can be left to the user. A similar method is already used by Dictionary.com, which offers the user the option

[89] COBUILD CD-ROM offers only audio pronunciation only, e-LDOCE offers audio pronunciation only for headwords and examples for words beginning with D and S, and NODE CD-ROM does not offer any pronunciation.

[90] The term "non-phonemic respelling system" is used by Fraser (1997) for a respelling system that does not attempt to use one symbol for one sound as phonemic respelling system does.

to switch between two written pronunciations (the IPA and Spelled pronunciation); in addition, audio pronunciation is available (Figure 53).

Figure 53. Pronunciations offered by Dictionary.com (part of the entry *law*).

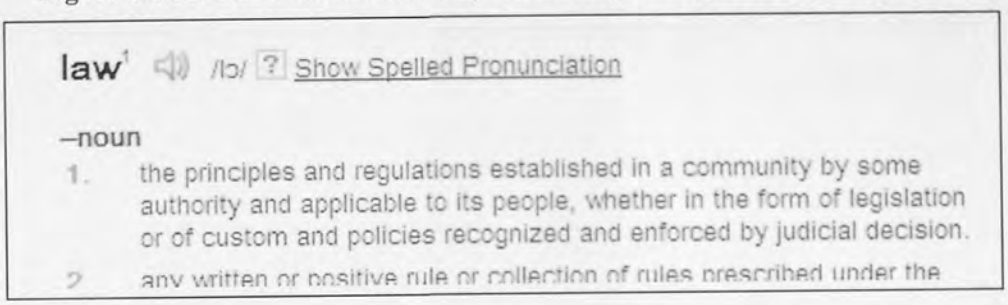
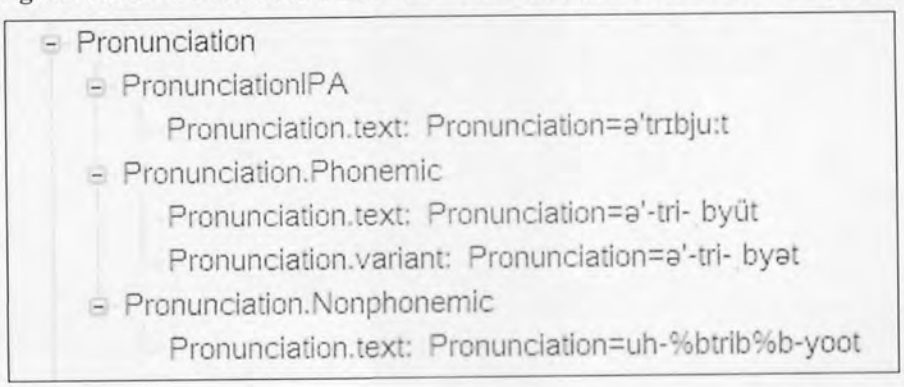


Figure 54. DOAE database: Pronunciation information for *attribute* (verb).



The pronunciation offered in the source dictionaries contains several additional features, such as the use of stress and pronunciation variants (e.g. see Figure 54), which could have been excluded for the purposes of this dictionary. Nevertheless, these features, especially stress, play a vital role in showing the user how to pronounce a word, and in some cases, how to distinguish between two homographs. Similarly, pronunciation variants (including pronunciations used in different varieties of English) point to alternative pronunciations which the users may be familiar with. Full pronunciation variants rather than abbreviated ones were recorded[91].

Audio pronunciation is even more important than written pronunciation, as any user interested in pronunciation essentially wants to **hear** rather than read how the word is pronounced. Audio pronunciation can be recorded in the database, as the TshwaneLex software offers the option of saving audio files[92].

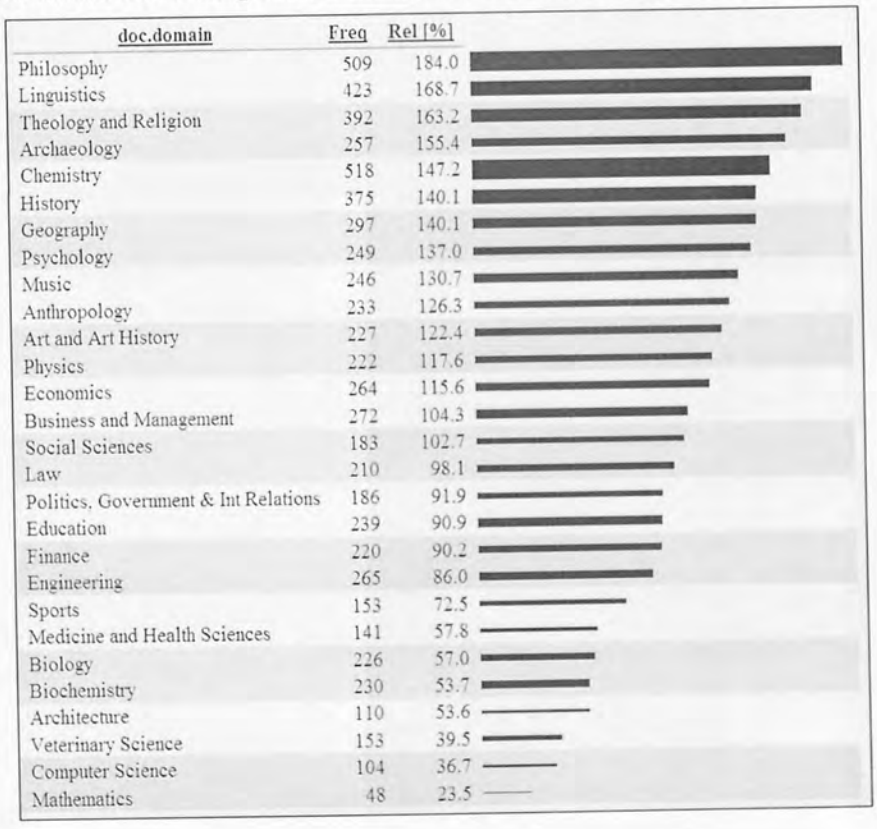
[91] Many dictionaries omit from the variant form the part of the pronunciation that is common to the main pronunciation form and the variant form, and replace it with a hyphen.
[92] Audio pronunciations for a few sample entries have been recorded in the database for exemplification purposes.

6.3.1.6 Domain distribution

The domain distribution is the frequency distribution of the headword, or one of its word classes, across the 28 domain subcorpora in CAJA. Domain distribution information has been useful for domain labelling, sense ordering, and example selection. The information was obtained by using the Doc IDs option of the Frequency function in Concordance. As TshwaneLex allows images to be saved in the database, a screenshot of the domain distribution was made and saved in the database entry.

It is important to order the domains by relative frequency rather than by raw frequency, because the CAJA subcorpora vary in size. For example, in the domain distribution of the verb *attribute* (Figure 55), Chemistry was the domain with the highest raw frequency, but was only fifth in order of relative frequency.

Figure 55. DOAE: Recording basic information - Domain distribution of *attribute* (verb).



doc.domain	Freq	Rel [%]
Philosophy	509	184.0
Linguistics	423	168.7
Theology and Religion	392	163.2
Archaeology	257	155.4
Chemistry	518	147.2
History	375	140.1
Geography	297	140.1
Psychology	249	137.0
Music	246	130.7
Anthropology	233	126.3
Art and Art History	227	122.4
Physics	222	117.6
Economics	264	115.6
Business and Management	272	104.3
Social Sciences	183	102.7
Law	210	98.1
Politics, Government & Int Relations	186	91.9
Education	239	90.9
Finance	220	90.2
Engineering	265	86.0
Sports	153	72.5
Medicine and Health Sciences	141	57.8
Biology	226	57.0
Biochemistry	230	53.7
Architecture	110	53.6
Veterinary Science	153	39.5
Computer Science	104	36.7
Mathematics	48	23.5

Screenshots of domain distribution are a very useful means to identify uneven domain distribution patterns. For example, the domain distribution of the noun *attribute* (Figure 56) indicated that almost a half (2772 out of 6053) of all occurrences were limited to three domains. Such departures from even distribution were noted down in the database entry.

Domain distribution diagrams of different word classes of the same headword can be compared when compiling the dictionary entry. An additional observation in the case of the headword *ATTRIBUTE* was that there is a considerable difference in the domains in which the verb use predominates, and the domains in which the noun use predominates.

Figure 56. DOAE: Recording basic information – Domain distribution of *attribute* (noun).

doc.domain	Freq	Rel [%]
Business and Management	1080	475.6
Economics	825	414.7
Computer Science	867	351.8
Psychology	266	168.0
Architecture	282	157.9
Philosophy	351	145.7
Finance	230	108.3
Theology and Religion	209	100.0
Art and Art History	152	94.1
Archaeology	130	90.3
Geography	160	86.7
Education	191	83.4
Social Sciences	121	78.0
Mathematics	131	73.5
Anthropology	114	71.0
Music	101	61.6
Engineering	154	57.4
Linguistics	122	55.9
History	106	45.5
Politics, Government & Int Relations	80	45.4
Law	78	41.8
Biology	96	27.8
Medicine and Health Sciences	50	23.6
Sports	42	22.9
Biochemistry	48	12.9
Chemistry	33	10.8
Veterinary Science	29	8.6
Physics	5	3.0

Higher-level domain groupings also need to be considered, as some headwords may be more common to a group of subjects. As shown in Figure 57 below, Arts and Humanities contained 97% of all the occurrences of AUTHORITY, whereas the Sciences domains contained only 3% of all occurrences and occupied the bottom 9 positions on the frequency list (aside from Geography which is in the bottom 13); in addition, Physics was not even on the list due to 0 occurrences of AUTHORITY. On the basis of this evidence, the database label 'Arts and Humanities' was added to the entry.

Note that any labels added to the database entry, or one of its sub-entries, are database labels (see 6.2.1 for a more detailed discussion on labelling). It is decided during the analysis whether the label should also be offered in the dictionary entry.

Figure 57. DOAE: Recording basic information – Domain distribution of the lemma AUTHORITY.

doc.domain	Freq	Rel [%]
Law	2269	500.8
History	2210	390.2
Politics, Government & Int Relations	1341	313.0
Anthropology	1201	307.7
Theology and Religion	1254	246.8
Architecture	705	162.4
Education	901	162.0
Philosophy	926	158.2
Social Sciences	579	153.6
Art and Art History	482	122.8
Economics	413	85.4
Archaeology	285	81.4
Music	295	74.1
Business and Management	391	70.9
Geography	230	51.3
Finance	258	50.0
Linguistics	233	43.9
Psychology	136	35.4
Sports	53	11.9
Engineering	33	5.1
Computer Science	30	5.0
Veterinary Science	38	4.6
Medicine and Health Sciences	20	3.9
Biology	15	1.8
Chemistry	6	0.8
Mathematics	3	0.7
Biochemistry	3	0.3

6.3.1.7 Etymology

Etymology, i.e. information about the origin of the word and its historical development, is often provided in monolingual dictionaries for NSs, but almost never in dictionaries for foreign learners. But as Atkins and Rundell (2008) point out, some electronic dictionaries for learners (e.g. LDOCE) now include etymological information, but in a separate pop-up window, which needs to be opened by the user, rather than as integral part of the main entry screen.

Research indicates that students very rarely consult etymology (Tomaszczyk, 1979; Béjoint, 1981; Hartmann, 1999), but we need to bear in mind that some students (especially students of Linguistics and related subjects) may require etymological information at some point in the course of their study. Etymology is therefore recorded in the database, and made available to students (as part of the customisable options).

Etymology was recorded using the LED CD-ROM structure of date, language, and origin (see Figure 58). Etymological information was obtained by consulting CED CD-ROM, Online Etymology Dictionary, and LED CD-ROM.

Figure 58. DOAE database: Etymology information for *significant*.

Headword:	
HeadwordSign	significant
Etymology:	
Date	1500-1600
Language	Latin
Origin	%b%isignificare%i (to signify)%b

6.3.1.8 Headwords with variant form(s)

Recording basic information for headwords with variant form(s) requires a more complex procedure. Particular care is needed with word class information, frequency rank, frequency, and domain distribution. The procedure for other types of information, such as inflected forms, pronunciation and etymology, is more or less the same as with headwords with a single form.

The headword STATE-OF-THE-ART and its non-hyphenated variant STATE OF THE ART are used as an example. The pronunciation is the same for both forms, and there are no inflected forms. However, information about frequency rank and information about frequency could be based on the lemma list alone, as only STATE-OF-THE-ART featured in it. The number of STATE OF THE ART concordance lines needed to be added, and although that did not significantly impact the frequency per million words, it resulted in a considerable change in the frequency rank (Table 50).

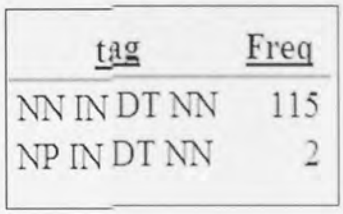
Table 50. DOAE: Recording basic information – Frequency information for STATE-OF-THE-ART and its variant STATE OF THE ART.

	state-of-the-art	*state of the art*	*state-of-the-art + state of the art*
raw frequency (per million words)	245 (2.6)	117	362 (3.8)
frequency rank	13478	-	10394

Identifying word classes using the method described in 6.3.1.1 was not problematic with STATE-OF-THE-ART as it was treated by Sketch Engine as a single word – all the occurrences are identified as adjectives. STATE OF THE ART, on the other hand, was treated as a sequence of four separate words (and can only be searched as a phrase), and the Node tags function

shows word class for each separate word (Figure 59). Thus, adjective, as the only automatically identified word class, was recorded in the database entry, and a note was left (for the analysis) that there may be other word classes.

Figure 59. DOAE: Recording basic information – Node tags display for STATE OF THE ART in Sketch Engine.



tag	Freq
NN IN DT NN	115
NP IN DT NN	2

Because it was not possible to do a concordance search for both variants at the same time, it was also not possible to obtain information about combined domain distribution. An examination of the domain distribution of each of the variants revealed that there was little difference between the distributions, so the domain distribution of STATE-OF-THE-ART (the more frequent form) was saved in the database, accompanied by a note saying that 117 concordance lines of STATE OF THE ART were not included in it, but showed very similar distribution across domains.

6.3.2 Meaning analysis

Meaning analysis is the single most important aspect of a lexicographer's work. During meaning analysis, the lexicographer describes the semantical, lexicogrammatical, and collocational characteristics of the headword. Meaning analysis has its own section (element Meaning.analysis) in the database entry, and information recorded in this section is used later for compiling the dictionary entry.

6.3.2.1 Meaning analysis with Word Sketch

The function of Sketch Engine that was at the core of the meaning analysis was Word Sketch. The word sketch provides lexicographers with a lexical profile of the word, which includes common grammatical relations and collocates. In this way, the lexicographers get a good insight into the word's typical patterns, and meanings, considering that meanings are associated with patterns (Hanks, 2008).

Word Sketch uses grammar codes for grammatical relations for the four major word classes (noun, verb, adjective, adverb), which are partly based on Frame Semantics (used by the FrameNet project[93]; see Atkins and Rundell, 2008). The lists of codes, which are a combination of the codes encountered during the analysis and the codes listed by Atkins and Rundell (2008), and the explanations of the codes, are provided in Table 119 (noun), Table 120 (verb), Table 121 (adjective), and Table 122 (adverb) in Appendix 8. The same grammar codes were used when saving information in the database.

The meaning analysis with Word Sketch consisted of three steps. Step 1 recorded the information provided by Word Sketch in the DOAE database. In Step 2, the recorded information was grouped by meanings. In Step 3, any as yet unidentified meanings were added by examining a random set of concordance lines.

The notion underlying this type of analysis is that of a pattern, and the close connection between patterns and meanings (Hanks, 2008). This is similar to CPA, used in compilation of PDEV (see 3.2.2.3); the important difference between the Hanks' approach and the approach used in this Model is that Hanks assigns a meaning to each individual pattern, whereas in this Model patterns were grouped if they had the same or similar meaning. This, as will be shown later, helped with creating and ordering senses in the dictionary entry.

Next, each step of the analysis with Word Sketch is presented in detail. The verb *attribute* (the word sketch of the verb *attribute* can be found in Figure 99 in Appendix 9) is used to exemplify the procedures. The verb *attribute* is frequent in CAJA with 6952 occurrences (frequency rank = 1408).

6.3.2.1.1 Step 1: Recording grammatical relations, collocates, and examples

Word Sketch identified several grammatical relations and collocates, however not all were relevant for analysis (see 6.3.2.3 for discussion on this). Thus, first, grammatical relations that would be recorded in the database needed to be selected. Different factors were taken into account – frequency of the relation, number of collocates, salience/frequency of the most salient/frequent collocates, etc. There was also a good deal of manual inspection of concordances involved, as I needed to determine whether the relation had been properly identified by Word Sketch. The results of such an analysis for the verb *attribute* are shown in Table 51.

[93] More information about the FrameNet project is available at http://framenet.icsi.berkeley.edu/.

Table 51. DOAE: Meaning analysis - The grammatical relations of *attribute* (verb), and their mode of inclusion in the database.

grammatical relation	% of total occurrences of *attribute*	most salient collocate	included in the analysis	comment
unary rels/passive	30	/	YES	
PP_*to*-i	52	Paisible	YES	
PP_PP_*between*-i	0.8	to	NO	included in PP_*to*-i relation
PP_PP_*in*-i	5.2	to	YES	
PP_PP_*of*-i	11.9	to	YES	
PP_PP_*by*-i	0.65	to	NO	included in PP_*to*-i relation
PP_PP_*because*-i	0.14	to	NO	included in PP_*to*-i relation
AVP_mod	11.5	mistakenly	YES	
NP_PP	17.2	to	YES	
PP_NP_Ving	1.2	to	NO	included in PP_*to*-i relation
NP	55.8	cm-1	YES	
PP_NP_Vinf_to	0.23	to	NO	included in PP_*to*-i relation
PP_PP_*during*-i	0.13	to	NO	included in PP_*to*-i relation
pro_object	1.2	it	YES	
PP_PP_to-i	0.95	in	NO	
PP_PP_*from*-i	0.45	to	NO	included in PP_*to*-i relation
PP_PP_*under*-i	0.07	to	NO	included in PP_*to*-i relation
PP_PP_*within*-i	0.1	to	NO	included in PP_*to*-i relation
PP_PP_*at*-i	0.33	to	NO	included in PP_*to*-i relation
PP_PP_*on*-i	0.43	to	NO	included in PP_*to*-i relation
pro_subject	3.1	he	YES	
subj_NP	20.5	divine	YES	
PP_PP_*as*-i	0.33	to	NO	included in PP_*to*-i relation
PP_PP_*with*-i	0.36	to	NO	included in PP_*to*-i relation
AVP	3	solely	YES	
PP_PP_*for*-i	0.22	to	NO	included in PP_*to*-i relation
PP_cl_wh	0.12	to	NO	included in PP_*to*-i relation
AJP	0.76	such	NO	
PP_*in*-i	0.62	part	NO	
PP_*on*-i	0.16	basis	NO	
and_or	0.43	denote	NO	

The more frequent relations were recorded in the database, whereas infrequent ones were excluded. Based on the sample entries built for this dictionary Model, 1% of total occurrences of the headword tended to be the cut-off point, although there were exceptions. Nonetheless, as demonstrated by the analysis of *attribute* in Table 51, many excluded relations

were still included in the analysis (and possibly in the database) as they were part of more frequent relations.

The next step in the analysis was to record in the database the selected grammatical relations, salient collocates of those relations, and examples of those collocates. Concordance lines of each of the salient collocates were examined first to exclude any incorrectly identified collocates (see 6.3.2.3.2.3), or to obtain additional information about the collocate. Examples of additional information (Table 52) were collocates that were specific to a particular domain or domain category, frequent syntactic patterns of the collocate, and preference for a particular word-form.

Table 52. DOAE: Meaning analysis - Detailed analysis of the collocates in the grammatical relation 'PP_to-i' (the verb *attribute*).

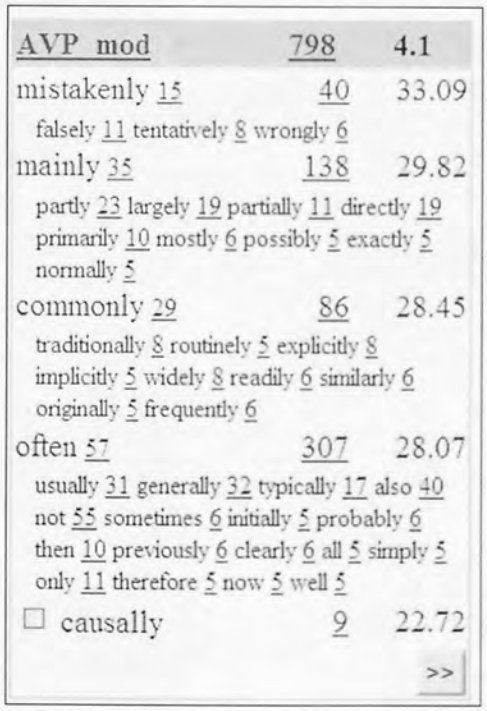
collocate	comment (mainly in passive unless noted)
Paisible	all occurrences from one Arts and Art History text
Galuppi	all occurrences from one Music text
fact	very frequently found in phrase *be + attributed to the fact (that)*
Kermit	all occurrences from one Philosophy text (all in active voice)
difference	plural form (*differences*) much more frequent
presence	mainly found in Science texts (frequently in *be attributed to the presence of*)
lack	frequently found in *attributed to the/a lack of*
Dryas	all occurrences from one Geography text
Michelangelo	all occurrences from one History text
erectus	all occurrences from one Archaeology text
physician	all occurrences from one Social Sciences text
formation	all occurrences from Sciences texts
factor	mainly found in plural form (*factors*)
variation	mainly found in singular form (*variation*)
messenger	all occurrences from one Politics text
increase	predominantly found in singular form
vibration	all but one concordance from Chemistry texts
effect	more common in Sciences than in Arts and Humanities
Kant	all but one concordance from Art and Philosophy text
Moore	all occurrences from one Philosophy text
failure	found only in singular
reduction	found only in singular, all but one concordance from Sciences texts
loss	found mainly in Sciences texts
cause	
change	many examples from Sciences (especially Chemistry and Biochemistry)

The TickBox Lexicography function was then used to select collocates, select examples, export them in XML format, and then import the XML file into the DOAE database. The entire process is illustrated in Figure 100 in Appendix 9. Several examples of each collocate were recorded, especially if the collocate was found in different forms, such as singular or plural, or in certain syntactic patterns, such as with the verb in the passive.

It was useful to group collocates of a particular relation whenever possible. The clustering function in Word Sketch was sometimes useful for this purpose. Lexicographers can also specify sub-group categories, for example the adverbs in the grammatical relation 'ADV_mod' of the verb *attribute* were divided into three groups (Figure 60):

1. Adverbs indicating frequency (e.g. *often, commonly, usually, generally*);
2. Adverbs indicating degree (e.g. *mainly, partly, largely*); and
3. Adverbs indicating negation (e.g. *mistakenly, falsely, wrongly*). These adverbs are often used when someone attributes something to somebody.

Figure 60. DOAE: Meaning analysis - Adverbial modifiers of the verb *attribute* (Word Sketch, clustered view, sorted by salience).



Collocates can be also grouped according to a common superordinate. Table 53 below shows an example of semantic grouping of collocates in one of the relations of the verb *attribute*.

Table 53. DOAE: Meaning analysis - Semantic grouping of collocates in grammatical relation 'PP_*to*-i' of *attribute* (verb).

person	activity	thing
Paisible	*fact*	*erectus*
Galuppi	*change*	*messenger*
Kermit	*increase*	
Dryas	*difference*	
Michelangelo	*lack*	
physician	*loss*	
Kant	*reduction*	
Moore	*failure*	
	vibration	
	effect	
	variation	
	formation	
	presence	
	cause	
	factor	

To distinguish between the two types of collocate groups, two different DTD elements have been created: 'Collo.type' is used when collocates were grouped by word class typology, and 'Collo.semantic' when collocates were grouped by semantic properties.

It was useful to use the same names for groups of collocates in different relations that belong to the same part of the syntactic pattern. It was thus better to group collocates after at least rough patterns were identified, which normally occured once all grammatical relations and collocates had been examined, and examples recorded in the database. For example, the examination of concordances of the salient collocates of the four most frequent grammatical relations of *attribute* ('NP', 'PP_*to*-i', 'NP_PP', and 'subj_NP'), pointed to three syntactic patterns (one pattern had two variants):

1. [Object 1] **be attributed to** [Object 2] – *passive voice*
   [Subject 1] **attribute** [Object 1] **to** [Object 2] – *active voice* (pattern variant)
2. [Subject 1] **attribute** [Object 1] [[no PP-*to*]] – *active voice*
3. [Object 2] **to which/whom** [Object 1] **be attributed** – *passive voice*

Subject 1 consisted of some of the collocates that occured in the grammatical relation 'subj_NP'. Object 1 were collocates in relation 'NP', and remaining collocates in relation 'subj_NP'. Object 2 were collocates in relation 'PP_*to*-i'. The grammatical relation 'NP_PP' contained collocates which were prepositions (mainly *to*) linking Object 1 and Object 2 in pattern 1 (the variant in the passive voice). Based on these syntactic patterns, I could start grouping collocates that were found in the four grammatical relations:

a) Subject 1 was normally a person (e.g. *author, consumer, scholar*).

b) Object 1 was a noun or a noun phrase denoting: finding/result (*difference(s), effect(s), increase, result(s), finding(s), outcome(s), decline*), level of success (*failure, success),* characteristic (*property(ies), value(s), meaning, importance*), authorship (*saying*), and blame/responsibility (*blame, responsibility*).

c) Object 2 (the noun or noun phrase that was part of the *to*-prepositional phrase) was normally a thing, a person, or an institution. The thing could be a finding, result or a fact (*fact, difference(s), effect(s), factor(s), lack, increase, change(s)*), a property (*characteristic(s), property(ies)*), or a physical object (*object*).

d) Both Object 1 and Object 2 were often found in the form of a (complex) noun phrase. The noun phrase was frequently found in the following two patterns: '(*THE*) + sth + *OF* + sth' (e.g. *the failure of development*), or '(*the*) + sth + *in* + sth' (e.g. *the increase in sleep disturbances*).

It was also important to record any syntactic patterns and their collocates that were not clearly pointed out by grammatical relations. One such a pattern, which is pervasive throughout the word sketch of the verb *attribute*, was the combination of *attribute* with a *to*-prepositional phrase. Especially frequent was the syntactic pattern in which *attribute* was followed by *to,* indicated not only by a very frequent relation 'PP_*to*-i', but also by the fact that *to* was the most salient and most frequent collocate in several other common grammatical relations, for example 'NP_PP', 'PP_PP_*between*-i', 'PP_PP_*in*-i', 'PP_PP_*of*-i', and 'PP_NP_Ving'.

As *to* was found in many patterns which were often distinguished only by the number of words between *attribute* and *to*, manual searches were conducted to examine the real extent of the co-occurrence of *attribute* and *to*. An examination of concordances revealed that *to* was found up to 10 words or more to the right of *attribute* (Table 54) because the thing being attributed was often a long complex noun phrase.

Table 54. DOAE: Meaning analysis - Different manual searches of co-occurrence of *attribute* and *to*.

search	percentage of all concordance lines of *attribute*
attribute + *to* (adjacent)	61.5 %
attribute + *to* (*to* 1-5 words to the right)	82 %
attribute + *to* (*to* 1-10 words to the right)	89 %

It was important to record any frequently occurring phrases, and their examples, that might need to be highlighted in the dictionary entry. At *attribute*, one such frequently occurring phrase was: *be attributed to the fact that* (100 hits). The verb *be* (found in various word-forms) was the only flexible part of the phrase. A more detailed discussion of the identification of multi-word items can be found in 6.3.2.2.

Before moving to Step 2, it was also important to consult any notes about hyphenated variants of the collocates identified during the creation of the headword list (see 5.1.3). Frequent variants needed to be mentioned under the relevant collocate, and an example provided.

6.3.2.1.2 Step 2: Identifying syntactic patterns, pattern elements, and meanings

In Step 2, the grammatical relations, collocates, and any syntactic patterns identified in Step 1 were grouped according to meanings, called 'meaning patterns' in the database. This was achieved by examining which collocates, or groups of collocates, appeared together in patterns.

The point of departure of Step 2 of the meaning analysis for the verb *attribute* was the three syntactic patterns identified in Step 1. The common elements in the patterns were Subject 1, Object 1 and Object 2, and collocates found in these respective elements of the patterns had already been semantically grouped within their grammatical relations. In Step 2, each pattern was examined separately, and divided into further patterns according to the collocates or collocate groups that represented each part of the pattern. Pattern elements were given new, more specific names which represented a superordinate of the collocate groups found in them[94]. For example, the pattern '[Object 1] **be attributed to** [Object 2]' was divided into the following three patterns:

A. [Event 1] **be attributed to** [Event 2]
B. [Abstract Entity] **be attributed to** [Physical Object | Human | Institution]
C. [Artifact | Event] **be attributed to** [Human | Deity | Institution]

The main difference between syntactic pattern A, and syntactic patterns B and C, lay in Object 2, and the distinction between syntactic pattern B and syntactic pattern C was made mainly by collocate groups that function as Object 1.

The same procedure was repeated with the other syntactic patterns identified in Step 1 (page 216). The variant of the first pattern ([Subject 1] **attribute** [Object 1] **to** [Object 2]) could be divided into three new patterns, while the second pattern and the third pattern could not be

[94] The Brandeis Semantic Ontology (see 3.2.2.3.1) is consulted for this purpose.

divided further. Nonetheless, the parts of those two patterns were given more specific superordinate descriptions.

One of the things noticed while assembling these new sub-patterns was that each pattern was associated with a specific meaning. At this point, three different meanings of *attribute* were observed (Table 55). The new syntactic patterns were therefore grouped according to these meanings, forming meaning patterns which were recorded in the database. Each meaning pattern contained syntactic patterns associated with it (called 'pattern definitions' in the database), a list of most salient collocates of each element of the syntactic pattern (these elements were called 'meaning pattern elements' in the database), an explanation of the meaning, any notes (on connotation, collocates, etc.), and examples:

Table 55. DOAE: Meaning analysis - Initial three Meaning patterns of *attribute* (verb).

MEANING PATTERN 1	
Pattern definitions	[Event 1] **be attributed to** [Event 2] [Human] **attribute** [Event 1] **to** [Event 2] [Event 2] **to which/whom** [Event 1] **be attributed** **be attributed to the fact that**
Pattern elements	Human: *scholar(s), author(s), consumer(s), researcher(s), he, we, they* Event 1: *difference(s), effect(s), increase, result(s), finding(s), outcome(s), decline, failure, success* Event 2: *fact, difference(s), effect(s), factor(s), lack, increase, change, property(ies)*
Meaning	If you attribute something, such as a finding, to something, that thing is perceived to be the cause of the finding.
Notes	Event 1 and Event 2 are often found in the form of a long noun phrase, or even a clause. Collocates listed are normally head nouns of the noun phrase. This pattern often refers to a finding/result that does not support the argument of the text, or a finding/result that was not expected.
Examples	Their review reports results from studies conducted in 29 out of 61 provinces in Vietnam and the authors **attribute** the differences between regions and provinces to differences in climatic and environmental factors. This lack of transfer from gain in muscle force and power into enhanced functional performance was **attributed** to the fact that a successful soccer kick depends on a precisely coordinated action of the leg muscles rather than on isolated muscle strength (Aagaard et al., 1993).

MEANING PATTERN 2	
Pattern definitions	[Human] **attribute** [Abstract Entity] **to** [Physical Object | Human | Institution] [Abstract Entity] **be attributed to** [Physical Object | Human | Institution] [Physical object] **to which/whom** [Abstract Entity] **be attributed**
Pattern elements	Human: *we, Foucault* Abstract Entity: *property(ies), value(s), meaning, importance* Human | Institution: *men, women, corporations, sisters* Physical Object: *object(s)*
Meaning	If you attribute a property, such as importance, to something or someone, you believe that it has that property.
Notes	Types Human | Institution are found in pattern with collocates such as *characteristic* and *quality* (as Abstract Entity), but not collocates such as *importance* or *value*.
Examples	None of the students had **attributed** much importance in the questionnaire to the use of strategies, effective or otherwise. Previous research particularly emphasizes that reading comprehension is an active and interactive process between the readers and the written materials in which readers **attribute** meanings to the text through their background knowledge (Goodman, 1995; Lewis & Wray, 2000; Rasinski & Padak, 2004; Zwiers, 2004).

MEANING PATTERN 3	
Pattern definitions	[Human 1 | Entity] **attribute** [Artifact | Event] **to** [Human 2 | Deity | Institution] [Artifact | Event] **be attributed to** [Human | Deity | Institution] [Human] **attribute** [Event] [[no PP-*to*]] [Human | Deity | Institution] **to which/whom** [Artifact | Event] **be attributed**
Pattern elements	Human 1 | Entity (none is prevalent): *she, we, scholars, study* Artifact | Event: *saying, statement, argument, painting, text, blame, responsibility* Human 2 | Deity: *Michelangelo, Kant, actor(s), supernatural beings, state*
Meaning	If you attribute something, such as a painting or a statement, to someone, you believe they are the author of it. If you attribute blame or responsibility (for something) to someone, you think they are to blame for that thing happening.
Examples	The paper distinguishes between two different senses of 'genius' found in Kant's Critique of Judgement, and criticizes an argument commonly **attributed** to Kant. Chavez's supporters and opponents have both **attributed** to him considerable responsibility for the resurgence of Latin America's left -- most recently with the election of Evo Morales in Bolivia.

Frequent phrases such as *be attributed to the fact that,* were recorded under the relevant meaning pattern as meaning pattern definitions.

The explanation of the meaning pattern could be a draft definition, a synonym, or a list of synonyms. But in any case, writing an explanation of the meaning pattern was not given priority over identifying the meaning pattern definitions, and meaning pattern elements.

After all the information had been recorded in the database entry, the meaning patterns needed to be re-examined for any further distinctions. In this case, examples of Meaning pattern 3 pointed to two separate meanings; one with no connotations ('to attribute authorship') and one with negative connotations ('to attribute blame/responsibility'). Furthermore, one of the pattern definitions, namely '[Human] **attribute** [Event] [[no PP-**to**]]', and certain meaning pattern elements (Institution and Entity) were specific to the meaning with negative connotations. The negative pattern nearly always contained *blame* or *responsibility* as Object 1 (the thing being attributed). Meaning pattern 3 was therefore split in two separate meaning patterns (Table 56).

Table 56. DOAE: Meaning analysis - Two meaning patterns, formed out of original Meaning pattern 3 of *attribute* (verb).

MEANING PATTERN 3	
Pattern definitions	[Human 1] **attribute** [Artifact | Event] **to** [Human 2 | Deity] [Artifact | Event] **be attributed to** [Human | Deity] [Human | Deity] **to which/whom** [Artifact | Event] **be attributed**
Pattern elements	Human 1 (rare and none is prevalent): *she, we, scholars* Artifact | Event: *saying, statement, argument, painting, text* Human (2) | Deity: *Michelangelo, Kant, supernatural beings*
Meaning	If you attribute something, such as a painting or statement, to someone, you believe they are the author of it.
Example	The paper distinguishes between two different senses of 'genius' found in Kant's Critique of Judgement, and criticizes an argument commonly **attributed** to Kant.

MEANING PATTERN 4	
Pattern definitions	[Human 1 | Entity] **attribute** [*responsibility* | *blame*] **to** [Human 2 | Institution] [*responsibility* | *blame*] **be attributed to** [Human | Institution] [Human] **attribute** [*blame* | *responsibility*] [[no PP-*to*]]
Pattern elements	Human 1 (rare and none is prevalent): *we, researchers, authors, study* Human 2 | Institution: *group(s), state, actor(s)*
Meaning	If you attribute blame or responsibility (for something) to someone, you think they are to blame for that thing happening.
Example	Chavez's supporters and opponents have both **attributed** to him considerable responsibility for the resurgence of Latin America's left -- most recently with the election of Evo Morales in Bolivia.

Step 2 was concluded by considering any changes to the order of meaning patterns. Frequency information for the meaning pattern definitions, grammatical relations, and collocates needed to be consulted. It has been observed that meaning patterns with a greater number of meaning pattern definitions tend to be more frequent.

A screenshot of the partial database entry for *attribute* containing meaning pattern information is offered in Figure 61.

Figure 61. DOAE database: Meaning pattern 1 of the verb *attribute* (without the examples).

HeadwordSign	attribute
Meaning.pattern:	1
Meaning	If you attribute something, such as a finding, to something, that thing is perceived to be th
Notes	Event 1 and Event 2 are often found in the form of a long noun phrase, or even a clause.
Domain.label.L1	.
References:	
Meaningpattern.definition:	
Pattern.definitio	[Human] attribute [Event 1] to [Event 2]
Domain.label.L1	.
Meaningpattern.definition:	
Pattern.definitio	[Event 1] be attributed to [Event 2]
Domain.label.L1	.
Meaningpattern.definition:	
Pattern.definitio	[Event 2] to which/whom [Event 1] be attributed
Domain.label.L1	.
Meaningpattern.definition:	
Pattern.definitio	be attributed to the fact that
Domain.label.L1	.
Meaning.patternelements:	
Pattern.element	[Human]: scholar(s), author(s), consumer(s), researcher(s), he, we, they
Meaning.patternelements:	
Pattern.element	[Event 1]: difference(s), effect(s), increase, result(s), finding(s), outcome(s), decline, failu
Meaning.patternelements:	
Pattern.element	[Event 2]: fact, difference(s), effect(s), factor(s), lack, increase, change, property, propert

6.3.2.1.3 Step 3: Adding any missed meanings, and significant collocates

Because not all occurrences of all the collocates in the word sketch could be analysed for practical reasons, there was a possibility that some meanings of the word had been missed. The meaning analysis thus needed to be concluded with an examination of concordance lines. Since many headwords were very frequent, not all the occurrences could be inspected. The searches for additional meanings looked at 250 randomly selected concordance lines. At some headwords, however, the number of concordance lines from which a random selection was made was (significantly) reduced by filtering out known collocates and/or syntactic patterns. For example, before the random selection of concordance lines of the verb *attribute* was created, the concordance lines with *blame* and *responsibility* occurring in the span +3-3 of *attribute* were excluded.

The second part of Step 3 was to use the Collocation function on all the occurrences of the headword, and record the most significant collocates (considering both T-score and MI^3 values). The default span in Sketch Engine is +5-5, however a larger span such as +7-7 was more suitable for some headwords, especially verbs that were surrounded by complex noun phrases. In addition, a larger span was often needed to compensate for the fact that punctuation is tokenised in the corpus. Collocates were recorded under the 'Collocates_significant' element in the Meaning analysis part of the database entry.

Recording the most significant collocates provided a reference that could be consulted to ensure that the most relevant collocational information had been included in the entry. Also, the list of collocates helped to identify any collocates that might not be noticed in the word sketch because they were dispersed across many different grammatical relations. For example, in the case of the verb *attribute,* the collocate *can* was among the top 5 significant collocates (excluding punctuation)[95], but was not among salient collocates in any of the frequent grammatical relations identified by Word Sketch.

6.3.2.1.4 An alternative approach to meaning analysis

Identifying syntactic patterns and related meaning patterns is usually easy with verbs, but is often problematic for other word classes. For certain headwords, an alternative approach to meaning analysis was used, which started by identifying meanings in a random selection of concordance lines, and then recorded grammatical relations and collocates under each meaning. The final step, recording the most significant collocates, remained the same. This alternative approach therefore used all the steps described in the previous sections, but in a different order.

One of the headwords for which this approach was used was the noun *authority.* First, 300 random concordance lines were examined and 7 different meanings were initially identified (Table 123 in Appendix 9). Examples of each meaning along with any notes (e.g. about *AUTHORITY* found mainly in singular or plural) were also recorded. Then, the word sketch of *AUTHORITY* was created. The most salient collocates of the more frequent grammatical relations were recorded; analyses of the two most salient grammatical relations, 'AJ_premod' (46% of all occurrences of *authority*) and 'object_of' (23%), are provided in Table 124 and Table 125 respectively.

[95] In fact, Word Sketch misses several modal verbs, with *cannot, may* and *could* being among the top 20 significant collocates (according to MI^3).

The majority of collocates were found to be associated with one or two meanings of *authority*. For many collocates associated with two meanings, the number (i.e. singular or plural) of *authority* can be used to disambiguate the meaning. Collocate type groupings were possible at only a few grammatical relations, 'object_of' being one of them.

The final distribution of grammatical relations and syntactic patterns by meanings (see Table 126, Table 127, Table 128, and Table 129[96] in Appendix 9) pointed to the following:

- Meaning 1 and Meaning 3 are the most frequent meanings of *authority*. Most salient (and frequent) collocates of most salient grammatical relations of *authority* normally co-occur with *authority* in one of these two meanings.
- The noun *authority* in Meaning 1 is found only in singular.
- The noun *authority* in Meaning 3 is found in both singular and plural, but plural use is more frequent.
- There are only few salient collocates found with Meaning 2 or Meaning 4 of *authority*. *authority* is found in both singular and plural in these two meanings.
- There were no salient collocates found that co-occur with Meaning 6 or Meaning 7 of *authority,* which suggests that those meanings are infrequent.
- Meaning 5 (another infrequent meaning) seems to be an extension of Meaning 3, the only difference seemingly being that Meaning 5 contains occurrences of *authority* as a proper noun[97].

The final step involved saving the list of most significant collocates. The collocate *local* (found in 1001 concordance lines) is particularly important as it is by far the most salient, frequent, and statistically significant collocate (T-score = 31.939)[98] of all collocates in grammatical relations identified by Word Sketch.

This alternative approach to meaning analysis proved more appropriate for nouns, adjectives, and adverbs. The approach is, however, accompanied by a potential pitfall; the lexicographers can be influenced by the initially identified meanings, and may try to include each collocate in one of the existing meanings rather than create a new meaning. This caveat should thus be included in the Style Guide.

---

[96] Tables for meanings 5, 6, and 7 are not provided as no collocates were found that co-occur with *authority* in these meanings.

[97] Based on this observation, the example containing the compound *police authority* was moved from Meaning 5 to Meaning 3.

[98] *Local* is the 8th on the list of collocates ordered by T-score (span -5+5), preceded only by *the, of, to, and, in, a* and *that.*

6.3.2.1.5 Grammatical relations in Word Sketch

While only more salient relations were recorded in the database entry, in fact the number of different grammatical relations offered by Word Sketch could be quite high at certain entries. It was noticed during the analysis of sample entries that some grammatical relations featured more frequently than others. The main factor affecting in the role of grammatical relations was word class.

Table 57 below shows the grammatical relations that were salient in at least two different sample entries for each of the four main word classes. There are significant differences between the grammatical relations of each word class. Lexicographers can benefit from these lists as they know which grammatical relations they should expect to encounter. Similarly, the absence of any of the relations may signal an important characteristic of the entry word.

Certain grammatical relations are of particular importance to individual word classes. 'NP' and 'subj_NP' are important to identify the subject and object of the verb respectively. 'AJ_premod', 'N_mod', and 'premod_N' are relations which are important for nouns when trying to identify compounds (see also 6.3.2.2.1). The relation 'premod_N' has a similar compound-identifying function for adjectives, but it can also provide additional grammatical information; for example, the percentage of an adjective's occurrences in this relation can indicate whether it is (predominantly) attributive.

'Unary rels' is an important type of relation for all word classes as it lists any salient relations specific to the word. For example, for the verb *argue,* the unary relation 'that_0' represented 65% of the occurrences of *argue.* Such information is vital for lexicographers, who need to ensure that the patterns like *argue that* are given a prominent role in the dictionary entry.

Although the relation 'and_or' was frequent only with adjectives, it did feature in other word classes. The value of this relation lies mainly in the fact that it often provides synonyms and/or antonyms of the word, which can then be used for writing quick explanations of meaning patterns and for compiling the synonym part of the dictionary entry (6.3.3.8).

It is worth noting that because of the limited number of sample entries compiled, certain grammatical relations were identified as specific to individual headwords, whereas they may be shared by several headwords, or a word class. The majority of these relations were prepositional phrases – for example, the relation 'PP_*to*-i' represents 52% of the occurrences of the verb *attribute*, 'PP_*for*-i' represents 15% of the occurrences of the noun *potential*, and 'PP_*from*-i' represents 14% of the occurrences of the verb *obtain.*

Table 57. DOAE sample entries: Frequent salient grammatical relations (by word class).

VERBS		
grammatical relation		**sample entries (% of all occurrences)**
unary rels	that_0	argue (65%)
	passive	obtain (50%), attribute (30%), justify (28%)
	it	argue (4%)
NP		obtain (74%), justify (62%), feature (61%), attribute (55.8%)
subj_NP		feature (50%), attribute (20.5%), justify (18%)
NP_PP		feature (20%), justify (18%), attribute (17.2%), obtain (14%)
AVP_mod		justify (15%), feature (13%), attribute (11.5%), obtain (5%)
AVP		feature (8%), argue (4.2%), attribute (3%), justify (2.8%), obtain (2.6%)
PP_Ving		obtain (6%), justify (4.4%)

NOUNS		
grammatical relation		**sample entries (% of all occurrences)**
AJ_premod		feature (61%), method (50%), authority (49%), assortment (45%), potential (40%), attribute (40%), ANOVA (30%), RNA (20%), CEO (11%), fact (9%)
V_subj		ANOVA (27%), authority (20%), CEO (17%), feature (12.7%), potential (10%)
object_of		potential (50%), method (40%), feature (37%), attribute (28%), authority (24%), assortment (23%), ANOVA (20%), fact (17%), CEO (12%)
N_mod		method (36%), potential (30%), assortment (14%), attribute (14%), authority (14%), feature (10%)
PP_of-i		feature (28%), assortment (25%), potential (15%), attribute (12%), method (7%), authority (7%)
premod_N		RNA (59%), CEO (40%), attribute (14%), feature (10%), authority (5%)
PP_obj_of-i		authority (17%), attribute (12%), feature (10%), potential (10%), method (9%), fact (5%)
unary rels	that_0	fact (50%)
	Vinf-to	potential (13%)

ADJECTIVES	
grammatical relation	**sample entries (% of all occurrences)**
premod_N	potential (95%), various (92%), subsequent (92%), state-of-the-art (80%), significant (70%)
and_or	potential (21%), various (20%), significant (15%)
AVP_premod	significant (23%)
comp_V	significant (20%)

ADVERBS	
grammatical relation	**sample entries (% of all occurrences)**
premod_VP	subsequently (62%), thus (26%), therefore (23%)
VP	subsequently (28%), therefore (7%), thus (5%)
unary rels (CL)	subsequently (20%)
AVP	therefore (5%), thus (5%)
AJP	therefore (5%), thus (3.3%)

6.3.2.1.6 Domain labelling during meaning analysis

Domain labelling during the meaning analysis is discussed separately as it played an important role in the meaning analysis. Domain labelling is based entirely on corpus data, and impacts on the use of domain labels in the dictionary entry. It is much easier to assign domain labels to dictionary entries and senses if there is information available on any domain specificity of parts of the meaning analysis.

Domain labels are recorded for collocates and collocate types, syntactic patterns, grammatical relations, meaning pattern definitions, and meaning patterns. Initially, domain labels for collocates, and possibly grammatical relations, are recorded on the basis of corpus information[99]. This helps distinguish between general collocates (no label), and domain-specific collocates (see Figure 62 below).

A domain label is assigned to collocates even if just a majority (rather than all) of the concordance lines fall under a certain label. For example, L2 label Business Sciences was assigned to the collocate *customer* of the adjective *potential* because 67 out of 86 concordance lines came from Business Sciences domains (Table 58 below).

Other types of information, such as collocate types, represent broader groupings of lower level information, and the same principle is used when assigning domain labels to them. So, a collocate pattern or grammatical relation is assigned a label if most of its collocates are limited to a specific domain or domain category. A similar procedure is used for meaning pattern definitions, and meaning patterns. For example, most grammatical relations of Meaning pattern 3 of the verb *attribute* (Figure 62) had collocates from 'Arts and Humanities' domains, especially from level 2 categories 'Arts' and 'Humanities'. Although collocates in 'AVP_mod' and 'pro_subject' relations did not have any domain labels, all examples came from 'Arts' and 'Humanities'. The analyses of random concordance lines in other domains confirmed this: the analysis of random 500 concordance lines from Sciences domains (250 Applied Sciences, 250 Life Sciences) identified only one example of this Meaning pattern, and the analysis of random 250 concordance lines from Business Sciences and Social Sciences identified only 2% of concordance lines in this Meaning pattern. Therefore, level 2 domain labels 'Arts' and 'Humanities' were assigned to the Meaning pattern of these grammatical relations.

[99] Domain labelling of collocates and grammatical relations could be (semi-)automatised, but that was not available in this research.

Figure 62. DOAE database: Domain labels assigned to collocates in grammatical relations of Meaning pattern 3 of *attribute* (verb).

- gramrel: grname=AVP_mod
  - Collo.type: collo.typename=modifying adverb (frequency)
    - collocation: collo=often
    - collocation: collo=usually
  - Collo.type: collo.typename=modifying adverb (degree)
- gramrel: grname=subj_NP
  - Collo.type: collo.typename=object 1 (the thing being attributed): authorship
    - collocation: collo=saying,Domain.label.L2=Humanities
    - collocation: collo=statement,Domain.label.L1=Arts and Humanities
    - collocation: collo=argument,Domain.label.L1=Arts and Humanities
- gramrel: grname=PP_to-i
  - collocation: collo=Paisible,Domain.label.L3=Art and Art History
  - collocation: collo=Galuppi,Domain.label.L3=Music
  - collocation: collo=Dryas,Domain.label.L3=Geography
  - collocation: collo=Michelangelo,Domain.label.L3=History
  - collocation: collo=erectus,Domain.label.L3=Archaeology
  - collocation: collo=physician,Domain.label.L3=Social Sciences
  - collocation: collo=messenger,Domain.label.L3=Politics, Government & Int I
  - collocation: collo=Kant,Domain.label.L1=Arts and Humanities
  - collocation: collo=Moore,Domain.label.L3=Philosophy
- gramrel: grname=subj_NP
  - collocation: collo=scholar,Domain.label.L2=Humanities
- gramrel: grname=pro_subject
  - collocation: collo=she

Table 58. DOAE: Meaning analysis (adjective *potential*) - Domain distribution of concordance lines of collocate *customer* in the grammatical relation 'premod_N'.

Subcorpus	Frequency
Business and Management	35
Economics	16
Finance	13
Linguistics	4
Law	3
History	3
Education	3
Architecture	3
Computer Science	2
Theology and Religion	1
Politics, Government & Int Relations	1
Geography	1
Biochemistry	1

All three levels of domain labels presented in 6.2.1 are used because broader domain categories, while not displayed in the dictionary entry, provide the basis for customisable functions of the dictionary, such as sense ordering.

6.3.2.2 Using Word Sketch to identify multi-word candidate headwords

Word Sketch is also very helpful in identifying multi-word items such as compounds, phrasal verbs (a type of verb compound), and phrases and idioms. Identifying compounds and phrasal verbs at this stage is particularly important as these two groups of multi-word items, while assigned headword status in this dictionary Model, do not feature on the initial headword list, so they need to be added to it. Identifying phrases and idioms, on the other hand, is more important for the entry being compiled, as that information is vital for compiling the dictionary entry.

6.3.2.2.1 Identifying compounds

Of the grammatical relations displayed by word sketches of sample noun entries, three were particularly important for identifying potential compounds: 'premod_N', 'N_mod', and 'AJ_premod'. These three relations look at the words immediately before or after the headword, focussing on noun-noun and adjective-noun combinations. Other relations also needed to be examined, but were not as nearly as compound-productive.

These three relations should be searched for compounds during the meaning analysis, (i.e. when recording these three relations, the headword's collocates, and the examples in the database) because otherwise, less frequent compounds may be missed because the collocates that were not salient enough in the grammatical relation have not been recorded in the database.

The main problem of identifying compounds, however, is not in frequency but in meaning. Atkins and Rundell (2008) divide compounds into three types (figurative, semi-figurative, and functional), according to their level of idiomacity, and point out that figurative and semi-figurative compounds are not difficult to identify, whereas functional compounds are often overlooked by lexicographers, or better said, mistaken for non-idiomatic compounds. Non-idiomatic compounds, if frequent enough, are also likely to feature in the entry, so it is important to ensure they are given enough prominence (see 6.3.3.3 and 6.3.3.5).

The level of idiomacity of a compound was a problem when deciding on candidate compounds in the sample entries for DOAE. However, by far the most problematic candidate

compounds proved to be those that exhibited various levels of technical meaning. These compounds were often found in only one or two related domains, or even in only one or two texts. The status of these compounds was difficult to determine without the necessary technical knowledge. One solution is to record all such compound candidates and let domain experts decide whether the candidates are legitimate compounds.

But for this dictionary Model, domain experts were not available, so it was decided instead to consult resources such as CAJA itself, large NS dictionaries (CED CD-ROM, NODE CD-ROM, MWCD CD-ROM, and Dictionary.com), and Wikipedia. All three resources might contain definitions/explanations of the compound. Another reason for consulting the dictionaries was to check whether the compound candidate had an entry status in them. Table 130 in Appendix 9 shows an example of such an analysis for collocates in the grammatical relations 'AJ_premod' and 'N_mod' in the word sketch of the noun *potential.* The findings of the analysis (definition(s), examples, and notes about the status of the compound were recorded in the database (Figure 63).

Figure 63. DOAE database: Information for collocate *action* in the grammatical relation 'N_mod' of *potential* (noun).

collocation:	
colo	action
Notes	Found in various domains (mainly Sciences), but mainly in Medicine, Biology, Biochemistry, Computer Science, Sports, and Veterinary Science.
Notes	Definition from NODE: "(Physiology) the change in electrical potential associated with the passage of an impulse along the membrane of a muscle cell or nerve cell."
Notes	Definition from CED: "a localized change in electrical potential, from –70 mV to +30 mV, that occurs across a nerve fibre during transmission of a nerve impulse"
Notes	CANDIDATE FOR ENTRY DUE TO FREQUENCY + FOUND IN ALL FOUR CONSULTED DICTIONARIES
Example: 1	
[illegible]	Biochemistry
Example	The voltage-gated para-sodium channel, the primary target of pyrethroids, is a large glycoprotein and the major engine of nerve action potentials.
[illegible]	
Source	Biochem_38_2006_sonodaetal
Example: 2	
[illegible]	Mathematics
Example	Applying a sufficiently strong stimulus to a cardiac cell leads to a prolonged elevation of transmembrane voltage v known as an action potential.
[illegible]	
Source	Math_48_2006_cain+schaeffer

It is essential that the headword list (including the headword variants) is consulted when identifying compounds to see whether a candidate compound is a) already in the database as a headword, or its variant, or b) a (not yet identified) variant form of database headword, in which case a note for the editorial team is inserted. Compounds that do not belong to group a or b, and were identified as candidates for the headword (such as *action potential* above) were recorded under Entry.Candidates.

6.3.2.2.2 Identifying phrasal verbs

Phrasal verbs, consisting of a verb and one more particles, are the most common verb compounds (Atkins & Rundell, 2008). Word sketches of verbs are the best place to look for phrasal verbs, as it may be easier to identify phrasal verbs there than in word sketches of particles. Many grammatical relations in the word sketch of a verb can help identify phrasal verbs, but the ones containing particles, prepositional phrases, or adverbs are especially relevant.

The verb *take* was very frequent (ranked among top 200 lemmas in CAJA) and known for being part of numerous phrasal verbs, which made it a good candidate to exemplify the process of identifying phrasal verbs with Word Sketch (the word sketch for *take* is provided on the CD-ROM, Appendix 13). 17 candidate phrasal verbs were identified by examining not only the grammatical relations but also concordance lines of relations and collocates of *take*. The candidates were found in 26 out of the 73 grammatical relations in the word sketch of the verb *take* (Table 59). Some phrasal verbs were discovered only after the settings in Word Sketch were changed. The default settings (minimum frequency of items = 7, maximum number of items in a relation = 25), revealed only 18 candidate phrasal verbs. After changing the settings to minimum frequency of items = 3, maximum number of items in a relation = 50, this increased to 26 candidates.

The grammatical relations in which candidate phrasal verbs were found were of two types: relations whose collocates were the particles, and relations that themselves constituted phrasal verbs. An example of the former type of relation for *take* is 'part_intrans', and of the latter type is 'PP_*over*-i' (Figure 64).

Figure 64. DOAE: Meaning analysis - Two types of relations with candidate phrasal verbs of *take*.

part_intrans	1659	13.0
over	364	73.95
up	854	71.36
off	158	53.06
away	52	42.36
out	173	39.21
on	18	24.63
down	24	24.32
along	7	17.3
		>>

PP_over-i	465	6.3
responsibility	12	18.01
leadership	8	15.89
management	12	15.24
role	16	14.65
function	19	14.58
business	8	13.54
company	6	9.91
position	7	9.21
part	7	7.91
firm	6	7.84
year	6	7.4
		>>

Table 59. DOAE: Meaning analysis - Candidate phrasal verbs of the verb *take*, ordered by the number of grammatical relations in which they are found.

candidate phrasal verb	grammatical relations
TAKE ON	part_intrans, PP_*on*-i, NP_part, PP_PP_*from*-i, PP_PP_*in*-i, PP_PP_*of*-i, PP_PP_*with*-i, PP_NP_Ving, PP_PP_*for*-i, PP_Ving, PP_PP_*because*-i, PP_PP_*after*-i, PP_PP_*as*-i, PP_PP_*at*-i, PP_PP_*to*-i, PP_PP_*within*-i, PP_PP_*between*-i, PP_cl_wh, (PP_PP_*by*-i), (PP_PP_*if*-i), (PP_PP_*during*-i), (PP_PP_*through*-i), (PP_PP_*than*-i)
TAKE OVER	part_intrans, PP_*over*-i, part_NP, NP_part, PP_PP_*from*-i, PP_PP_*in*-i, PP_PP_*of*-i, PP_NP_Ving, PP_PP_*for*-i, PP_Ving, (PP_PP_*with*-i)
TAKE UP	part_intrans, Part_*up*-x_obj, AVP, part_NP, NP_part
TAKE TO	PP_*to*-i, PP_PP_*in*-i(?), PP_PP_*of*-i(?), (PP_PP_*with*-i), (PP_PP_*on*-i)
TAKE OFF	part_intrans, part_NP, NP_part, (Part_*off*-x_obj)
TAKE AWAY	part_intrans, AVP, part_NP, NP_part
TAKE OUT	part_intrans, Part_*out*-x_obj, part_NP, NP_part
TAKE DOWN	part_intrans, part_NP, NP_part
TAKE ALONG	part_intrans, PP_*along*-i, (NP_part)
TAKE IN	NP_part(?), PP_PP_*from*-i, (part_intrans)
TAKE THROUGH	(part_intrans)
TAKE TOGETHER	AVP(?)
TAKE ABACK	AVP
TAKE BACK	AVP
TAKE FORWARD	AVP
TAKE APART	(AVP)
TAKE ASIDE	(NP_part)

Key:

( ) – grammatical relations in brackets were identified only after setting the minimum frequency of items to 3, and the maximum number of items in a relationship to 50.

(?) – means that while the phrasal verb occurred in the concordance lines of the grammatical relation, most concordance lines were examples of non-phrasal use.

The relations 'part_intrans' and 'AVP' helped to suggest the greatest number of phrasal verbs; it is noteworthy that these relations suggested different candidate phrasal verbs. The relations 'part_NP' and 'NP_part' suggested several candidate phrasal verbs that were found in the 'part_intrans' relation. The relation 'NP_part' was particularly important because it suggested candidate phrasal verbs that were separable, i.e. could have an object or a pronoun between the verb and the particle.

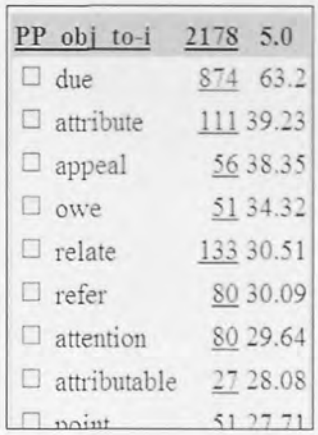
Some candidate phrasal verbs, for example *take on* and *take over,* were found in several different grammatical relations. In fact, most candidate phrasal verbs were found in up to five different relations. A few phrasal verbs were found in only one relation. Nonetheless, considering that word sketches for phrasal verbs are not available, any grammatical relation which suggests candidate phrasal verbs is valuable and examples of it should be saved using the TickBox Lexicography function. This will save time later, when using the Concordance function, to analyse each candidate phrasal verb in more detail.

The number of different grammatical relations a candidate phrasal verb occurs in is a good indication of its frequency. For example, the candidate phrasal verbs *take on, take over,* and *take up* were found in the greatest number of relations and were also the most frequent (of all the) phrasal verbs of *take.* But the frequency of a phrasal verb cannot be obtained by adding occurrences of the phrasal verb in different relations, as grammatical relations may overlap. A good idea of the frequency of a phrasal verb is obtained by combining the frequency of grammatical relations specific to the phrasal verb (grammatical relation = phrasal verb) with the frequency of occurrences in the grammatical relation 'NP_part'. These results always need to be checked by conducting a Concordance search.

6.3.2.2.3 Identifying phrases and idioms

Phrases (including transparent collocations) and idioms often consist of more than two words so Word Sketch does not identify them directly; these multi-word items are normally identified by examining concordances of the collocates. Sometimes the identification of a phrase is very straightforward – the collocate is very salient in a particular relation, and is predominantly found in a particular phrase with the headword. For example, *due,* the most salient collocate by far in relation 'PP_obj_*to*-i' of the noun *fact* (see Figure 65), was found in the phrase *due to the fact that* in 806 out of 874 concordance lines.

Figure 65. DOAE: Meaning analysis - Top collocates of the grammatical relation 'PP_obj_*to*-i' of *fact* (noun).



PP_obj_to-i	2178	5.0
☐ due	874	63.2
☐ attribute	111	39.23
☐ appeal	56	38.35
☐ owe	51	34.32
☐ relate	133	30.51
☐ refer	80	30.09
☐ attention	80	29.64
☐ attributable	27	28.08
☐ point	51	27.71

But often, phrases are easier to miss because they are hidden within relations where one would not necessarily expect them. The noun *fact* can again be used to exemplify this. Relation 'object_of' listed verbs that had *fact* as an object, *give* being one of the most salient verbs. But a quick look at the list of collocates in the grammatical relation did not reveal that *given the fact that* occurs in 166 out of 260 concordance lines of all the collocates in the grammatical relation; this was only discovered after using the Collocation function on the 260 concordance lines, and confirming the findings by doing a search for a phrase *given the fact that* in Concordance.

Meaning pattern definitions, formed during the analysis, were also very useful for identifying phrases, especially frequent ones with a fairly transparent meaning. These phrases were added when the dictionary entry was being compiled (see 6.3.3.3 for more).

Idioms share the same problems of identification as phrases, but cause much more difficulties in terms of dictionary treatment. Their meanings are not a sum of their parts so they need to be accompanied by an explanation.

6.3.2.3 Some limitations of using Word Sketch

Word Sketch is undoubtedly a very useful function in the Sketch Engine for meaning analysis. The analysis of sample entries has however pointed to some underlying problems of analysis with Word Sketch, which lexicographers should be made aware of.

6.3.2.3.1 Things missed by Word Sketch

First and foremost, it should not be assumed that the analysis of grammatical relations and their collocates in the word sketch will account for all the meanings of the word. Meaning

pattern definitions will cover most of the meanings of the word, but the less frequent meanings may not have syntactic patterns and/or collocates that are salient or frequent enough to be picked up by Word Sketch. Low frequency was indeed the most common cause of missed meanings in the analysis for the sample entries, but in total not many meanings were missed by doing the analysis with Word Sketch (see 6.3.3.10, and Table 136 in Appendix 9).

In this Model, the issue of potentially missed meanings was addressed in two ways: by the introduction of Step 3 in the analysis where random concordance lines were examined, and by consulting other corpora and dictionaries in the final stages of compiling the dictionary entry (see 6.3.3.10).

A common problem encountered during the analysis with Word Sketch was missed (high frequency) syntactic patterns. Some high frequency syntactic patterns were identified only after using the combination of intuition and observation analysis of the concordances of collocates in various grammatical relations. An example of such a syntactic pattern is '**as** [Human] **argue**', as in:

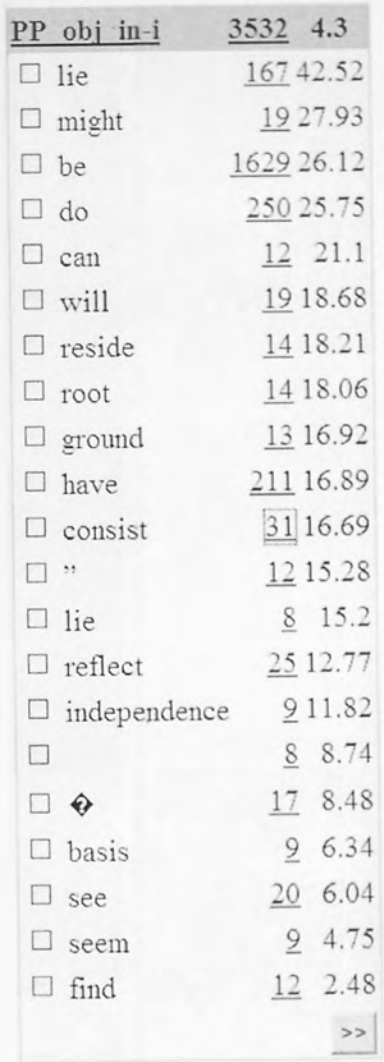
> *Physical matter, as Rose (1995)* ***argues****, is always an 'active ingredient' in the daily churn of human practice and different substances clearly offer us some radically different possibilities.*
>
> (corpus file: Anthrop_11_2006_hitchings)

> *As we* ***argued*** *in Costa and Santesteban (2004b), we believe that this contrasting pattern of results has its origin in the development of language-specific selection mechanisms by highly proficient bilinguals.*
>
> (corpus file: Psych_32_2006_costaetal)

The pattern was only suggested (as very salient) because of my personal experience in academic writing, and the analysis of the concordances of the collocates in the grammatical relations 'pro_subject', 'subject_NP' and, to a lesser extent, 'AVP_mod' confirmed its high frequency (around 1500 occurrences, i.e. 5% of all occurrences of *argue*).

Syntactic patterns can thus be distributed among several grammatical relations and therefore not be signalled by Word Sketch. A syntactic pattern can be also missed if attention is paid to the collocates in grammatical relations, but not the grammatical relations themselves. The pattern *in fact,* for example, was represented in the word sketch of the noun *fact* by the grammatical relation 'PP_obj_*in*-i' (see Figure 66). The pattern was also obscured by the fact that not all collocates in the relation were found in this pattern – *lie* (the most salient collocate), *root, ground,* and *consist* appeared in the syntactic pattern 'VERB + *in the fact (that)*'.

Figure 66. DOAE: Meaning analysis - Grammatical relation 'PP_obj_*in*-i' of *fact* (noun).



PP_obj_in-i	3532	4.3
☐ lie	167	42.52
☐ might	19	27.93
☐ be	1629	26.12
☐ do	250	25.75
☐ can	12	21.1
☐ will	19	18.68
☐ reside	14	18.21
☐ root	14	18.06
☐ ground	13	16.92
☐ have	211	16.89
☐ consist	31	16.69
☐ "	12	15.28
☐ lie	8	15.2
☐ reflect	25	12.77
☐ independence	9	11.82
☐	8	8.74
☐ ◆	17	8.48
☐ basis	9	6.34
☐ see	20	6.04
☐ seem	9	4.75
☐ find	12	2.48
		>>

While *in fact* was still found in the word sketch of *fact*, its sentence-initial variant *In fact* was absent from the word sketch. This was particularly significant considering that *In fact* had 7028 occurrences (over 17% of all occurrences of *fact*). The variant had been identified when adding significant collocates (Step 3 of the analysis); *In* was the fourth most significant collocate on the list (ordered by MI^3 values).

Variants of syntactic patterns may also be overlooked because of the omission of one of the constituent parts of the pattern. The verb *argue* for example was very frequently followed by *that* (*that*-clause), however, there were instances not picked up by the word sketch of *argue* followed by a [zero *that*]-clause (i.e. with *that* omitted):

*I **argue** the implications of such shifts in strategies and scales of resistance reflect an "aboriginal social imaginary," which holds promise for the survival of aboriginal languages as well as meaningful participation in the "modern social imaginary" called modernity.*

(corpus file: Anthrop_13_2006_perley)

This variant was identified during Step 3 of the analysis and while it was much less frequent than *ARGUE that*, it was important to have it recorded in the database entry.

6.3.2.3.2 Problems with statistical information in Word Sketch

Statistical information about the grammatical relations and collocates needed to be handled with care. Sometimes the statistics were only partly correct, whereas in other cases the statistics were completely incorrect due to the nature of information on which they were based. Both types of issues are discussed in this section.

6.3.2.3.2.1 Under-represented grammatical relations and collocates

The frequency of grammatical relations is important information that is recorded in the database, as it influences the decisions made during compiling the entry, such as the ordering of senses, and the inclusion of syntactic patterns and phrases in the entry. The analysis revealed instances of relations being under-represented, i.e. being more frequent in the corpus than shown in the word sketch. Two examples of such relations are shown in Table 60. The case of the passive for *attribute* was not that problematic as the relation was already quite prominent even with 30%. On the other hand, CL (clause) relation for *subsequently* demanded a much more important role in the entry than indicated by Word Sketch.

Table 60. DOAE: Meaning analysis - Two under-represented relations identified in the analysis.

headword	relation	(Word Sketch statistics) frequency + percent of all occurrences of headword	manual search
attribute (verb)	passive (unary relation)	2099 (30.2%)	*BE + attributed* 2879 (41.4%)
subsequently (adverb)	CL (unary relation)	210 (4%)	*Subsequently* 1052 (20%)

Collocates can also be under-represented, but this is rarely due to relations not being identified properly; it is usually because the collocates are dispersed among different relations. An example of such a collocate has been provided in 6.3.2.1.1 (collocate *to* of the verb

attribute). Recording significant collocates at Step 3 was an efficient method of identifying previously under-represented collocates.

The under-representation of collocates can also be caused by variant spellings. Sometimes variant spellings are frequent enough to be easily spotted on the list, as was the case with *favour* and *favor* in the relation 'PP_*in*-i' of *argue* (see Figure 67). In other cases, only one variant spelling may be frequent enough to be listed in the word sketch – it is useful to identify the other variant spelling, if it features on the list; the clustering function in Word Sketch can be used for this purpose. In this Model, variant spellings were recorded as one collocate, with the British English spelling being the main form, and the American English spelling being the variant. Any differences in the frequency of the two spellings were noted.

Figure 67. DOAE: Meaning analysis - The most salient collocates of the grammatical relation 'PP_*in*-i' of *argue* (verb).

PP_in-i	617	1.1
☐ favour	67	48.96
☐ favor	47	42.72
☐ section	67	31.24
☐ paper	35	25.23
☐ article	28	23.69
☐ proof	15	20.08
☐ essay	10	19.6
☐ Section	11	18.91
☐ detail	11	15.44
☐ way	19	14.31

Collocates found mainly in one word form in a particular relation are under-represented in their own way. This has been commented on in 6.3.2.1.1 (see Table 52). The preference for a particular word form indicates a stronger collocation (as opposed to the strength of collocation with the collocate as a lemma), and such information cannot be obtained by looking at the word sketch alone, because all the collocates in the word sketch are listed in their lemma form.

6.3.2.3.2.2 Over-represented grammatical relations and collocates

Collocates can be considered to be over-represented if they come from a small number of texts, or a small range of texts (i.e. a single domain). Collocates displaying these properties need to be pointed out in the database, and concordances more closely examined.

Moreover, relations often contain several collocates that come from a single text. Whereas one collocate from a small number of texts may suggest a technical collocation,

several collocates from a single text may suggest the idiolect of an author or editor. Rather than always creating a note, and to make this feature distinct from other notes, an optional DTD element 'All.from.one.text' with the fixed value 'All concordance lines from a single text.' was created.

If a grammatical relation contains several collocates (or one very salient one) that occur in a small range of texts, that undoubtedly affects the salience of the relation. The relation 'N_mod' of the noun *feature* was one such a relation, as the list of 25 most salient collocates contained 17 collocates from a small range of texts (greyed items in Table 61). In addition, 9 of the 17 collocates were found in a single text only. These 17 collocates represented 35.5% of the total concordance lines of all the 25 collocates – this suggested that in the database entry the prominence of relation 'N_mod', the fourth most frequent relation of the noun *feature,* had to be slightly reduced.

Table 61. DOAE: Meaning analysis - Domain distribution of texts for the grammatical relation 'N_mod' of *feature* (noun).

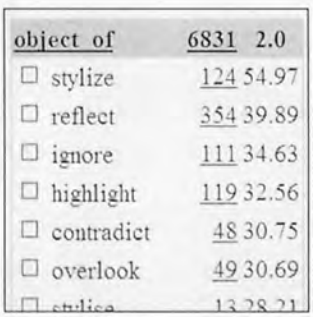
collocate	Frequency	Comments
Epp	33	3 Linguistics texts
dwelling	40	38 out of 39 concordance lines from a single Architecture text
HLAC	15	1 Engineering text
Gabor	15	12 concordance lines from one Computer Science text
webcast	10	1 Economics text
design	111	various
MFCC	6	1 Computer Science text, 2 Engineering texts
MBC	9	1 Computer Science text
GZK	6	1 Physics text
landscape	38	various
novel	27	various
LSI	6	1 Computer Science text
input	41	various
front-end	7	1 Computer Science text, 1 Engineering text (same authors)
DOS	7	1 Chemistry text
texture	16	8 out of 16 examples from a single Computer Science text
surface	48	various
filler	8	1 Business text
ground	28	26 out of 28 concordance lines from one Archeaology text
speech	26	18 out of 26 concordance lines from a single Engineering text
multi-level	6	1 Computer Science text
definiteness	6	2 Linguistics texts
case	82	various
Community	12	1 Business text
privacy	11	10 out of 11 concordance lines from one Architecture text

Another reason for over-representation of relations and collocates is errors in the identification of the relation, or errors in assigning the collocate to the relevant relation. This is discussed next.

6.3.2.3.2.3 Incorrectly identified grammatical relations and collocates

The main reason for relations, and consequently collocates in those relations, being incorrectly identified by Word Sketch was tagging errors. Tagging is the basis for word sketches, so any errors in tagging are reflected in the word sketch. An example of a frequent tagging error was mistaking a noun for an adjective in complex noun phrases. The word sketch for the adjective *potential,* for example, showed a relation 'N_premod' in which *potential* was always used as a noun, but still labelled adjective. Occurrences of incorrect identification of relations were not difficult to spot. It was more important, however, to make a note if the frequency of incorrectly tagged items was high, as with *potential.*

Collocates listed in the wrong relation, i.e. given the wrong word class, were also quite common. The most common mistake in identification was associated with adjectival uses of verbs. Two examples are presented below, namely *stylized* and *defining*; both were listed as verb lemmas under the relation 'object_of' in the word sketches of the nouns *fact* and *feature* respectively, and both were always used as adjectives.

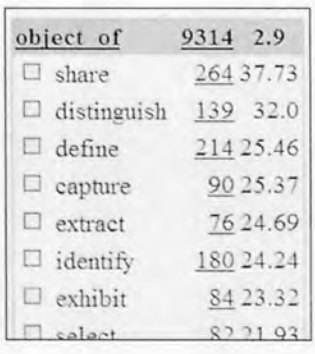


object_of	6831	2.0
☐ stylize	124	54.97
☐ reflect	354	39.89
☐ ignore	111	34.63
☐ highlight	119	32.56
☐ contradict	48	30.75
☐ overlook	49	30.69
☐ stylise	13	28.21

a. *We document this new **stylized fact** in the data for all the G7 countries.*
(corpus file: Econ_74_2007_albuquerqueetal)

b. *Yet it is a **stylized fact** that almost all NFP hospitals have debt obligations (p. 21).* (corpus file: Fin_22_2006_jegers+verschueren)

c. *This captures the **stylized fact** that the larger part of public consumption is directed towards domestic markets* (corpus file: Fin_75_2007_spange)



object_of	9314	2.9
☐ share	264	37.73
☐ distinguish	139	32.0
☐ define	214	25.46
☐ capture	90	25.37
☐ extract	76	24.69
☐ identify	180	24.24
☐ exhibit	84	23.32
☐ select	82	21.93

a. *This relational context is a common **defining feature** of social capital.*
(corpus file: Educ_55_2007_mcgonigaletal)

b. *One of the **defining features** of representative democracies is periodic elections.* (corpus file: Econ_121_2006_list+sturm)

c. *The iterative, two-way flow of communication becomes as much a **defining feature** of the relationship as the flow of products and money.*
(corpus file: Business_46_2004_dawar)

Incorrectly tagged collocates like this may still need to be recorded in the database, but under the appropriate relation. *Stylized* and *defining*, for example, were recorded under the relation 'AJ_premod'.

6.3.2.3.3 Limitations of selecting examples with TickBox Lexicography

TickBox Lexicography is a useful function in Word Sketch, as it allows information on grammatical relations and collocates to be saved in the dictionary database. Nonetheless, certain problems with TickBox Lexicography arose during the analysis.

One of the problems concerned the quality of examples offered by TickBox Lexicography. It appeared that the examples offered by TickBox Lexicography often came from a single domain, or even a single text[100]. The problem seems to be that examples are not selected randomly, but according to the alphabetical order of the subcorpora and filenames (thus, Anthropology examples were more likely to be selected than Veterinary Science examples). The representation of examples could only be established once they were copied into an XML file, since TickBox Lexicography does not provide Doc ID information (as the Concordance display does). Furthermore, the GDEX (Good Dictionary Examples) function, which would have been useful, could not be used in Word Sketch.

There was no easy solution to this problem – concordances of the problematic collocates (where examples provided by TickBox Lexicography were not useful) were examined, and examples were copied and pasted into the database. This proved to be very time-consuming, and should be avoided in the design of the actual dictionary.

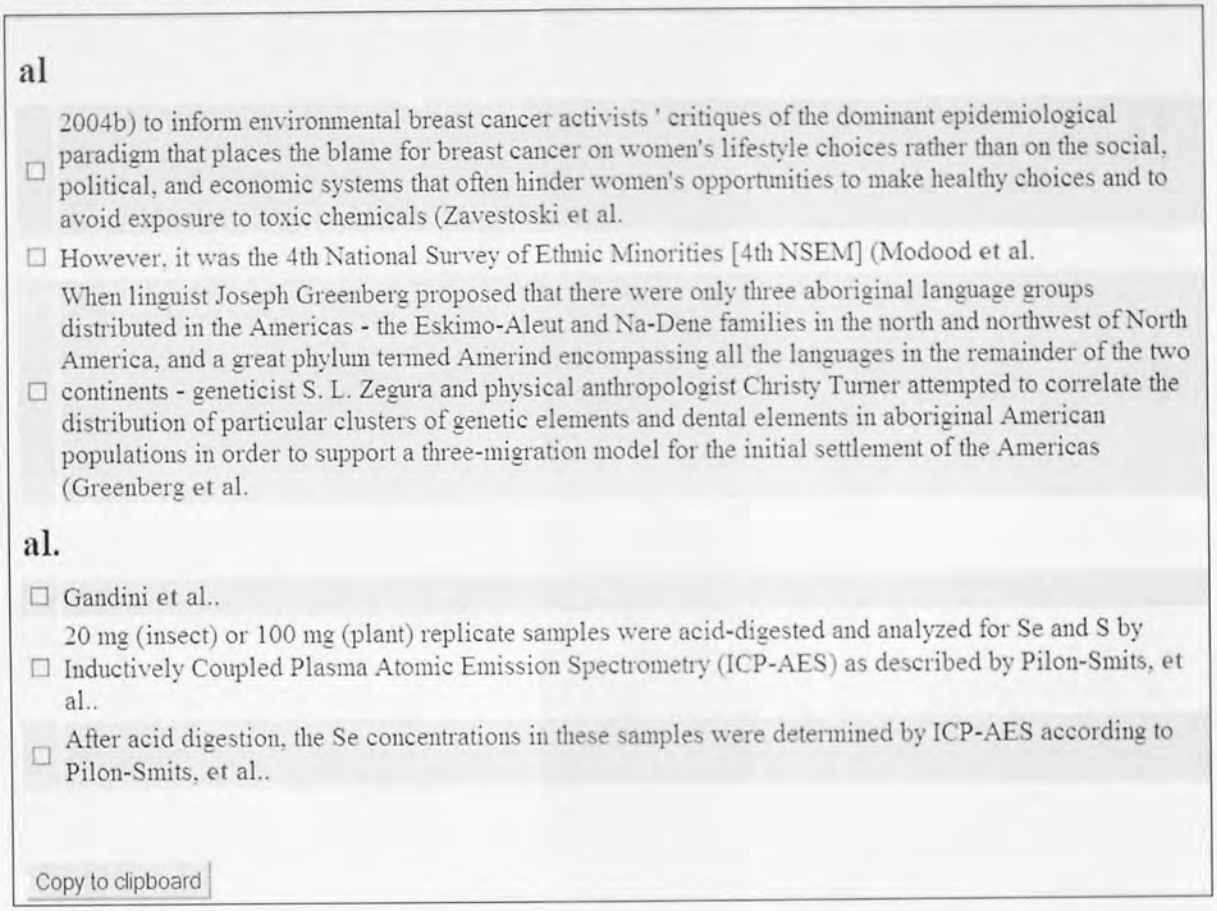
Other problems were technical in nature. These problems were mainly encountered when transferring data from Word Sketch to the database. One problem occurred during the first step, i.e. before copying the example(s) to Clipboard. Sketch Engine considers full stops that are followed by a space as sentence breaks. Also, if a collocate is followed by a full stop, such as the collocate *al.* of the noun *et*, example sentences are not saved in full (Figure 68 below).

Another problem is that TickBox Lexicography allows only the sentence containing the collocate to be saved. This can be problematic when the preceding or following sentence is essential for exemplifying the usage of the headword. An example is the adverb *thus*, which was often sentence-initial. This was indicated by the unary relation 'CL' in the word sketch of *thus*. However, TickBox Lexicography saved only examples containing *thus*, so the preceding

[100] Of course, this is not a problem with less frequent collocates as all the examples, rather than just a supposedly random selection of examples are provided.

sentence or paragraph of text needed to be added by conducting a phrase search for a (longer) sequence of words from the example, and only then copying and pasting the text in the database and joining it to the *thus* example.

Figure 68. DOAE: TickBox Lexicography - Examples for collocates *al* and *al.* of the noun *et.*



Technical problems can also occur when importing data from Sketch Engine into the TshwaneLex database. Double quotes, for example, are used by XML for indicating element/attribute values and are therefore not allowed in the text. If the text contains double quotes, the TshwaneLex software issues an error message when importing the XML file into the database. The XML file (Figure 69) of the collocate *calculation* of the noun *method* can therefore only be saved into the TshwaneLex database once all the double quotes in the example are removed or, as was done in this research, are changed to single quotes. If more than one example is being saved at the same time, only the examples up to the one containing the error are saved in the database.

Figure 69. DOAE: TickBox Lexicography - The XML file for the selected example for the collocate *calculation* of *method* (noun).

```
<Dictionary>
 <Language>
 <Headword HeadwordSign="method">
 <gramrel grname="PP_for-i">
   <collocation collo="calculation">
    <Example Example.number="1" Database.DomainLabel="Biochemistry"
     Example="Limits of individual SH3 domains were determined by a Pfam database
     HMM search Subsequent analyzes were performed with version 3.1 of the "Molecular
     Evolutionary Genetics Analysis" (MEGA) package The initial alignment was created
     with the incorporated ClustalW algorithm The phylogenetic tree was constructed by
     using the "neighbor-joining" algorithm with "number of differences" as method for
     calculation of distances between the sequences and "complete deletion" for treating
     gaps in the alignment." Source="Biochem_14_2006_harkiolakietal" />
   </collocation>
 </gramrel>
 </Headword>
 </Language>
</Dictionary>
```

6.3.2.4 Analysis of infrequent headwords with Word Sketch

Despite some problems, word sketches are very useful for meaning analysis, especially in the case of very frequent items. But what about low frequency items? Low frequency may adversely affect on the usefulness of word sketches, as they are based on a statistical measure that contains a variable based on the joint frequency of three different elements (headword, grammatical relation, and collocate)[101]. To test this, several low-frequency headwords were analysed.

The sample headword used as an example here is *assortment* (frequency = 346). Word Sketch produced 18 different grammatical relations (Figure 101 in Appendix 9). Several grammatical relations were quite frequent, for example 'AJ_premod' (found in 45% of occurrences of *assortment*), 'PP_*of*-i' (25% of occurrences), 'object_of' (23 % of occurrences), and 'N_mod' (14% of occurrences). Not many collocates stood out in terms of frequency; in fact, it was more common that the relation contained many collocates of similar frequency. Most collocates that did stand out, such as *large* and *small* in 'AJ_premod', and *impact* in 'PP_obj_*of*-i', came from a single text so their salience was somewhat misleading.

[101] See 3.3.1.2.2 for details.

The meaning analysis (Step 2) identified one meaning (a selection of things), a related syntactic pattern '*assortment of* something'. A few salient collocates (*product(s), choice(s), impact(s), large, consumers*)[102] were recorded (Step 3). Furthermore, due to their low number, all the concordance lines could be examined for any missed meanings; no additional meanings were found.

The main point is that all the information about *assortment* was obtained without adapting or expanding the three-step meaning analysis process. The same is true for other low-frequency sample entries that were analysed. An alternative approach to analysis with Word Sketch, similar to the one described in 6.3.2.1.4[103], also proved useful at certain entries.

6.3.2.5 Analysis of headwords partially covered by Word Sketch

Some headwords have variant single- and multi-word forms, of which only the forms without spaces (found in the lemma list) can be analysed with Word Sketch. *State-of-the-art* (frequency = 245) is an example of such a headword; identified as an adjective, it can be analysed with Word Sketch. This is not the case with its variant *state of the art* (frequency = 117). The analysis was conducted in two steps: first, the word sketch of *state-of-the-art* was analysed using the three-step process, followed by the analysis of the entire concordance of *state of the art.* The word sketch pointed to one predominant grammatical relation ('premod_N'), indicating that *state-of-the-art* is predominantly used attributively, and is used to describe things (e.g. *technology, product*) or activities (*method, technique*). In addition, examination of the concordance revealed that *state-of-the-art* is also used as a noun. An analysis of the concordance for *state of the art* confirmed noun uses; in fact, noun uses were more typical of the form *state of the art*. Collocates of *state of the art* were compared to the ones of *state-of-the-art,* and examples and statistical information of the most significant collocates of *state of the art* were recorded in the database[104].

But headwords with variants are not the only ones requiring the analysis of some occurrences with Word Sketch, and others with Concordance. A similar approach is needed for

---

[102] A note is made about the fact that all but two occurrences of the collocates *large* and *impact* come from one Economics text.

[103] The approach involves first examining all the whole concordance for meanings, and then using Word Sketch to assign relations and collocates to those meanings.

[104] If the collocate of *state of the art* was the same as the collocate of *state-of-the-art*, statistical information was combined.

nouns that can be used as proper nouns. Such an example is the noun *authority*. The proper noun *Authority* had 541 occurrences, and since capital initial forms could not be analysed using Word Sketch, these concordance lines needed to be analysed separately. Sometimes this is not a problem as the noun and the proper noun are separate headwords. In the case of *Authority* and *authority*, the proper noun was simply a variant of the noun, or rather one of its senses. The procedure used here was the same as described for *state-of-the-art;* a Word Sketch analysis of *authority* was conducted first, followed by an analysis of *Authority* using Concordance which included comparing meanings, syntactic patterns, and collocates with the ones identified for *authority*.

6.3.2.6 Analysis of headwords not covered by Word Sketch

Some headwords cannot be analysed with Word Sketch. These include proper nouns, pronouns, and conjunctions. Word Sketch can be of some use for multi-word items such as compounds and phrasal verbs (see 6.3.2.2.1 and 6.3.2.2.2). Most headwords that cannot be analysed with Word Sketch tend not to be highly polysemous. The analysis of such headwords relies completely on Concordance and its functions, especially the Collocation function. This needs to be used extensively as it represents a form of word sketch, except that collocates are not ascribed to specific grammatical relations.

The conjunction *albeit* is used to exemplify the analysis without Word Sketch. It has 2060 occurrences in the corpus. First, the list of collocates (span -5+5) was created using the Collocation function. Comma was the most significant collocate by T-score or MI^3 value. Examination of concordance lines with commas showed that a comma was very often found immediately before *albeit*. The list of collocates in the span -1+0 confirmed this pattern; in fact, in 75% of concordance lines, a comma was found one position to the left of *ALBEIT*. What is more, *albeit* was always preceded by punctuation (Table 62 below). Several examples with a comma found immediately before *albeit* were recorded in the database, as well as a few examples with an open round bracket preceding *albeit*, for example '*We later show how to extend the algorithm to higher dimensions (albeit with weaker performance bounds')*.

Table 62. DOAE: Meaning analysis - Collocates of *albeit* (span -1+0), ordered by MI[3] score.

	Freq	T-score	MI	MI3
,	1541	36.666	3.922	25.102
(	315	16.152	3.475	20.073
◆	20	4.365	5.390	14.034
-	8	2.787	6.095	12.095
-	17	3.845	3.891	12.066
--	4	1.961	5.694	9.694
.	29	-7.617	-1.272	8.444
;	6	-0.150	-0.086	5.084
"	6	-0.280	-0.156	5.014
)	6	-9.997	-2.345	2.825

Settings used: minimum frequency of a collocate in a corpus = 5
minimum frequency of a collocate in the span = 3

The focus of the analysis then shifted to the collocates that were found after *albeit*. The list of collocates (see Table 131 in Appendix 9 for the list of top 48 collocates) revealed that *albeit* was often associated with words having negative connotations, i.e. *albeit* had a negative semantic prosody (Sinclair, 1991; Louw, 1993). *Albeit* was often followed by an adjective or adverb that indicates a thing or action that is incomplete (e.g. *limited, somewhat, slightly, small*), diminished (e.g. *less, lesser, lower, weaker, smaller, reduced*), negative (*not, inefficiently, without)* or shows some degree of negativity in manner (*reluctantly, redundantly, weakly, only, weak, tentatively*). Concordances of these collocates were examined, and examples saved in the database. Another finding, revealed by both the list of collocates and manual inspection of concordances was that nouns were rarely found among the words following *albeit*.

The meanings of *albeit* were noted during the examination of concordances of the collocates. These meanings were confirmed, and any new meanings added, by the analysis of a random set of concordance lines for *albeit*. The number of random concordance lines depends on the frequency of the headword – the more frequent it is, the greater the number of random lines that need to be analysed. But even with less frequent headwords such as *albeit*, at least 25% of the concordance lines should be examined.

The analysis remains broadly the same, whether or not Word Sketch can be used. The focus remains on collocates, and their connection to meaning. The difference is the more automatic steps of the Word Sketch analysis (e.g. grammatical relations) need to be done manually, such as saving examples.

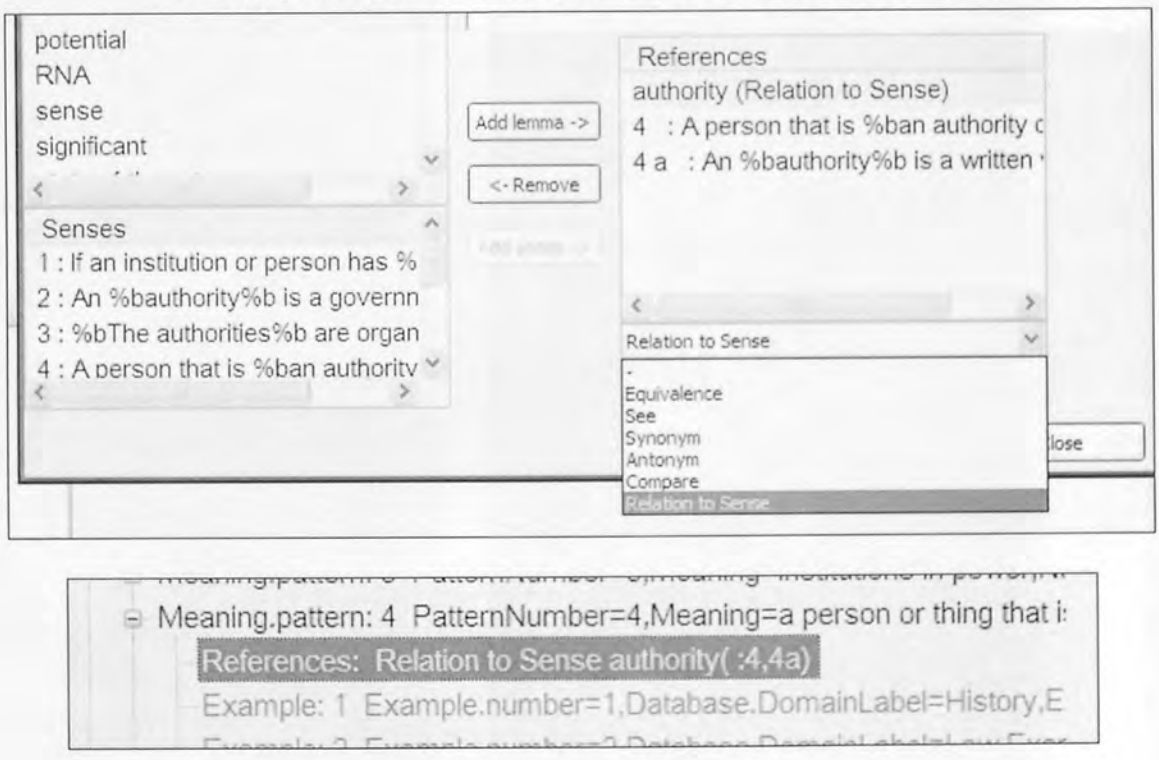
6.3.3 Compiling dictionary entry

Meaning analysis produces a great deal of information about the headword, but that information is in a form that is useful for lexicographers and not the target users. To make it user-friendly, the information needs to be carefully selected, modified, and complemented.

6.3.3.1 Converting meanings into senses

The conversion of meanings into senses starts with making each meaning pattern into a sense, which involves copying all the information (grammatical relations, meaning pattern definitions, pattern elements, collocates, and examples) into the dictionary entry. The information is likely to be merged or reduced (e.g. less salient collocates and examples will be discarded), and it is much easier to copy all the data from the meaning analysis and edit it, than going through the data in the meaning analysis part of the database and selecting what will end up in the entry. Also, working on a copy of the data prevents any loss of information.

Figure 70. DOAE database: Establishing cross-reference link between a meaning pattern and sense(s) (above). The link dispayed in the database entry (below).



As soon as senses are created, the link between the meaning pattern and its corresponding sense(s) needs to be established. This is done by opening the element Reference

in the meaning pattern, selecting the relevant sense, and selecting cross-reference type 'Relation to Sense' (Figure 70 above). By doing this, any subsequent changes to the sense ordering are immediately updated in the Reference, and the link between the meaning pattern and its related sense(s) is always maintained[105].

The process therefore starts with the assumption that the meaning patterns equal senses in the dictionary entry. This is in fact a justified approach considering that frequencies of meaning pattern definitions, grammatical relations, and collocates, are used to determine the order of meaning patterns (see 6.3.2.1.2).

6.3.3.2 Definitions

Definitions lie at the centre of a lexicographer's work, and are the main feature by which the dictionary is judged. Attributing a high importance to the accuracy and clarity of a definition is justified, as research shows that the definition is the entry feature that students consult most frequently (see 2.3.2.1 and 4.1.1.7).

Various types of definition available to lexicographers were considered in the selection of type(s) of definition to be used in DOAE. Special attention was paid to the types of definition found in dictionaries that students currently use. Two groups of definition emerged: main definitions, and supporting definitions (they provide additional information, and are dependent upon the main definition).

6.3.3.2.1 Main definitions

The two types of definition considered for main definitions in DOAE were full-sentence definitions and traditional definitions. Lexicographic literature (e.g. Zgusta, 1971; Ilson, 1987; Svensén, 1993; Landau, 2001) mentions other types of definition (see Table 63 for examples) such as paraphrases (definition by (near-)synonym), ostensive definitions (e.g. illustrations) and extensional definitions (a list of (all) concepts related to the word defined), but these were considered to be more appropriate for supporting definitions (see 6.3.3.2.2) or were not considered suitable for DOAE. Nevertheless, sometimes other types of definition may still be used in DOAE, if they are more effective than the full-sentence definition or the traditional definition – for example, the sense of the noun *colour* defined in Table 63 (the sense is also

[105] If a sense is later omitted from the dictionary entry, the cross-reference link is also deleted, and the empty Reference element in the meaning pattern indicates that the meaning pattern is not represented in the dictionary entry.

found in CAJA) may be best defined with an extensional definition similar to the one used by e-LDOCE.

Table 63. Examples of types of definition not used in DOAE.

paraphrase	**tired** *adj.* 1. weary; fatigued (CED CD-ROM)
extensional definition	**motor vehicle** *n.* car or motor cycle or moped or van or… (Svensén, 1993:124) **colour** *noun* 1. red, blue, yellow, green, brown, purple etc (e-LDOCE)
'When' definition	**analysis** *noun* when you analyse something (e-CALD) **disposal** *noun* when you get rid of something (e-LDOCE)

'When' definitions, used in some learners' dictionaries were also considered. This type of definition is basically a shorter version of the full-sentence definition (without the first part), but still avoids the abstraction of the traditional definition (Osselton, 2007). A study by Lew and Dziemianko (2006) reveals one of potential problems of using 'when' definitions, as their subjects had problems in identifying the word class of the words being defined. Hence, Atkins and Rundell (2008) advise against using this type of definition until more research is conducted on its benefits and pitfalls for dictionary users.

6.3.3.2.1.1 Full-sentence definition

Full-sentence definitions, as the name suggests, provide the definition in the form of a complete sentence, with the word defined (definiendum) embedded in it. Full-sentence definitions were first used in the first edition of the Cobuild dictionary, published in 1987. This style of definition has now been adopted by other monolingual learner's dictionaries, however the Cobuild dictionaries remain the only ones to use full-sentence definitions consistently throughout the dictionary.

A full-sentence definition consists of two parts which are joined by a 'hinge' (Barnbrook, 2002). A hinge is most often a verb, however a conjunction (e.g. *if*, *when*) is sometimes used. Each part of the definition has a different function, the first part showing the typical usage, and the second explaining the meaning (Hanks, 1987). The usage component represents the biggest departure from the traditional defining style, as it incorporates the encoding information in the definition (Hanks, 1987; Béjoint, 2000). Barnbrook (2002) identifies seventeen definition types and groups them into four categories, while Hanks (1987) provides a detailed analysis of the definition strategies.

> If you **are entitled** to something, you have the right to have it or to do it.
> (*entitle*, sense 1 - COBUILD CD-ROM)

As demonstrated by the definition of *entitle*, the definiendum in a full-sentence definition is found in the most frequently used form. This follows Sinclair's idea that "every distinct sense of a word is associated with a distinction in form" (Moon, 1987:89). A full-sentence definition shows the lexical item in its typical use, showing its 'selection preferences'[106], i.e. the syntactic patterns the item normally occurs in. The definition of *entitle* thus tells the user that the subject is a person, *entitle* normally occurs in passive, is followed by *to,* and the object is a thing.

Full-sentence definitions are particularly well-suited to the needs of students. Many important types of information that students rarely consult (grammar, usage, collocates) are incorporated in the definition. Moreover, the full-sentence definition can offer information which would normally be provided in the form of grammar labels such as *intransitive*, *modifier*, and *with adverbial of direction*, which can be very user-unfriendly and perhaps overlooked or not understood by users (Atkins & Rundell, 2008).

Full-sentence definitions have been closely scrutinized by lexicographers (Fillmore, 1989; Hausmann & Gorbahn, 1989; Herbst, 1996; Landau, 2001; Sinclair, 2004a; Rundell, 2006; Atkins & Rundell, 2008). Most frequently mentioned problems are length (or non-economical nature if full-sentence definitions are applied throughout the dictionary) and the fact that full-sentence definitions do not always work. Addressing the criticism of length, Sinclair (2004a) argues that full-sentence definitions provide a great deal of information that other dictionaries offer in the form of notes.

Full-sentence definitions work better for some words than others. They are very effective for defining intransitive verbs, some reflexive verbs, phrasal verbs, and many adjectives (Landau, 2001; Rundell, 2006; Atkins & Rundell, 2008). In addition, full-sentence definitions are effective for defining words with limited selection preferences, for example for verbs occurring mainly in the passive voice (Herbst, 1996; Rundell, 2006; Atkins & Rundell, 2008).

6.3.3.2.1.2 Traditional definition

The traditional definition, or the lexicographic definition, is "the most prestigious type of definition" (Béjoint, 2000:198). The traditional definition is found in all monolingual dictionaries of English, but learner's dictionaries use it to a much lesser degree. The definition

[106] Term used by Hanks (1987), deemed to be more appropriate than 'selection restrictions'.

uses a genus-and-differentia model: the genus part contains a superordinate term of the defined word, and the differentia (or differentiae) part includes one or more features that distinguish the defined word from other similar words. In the definition of *scalpel* from NODE CD-ROM, 'a knife' is the genus, and the rest of the definition represents the differentiae.

scalpel *n.* a knife with a small, sharp, sometimes detachable blade, as used by a surgeon

When selecting the genus, "the lexicographer must be looking to strike an ideal balance between maximum defining power of his genus word (not too general) and comprehensibility (not too specific)" (Ayto, 1983:90). The same is true of the differentia where the right number of distinguishing features need to be listed to avoid making the definition too general (too few features), or too specific (too many features).

The traditional definition does not produce the best results for all categories of words. As Atkins and Rundell (2008:415-416) point out, it is effective for most nouns (especially for the ones referring to concrete objects) and many classes of verbs, but less effective for most adverbs and adjectives.

There is therefore no single type of definition that is the most effective for all words in the dictionary. So why not select the most appropriate type of definition for individual entries, or parts of entries, rather than the entire dictionary?[107]

One argument against using more than one type of definition in the dictionary would be that users might be confused by the lack of consistency. But users rarely know, or want to know, their dictionary in such detail; they are more interested in decoding or encoding meanings. Besides, dictionaries are normally consulted for one item, maybe two, at a time, so the likelihood of users remembering the type of definition encountered in the last consulted entry is very low.

Full-sentence definitions and traditional definitions are considered the most appropriate types for the main definition in DOAE. The selection of the type of definition for a specific sense or word class takes into account not only the effectiveness of the definition, but also the type of definition used in other senses, or related entries.

[107] A similar approach was suggested by Béjoint (2000), and has already been used by many monolingual learner's dictionaries (Rundell, 2006).

6.3.3.2.2 Supporting definitions

Three types of supporting definition used are brief definitions, definition by synonym or antonym, and illustrations. Synonyms and antonyms, and illustrations are substantially different and therefore discussed later, so only brief definitions are presented here.

6.3.3.2.2.1 Brief definitions

A brief definition is a shorter version of the main definition, and can help the user to find the relevant sense more quickly, or can even act as an independent definition for users who only want to get a quick idea about the meaning of the lexical item. The form and the content of brief definitions are very similar to a definition in the form of a paraphrase, with one important difference: a brief definition's main role is to refer the users to the main, and longer, definition, not to act as a self-sufficient definition.

Two types of brief definitions are used in this Model[108]: quick definitions[109], and mnemonics (in menus). Quick definitions are currently found only in the Encarta World English Dictionary Online, a dictionary for NSs. Quick definitions use many words from main definitions, and reflect the word class of the word defined (Figure 71). In addition to helping the users to find the relevant part of the entry quicker, quick definitions can act as standalone definitions for users who do not require more information about the item.

Figure 71. Quick definitions in the entry for *elbow* (Encarta World English Dictionary Online).

[108] Signposts, another type of brief definition, were also considered useful but were not used due to their incompatibility with quick definitions.
[109] Expression from the *Encarta World English Dictionary* (as cited by Atkins and Rundell, 2008).

Quick definitions are thus a useful alternative for students who will not need, or want, to use the main definition. Quick definitions will be provided with most full-sentence definitions, but only with longer traditional definitions; there is no point in offering a quick definition if the main definition is already short (e.g. see sense 2 in Figure 71).

Quick definitions may also be a useful basis for mnemonics in menus. Mnemonics in menus are supposed to give the users a quick idea about the content of the main definition, helping them to find the relevant part of the entry more quickly (for more on menus, see 6.3.3.9.1).

6.3.3.2.3 DOAE sample entries: Examples of writing definitions

6.3.3.2.3.1 Example of a full-sentence definition: sense 1 of *attribute* (verb)

The first step in writing a definition for sense 1 included recalling the information for the Meaning pattern 1 (Table 64).

Table 64. DOAE database: Pattern definitions, pattern types, and meaning of Meaning pattern 1 of *attribute* (verb).

Pattern definitions	[Event 1] **be attributed to** [Event 2] [Human] **attribute** [Event 1] **to** [Event 2] [Event 2] **to which/whom** [Event 1] **be attributed** **be attributed to the fact that**
Pattern elements	Human: *scholar(s), author(s), consumer(s), researcher(s), he, we, they* Event 1: *difference(s), effect(s), increase, result(s), finding(s), outcome(s), decline, failure, success* Event 2: *fact, difference(s), effect(s), factor(s), lack, increase, change, property(ies)*
Meaning	If you attribute something, such as a finding, to something, you perceive that thing to be the cause of the finding.

The explanation of the meaning pattern was already written in full-sentence style and reflected the first pattern definition. But the verb *attribute* was often found in the passive voice (42% of all occurrences), and earlier analyses of the most salient collocates and the grammatical relations, and 250 randomly selected concordance lines, had shown that the passive was especially prevalent in Meaning pattern 1. On the basis of this information, the explanation of the meaning pattern needed to be revised so it reflected the second pattern definition:

> *If something, such as a finding,* ***is attributed to*** *something, that thing is perceived to be the cause of the finding.*

In accordance with the COBUILD style, the typical form of the verb *attribute* is displayed in bold. The preposition *to* is also displayed in bold because it is a very strong collocate of *attribute*.

The left side of the definition needed improving as it contained *something* twice (indicating Event 1 and Event 2 respectively). This formulation was likely to confuse the user, so a more specific wording was needed for at least one Event. Event 1 was already partly specified with *finding*, which was one of the collocates. However, a closer inspection of concordance lines of Event 1 collocates showed that *finding* was also one of the appropriate superordinate terms, the other one being *result*. In other words, Event 1 predominantly denoted some finding or result of a study.

Finding a suitable superordinate term for Event 2 was slightly more problematic, because collocates in the position of Event 2 indicated a variety of concepts (e.g. *fact*, *finding*, *property*). But since the wording for Event 1 was now more specific, *something* was used for Event 2. So, after revising the left side, the definition was:

> *If a finding or result* ***is attributed to*** *something, that thing is perceived to be the cause of the finding.*

The right-hand part of the definition did not match the structure of the left-hand part, so *that thing*, which referred to the object *something* in the left-hand part, was a subject in the right-hand part of the definition. Furthermore, the verb *perceive* needed to be replaced with a more frequent verb with a similar meaning, for example *consider, believe* or *see*. After the changes, the definition was:

> *If a finding or result* ***is attributed to*** *something, the finding or result is believed to be caused by that thing.*

The right-hand part now contained a rather complex phrase *is believed to be caused by*. The phrase served two functions: it explained the meaning of *attribute* and implied the presence of the subject. It was considered to be better to present these two pieces of information separately. The phrase was also rare (5 occurrences in the BNC, 9 in CAJA); the same was true of other candidate phrases; *is considered to be caused by* (0 occurrences in the BNC, 2 in CAJA), and *is seen to be caused by* (1 occurrence in the BNC, 2 in CAJA). Hence, it was better to split the phrase, and use the two verb phrases in separate clauses:

> *If a finding or result* ***is attributed to*** *something, it is believed that the finding or result was caused by that thing.*

The final definition now contains the phrase *it is believed*, which is frequent (338 occurrences in the BNC, and 248 in CAJA), and offers the verb *cause* in the passive voice, and in past tense, which more accurately reflects the meaning of *attribute*.

6.3.3.2.3.2 Example of a traditional definition: *method* (noun)

Although the noun *method* had only one meaning pattern, it was very frequent and was found in many different grammatical relations. *Method* was frequently modified by an adjective or a noun, and/or was an object of a verb, however there were no groups of collocates that could have been used to write a full-sentence definition. Two frequent phrases were found, '*method for doing something*' and '*method of doing something*', but those were offered as constructions under the sense.

Examples indicated that a method was a way of doing some complex activity, such as *analysis* or *inquiry*. So this was used as the point of departure for the definition:

a way of doing some activity

Some activity was rather unclear and needed to be specified. Help was sought in collocates of the grammatical relations 'AJ_premod', 'N_mod', and 'PP_of-i'. It quickly became obvious that the activity is normally research-related (e.g. *analytical method, qualitative method, Monte Carlo method, method of inquiry, method of analysis, method of collection*, etc.), and uses a specific set of steps, so the definition was made more specific:

a systematic way of doing research

Nonetheless, there were examples of collocates which did not necessarily imply research activity, such as *new method, contraceptive method, teaching method, method of assessment*. To allow for such instances, the definition had to be extended:

a systematic way of doing research or some other activity

Because other activities, and even some research, are not always done systematically, the decision had to be made whether to indicate this somehow to the user, or to omit *systematic* altogether. The former option was taken, so the final definition is:

a (systematic) way of doing research or some other activity

This definition still stresses the predominant use of *method* in research-related activity, but also covers its use for other, non-research, activities.

6.3.3.2.3.3 Example of quick definitions: *authority (noun)*

The noun *authority* has nine senses, and the full-sentence definition form has been used for all of them. As some of the main definitions were quite long, shorter versions, i.e. quick definitions, were written. All the definitions are presented in Table 65.

The wording of each quick definition has been derived from the wording of the main definition, more specifically its right-hand part. Most information from the main definition has been retained, although some of it may only be implicit. The essential information refers to what the sense is about, so what would be the differentia(e) in the traditional definition. Thus, the quick definitions tell the user that Sense 1 is about power/control (something abstract), Sense 2 and Sense 3 refer to an organization and organizations, Sense 4 and 6 refer to a person, Sense 5 refers to a piece of writing, Sense 7 and Sense 8 refer to abstract notions of permission and personal quality respectively, and Sense 9 refers to a type of internet page.

Table 65. DOAE: Quick definitions and main definitions of all nine senses of *authority* (noun).

Sense	quick definition	main definition
1	the power to control people or activities	If an institution or person has **authority**, they have the right or power to control people or activities.
2	a government department	An **authority** is a government department or official organization that is responsible for certain area of activities, and has the power to make decisions.
3	**(the authorities)** organizations in charge of a country	**The authorities** are organizations or people that are in charge of a certain country or area.
4	an expert	A person that is **an authority on** something is considered to be an expert on a particular subject.
5	an important written work	An **authority** is a written work that is often cited in support of a particular argument.
6	a person with power	An **authority** is a person in a position of power.
7	official permission	**Authority** is official permission to do something.
8	personal quality	if someone has **authority**, they are knowledgeable or behave in a way that other people listen to them.
9	a type of internet page	An **authority** is an internet page that has many citations pointing to it.

Quick definitions tend to be short; some are significantly shorter than the main definitions, like the quick definition for Sense 2. For this particular quick definition, it was thought that 'a government department' would suffice since the phrase already implies some type of control and power to make decisions.

There is a similarity between Sense 2 and Sense 3, which can also be observed in the wording of the quick definitions. The key to making the distinction is the phrase *the authorities* (provided in bold) in the quick definition of Sense 3. This information from the left-hand of the definition not only helps the user to distinguish between the meanings, but also presents the phrase or form in which Sense 3 is found.

6.3.3.2.4 Principles of definition writing

There is no standard procedure for writing a good definition, but according to the lexicographic literature, several rules need to be followed to achieve good defining practice. The most frequently mentioned rules are substitutability (the definition can substitute the defined word in any given context), reflection of grammatical function (the definition reflects the word class of the word being defined), no circularity (do not define the word with related words[110]), and closed dictionary (all words in the definition must be defined in the dictionary). Substitutability is the most contentious of the four rules because it is often impossible to apply, for example when defining function words (Stock, 1988; Béjoint, 2000) and scientific terms (Landau, 2001).

In addition, there are many principles that refer specifically to the wording of the definition. Zgusta (1971:257) pointed out that the definition should not "contain words more difficult to understand that the explained word itself". The principle is also mentioned by modern books about lexicography (Landau, 2001; Atkins & Rundell, 2008), but with a caveat that it cannot always be applied, especially for more frequent words.

Definitions also need to be brief. Brevity is mainly dictated by the need to save space in the dictionary. Stock (1988:81) rightly states that "(l)exicographic definitions are generally competing for space against other elements in the dictionary — pronunciations, etymologies, etc — and more particularly against the extent of the headword list itself." There are cases when it is neither possible nor desirable to produce a brief definition because that would result in breaking one or more of the principles mentioned in the earlier paragraphs. Therefore, rather

[110] This is a simplified version of Landau's (2001:158) rule: "No word can be defined by itself, and no word can be defined from its own family of words unless the related word is separately defined independently of it."

than stressing brevity, it is better to say that "definitions should not waste words" (Landau, 2001:170).

Lexicographers therefore need to observe several principles when writing definitions. Yet, in reality, there is often no time nor space to accommodate all these requirements – the main problem of lexicographers is "to pack enough information into a definition to suit the needs of a wide range of users" (Atkins, 2008:45). The users are the most important judges of the usefulness of the definition; if the users fail to understand it, the dictionary fails. In the words of Bolinger (1965:572), "dictionaries do not exist to define, but to help people grasp meanings".

6.3.3.2.5 Some considerations

In addition to lexicographic principles, there are some other considerations that need to be borne in mind when writing definitions. Some of these considerations are addressed in this section.

6.3.3.2.5.1 Considering other senses

One important item that is not mentioned among lexicographic principles of good definition writing is that the process of writing a definition needs to take other senses of the entry into consideration. This is necessary to ensure that a distinction between the senses is made, which in turn avoids confusing the user with definitions that are very similar.

The principle of considering other senses was frequently followed when compiling the sample entries. The process is exemplified by Sense 1 and Sense 2 of the noun *attribute*. The similarity between the senses became immediately apparent from the explanations written during meaning analysis:

1. *characteristic or feature*
2. *positive characteristic or feature*

The same synonyms were provided in both explanations, and synonyms rather than superordinates had been identified as candidates for the genus word in both definitions. An analysis of synonym candidates offered by the Thesaurus function (Table 133 in Appendix 9) and the grammatical relation 'and_or' in the word sketch (Table 134) identified five words that are synonymous to one or both senses: *characteristic, feature, quality, trait,* and *property. Trait* was discarded on the basis of low frequency, and *property* was discarded because its core

meaning (assets or belongings) was more common than the 'feature' meaning, and because it exhibited limited similarity to *attribute*[111]. *Quality* seemed the best candidate because it was synonymous to both senses – but this made it a good synonym, not necessarily a good genus word. To make a clearer distinction between the senses, it was better to avoid using the same genus word in both definitions.

Characteristic and *feature* seemed to be good genus words for Sense 1, as they were single-word synonyms of *attribute*, and were very frequent. Similarly, *quality* was a good genus word for Sense 2 as it was the only single-word synonym of *attribute* of the three synonyms (*characteristic and feature* need a modifier *positive*). But since *quality* was also synonymous to Sense 1 of *attribute*, another genus wes needed not only to make the definition clearer, but also to specify which meaning of *quality* is used in the definition. This was best achieved by using the phrase *positive characteristic*, which had been used in the original explanation. The phrase showed there was some similarity between Sense 1 and Sense 2, and at the same time points out the difference between the senses. The definitions at this point were:

1. *a characteristic or feature*
2. *a quality or positive characteristic*

The final step was to add differentiae using the information obtained by analysing pattern definitions and pattern elements (i.e. collocates) of each sense. The analysis showed that most frequently used collocates of *attribute* in Sense 1 were predominantly things (*product, object, brand*) or people (*person, individual, people*), whereas most frequently used collocates of *attribute* in Sense 2 were people (*scientifically literate person, females, personal*) or institutions (*hospital*, *country*, *supermarket*, *Banana Republic*). Incorporating this information produced the following final definitions:

1. *a characteristic or feature of a thing or person*
2. *a quality or positive characteristic of a person or institution*

In this case, the definition of Sense 1 remained much more faithful to the explanation provided during meaning analysis, than the definition of Sense 2. Many similarities between the definitions remain – but it is the (sometimes subtle) differences that will help the users distinguish between the definitions, and, ultimately, senses.

[111] Information on the meaning of synonyms was obtained from other dictionaries.

6.3.3.2.5.2 Defining the function of a word

Grammatical words cannot be defined using the standard types of definition because they do not normally have meaning. The function of these words is more important and should be described to the users, and examples of use provided. An example of such an entry is the adverb *therefore,* where the following definition has been written:

> *used to introduce the result or conclusion of something that has just been mentioned*

The meaning of *therefore* is indicated by providing cross-references to synonyms, such as *thus.* Because these function words are very frequent, the users are expected to be familiar with at least one of the synonyms.

A slightly different case is the entry *et al.*, where the description of the function needed to be preceded by the explanation of the meaning:

> **et al.** *means* and others *and is used to save space in academic writing when you are providing a reference for a book or article with three or more authors, but you do not want to name all the authors*

The explanation of the meaning is required because the headword is an abbreviation, and the abbreviation of a Latin phrase rather than an English one.

6.3.3.2.5.3 Defining technical terms or senses

Definitions of technical terms are often the result of a joint effort between an expert and a lexicographer, where the expert writes the initial definition and the lexicographer makes amendments to it. As a result, the definitions can be too demanding for the users and useful only to other experts in the field (Landau, 2001).

A different approach to the writing definitions of technical terms is proposed by Norman (2002) who suggests that scientific lexicographers (people with expertise in lexicography and certain technical fields), rather than experts, are much better equipped to produce definitions that users will understand. Supporting Norman's view are studies such as Pearson (1996) and Chung and Nation (2003) which show that definitional information can be often found in corpus texts, thus reducing the need for expert input.

Corpus-based defining was tested for the purposes of this Model. The approach is already used to some extent in the identification and explanation of compounds (see 6.3.2.2.1), the difference being that the focus at that stage was not on writing definitions.

Table 66. DOAE: Definitions - Pearson's patterns to find useful examples for writing definitions.

Pattern	Search
X is/are Y+ distinguishing characteristic	"attribute" [] {0,4} "is" "attributes" [] {0,4} "are"
X consist(s) of Y + distinguishing characteristic	"attribute" [] {0,4} "consists" "of" "attributes" [] {0,4} "consist" "of"
X is/are defined as Y + distinguishing characteristic	"attribute|attributes" [] {0,6} "is|are" "defined"
Y + distinguishing characteristic is/are called X	"is|are" "called" [] {0,6} "attribute|attributes"

Patterns identified by Pearson (1996:822-823) in her study (see Table 66 above) were used to conduct searches for concordance lines containing explanations of technical senses of sample entries. Pearson's patterns proved useful in some cases, for example when writing the definition of the Computing sense of the noun *authority* (sense 6). Pattern searches were limited to the Computer Science subcorpus, and identified the following example that was used to form the definition:

> ***Authorities*** *are pages that have many citations pointing to them, whereas hubs represent pages that have a lot of outgoing links.*
>
> (corpus file: Comp_9_2006_lauetal)

In other cases, such as in the Computing sense of the noun *attribute* (sense 3), Pearson's patterns did not identify any examples with definitional information[112]. Limited success of Pearson's patterns was not completely unexpected, given that Pearson's (1996) study showed that texts written by experts for their peers rarely contain definition statements.

However, during the analysis of sense 3 of *attribute* with Pearson's patterns, two other patterns were identified that were considered useful for writing the definition of sense 3 (Table 67). The first pattern consists of *attribute* followed by a word or phrase in brackets, which is often a synonym or a paraphrase. The other pattern is "*attribute* + *describe*", and tells us something about the function of an attribute.

[112] Concordance searches were limited to Computer Science, Engineering, and Mathematics as examples of Meaning pattern 3 (i.e. sense 3) of *attribute* came from these three subcorpora.

Table 67. DOAE: Definitions – Other patterns useful for writing the definition of sense 3 of *attribute* (noun).

Other patterns	Examples
attribute + *feature/item/variable/properties*	The data set contains 32711 instances (transactions) with 294 **attributes** (items); each attribute is an area of the www.microsoft.com web site. Attribute validity: the entity (ontology) has the **attribute** (properties), a measure of the number of properties in the ontology. Note that we could have used other clustering methods, such as semisupervised clustering methods based on partitioning data on a selected **attribute** (variable).
attribute + describe	Conditions in a clause often refer to attributes. **An attribute** describes a single or list-valued invariant (constant variable with constant value).

The definition of sense 3 was then written by using the information from examples. The analysis of examples under sense 3, especially examples of the two patterns shown in Table 67, indicated that *attributes* ***describe properties*** of an ***entity*** or a ***file*** (key words are provided in bold), and that each attribute had a certain value. Furthermore, the analysis of the Word Sketch relations of *ATTRIBUTE* (for the Computer Science domain only) showed many examples of *attributes* describing ***objects***. All this information was added to the explanation of the meaning ('a data item with a certain value') to create the definition:

3. *a data item with a certain value that describes a property of an object, entity, or file*

Corpora are therefore useful when writing definitions of technical terms or senses. It has been shown that the list of patterns offered by Pearson is by no means extensive – more patterns, perhaps specific to a word or group of words, can be discovered by looking at examples, or list of collocates.

6.3.3.2.5.4 Defining multi-word items

A multi-word item can be assigned sense status, which means that it requires a definition. One option is to embed the multi-word item within a full-sentence definition. However, the item may then not be prominent enough in the entry. It is more efficient if it is immediately signalled to the user that the entry contains multi-word items with sense status. This is achieved by offering the multi-word item first, followed by the description of usage, as in Sense 3 of *fact*.

3. the fact of the matter is / the fact is
used to introduce a statement in which you make an important point about what has just been said

Sometimes, even a multi-word item provided within the sense needs a definition, which acts as a quick explanation:

6. an actual occurrence; an actual event
after the fact *after the event that has just been mentioned*

Because the definitions of multi-word items often focus more on the function than the meaning of the phrase or construction, they rely heavily on examples to provide complementary information.

6.3.3.2.5.5 Definitions within examples

A form of definition is sometimes required within an example. Such a need arises when the example, which is typical of the word's usage in a particular sense, lacks certain contextual information that cannot be provided because it does not immediately precede or follow the example, or because it is provided in non-textual form (e.g. as a table). The explanation of the relevant context needs to be provided for the user, next to the relevant item.

as someone argues somewhere / as argued somewhere
No doubt state leaders in new states often follow the ethnic model of nation-building, but, as I argue below ***(=later in the article)****, this is not the only possible solution, and normally not the best one in order to survive and flourish as nations.*

As shown by this example of the construction containing the verb *argue,* the definition/explanation of *below* is inserted in the main example, but is at the same time differentiated from the example text by being offered in brackets and bold font[113].

6.3.3.3 Multi-word items

The group of multi-word items within the entry are phrases and idioms, and represent the phraseology of the headword. Phrases contain fixed and semi-fixed phrases, transparent collocations (see 6.3.3.5), formulae (similes, catch phrases, and proverbs), quotations, and support verb constructions. Phrases are treated in one of the following four ways: as a separate

[113] A similar approach is found in some learners' dictionaries and dictionaries for students, for example LED CD-ROM.

(sub)sense (see 6.3.3.2.5.4), as a part of the definition, as a construction with examples within the sense, or as a highlighted item in the example under the sense (see example for *assortment* below).

> **assortment** *(noun)*
> *The main retail chains did not want to incorporate new salads into their* ***product assortment****.*

Figure 72 shows extracts from the entry *argue*, showing how phrases are embedded in the definition (senses 2 to 5), and phrases offered as constructions (the multi-word items in bold under Sense 1, and Sense 6). The entry *argue* is also a good example of how prevalent phraseology can be in certain entries. The use of full-sentence definitions in such cases is not only highly appropriate, but also user-friendly, as phraseology is presented within the most frequently consulted part of the entry (i.e. definition).

Figure 72. DOAE: Multi-word items in the entry *argue* (verb).

argue *(verb)*
1 If someone **argues** a view or an idea in an article or book, they present the idea and support it with evidence. Note: **argue** is very often followed by a ***that***-clause.
article/paper/essay argues that...
it could/can/might/may be argued that...
one could/can/might/may argue that...
as someone argues somewhere / as argued somewhere
as argued by someone / as someone argues
2 If you **argue for** or you **argue in favour of** an idea or theory, you provide evidence that supports it.
3 If you **argue against** an idea or theory, you provide evidence that opposes it.
4 If you **argue with** someone about/over something, you discuss it because you have different opinions.
5 If you **argue with** someone or someone's view, you disagree with it.
6 if people **argue**, they talk angrily to each other because they disagree.
argue about/over something

The treatment of a phrase depends on its transparency, frequency and salience. For example, more transparent phrases (e.g. *obtain consent*) are normally shown within an entry, or even just offered in an example, if they are not that frequent.

Idioms differ from phrases in that their meaning is not a sum of their parts. Hence, idioms need to be accompanied by a definition which means they cannot be offered only as an example.

Other multi-word items such as compounds and phrasal verbs are given headword status, and feature in single-word dictionary entries in the form of cross-references (see 6.3.3.9.2).

There are cases when compounds and phrasal verbs need to be listed under single-word headwords, for example when they represent one of the most salient patterns of the headword.

Phrases and idioms often have variants. Once all the variants are identified, the canonical form(s) that will be used in the dictionary need to be selected. The decision on what represents the canonical form of the multi-word item, and which variants to show in the entry, should reflect the typical variant forms found in the corpus (Moon, 1996).

Phrases and idioms can also have one or more parts that can be used in different word-forms, or can have parts that are optional. For example, in the phrase *the fact of the matter is*, the verb *be* is also found, albeit rarely, in past tense (*was)*. Similarly, in the phrase *given the fact (that)*, *that* is sometimes omitted. But the aim is to present typical usage, so the most frequent phrase forms are offered in the dictionary.

This variability of parts of phrases and idioms was recorded in the database by using <fix> and <flex> tags, which indicate fixed and flexible part(s) of the phrase/idiom respectively (see example below). Tagging parts of multi-word items, while useful for lexicographers, can also be utilized to improve the successfulness of multi-word searches.

<fix>given the fact</fix> <flex>(that)</flex>

6.3.3.4 Examples[114]

Examples are an extremely important part of a dictionary entry, because they show how the word is actually used, thus returning the word to its natural environment in text after decontextualizing it by listing it in isolation. The main function of examples is to complement the definition; sometimes, the definition is hard to understand without reading the examples (Atkins & Rundell, 2008:454). Examples can also be of great help for navigating through longer entries, where the users can "identify the particular sense they are seeking by finding examples that are similar to the one they need or have in front of them" (Fox, 1987:137).

Examples are consulted frequently by students, so they must be given a prominent role in DOAE. Examples should convey as much syntactic and collocational information about the word as possible, as the findings of the survey indicate that students, especially NSs, are less likely to consult such information if it is provided separately. Frequent collocational patterns in examples are highlighted in bold, so their importance is more prominent.

[114] See 6.2.4 for discussion on how examples are saved in the database.

6.3.3.4.1 What makes a good dictionary example?

Most frequently mentioned characteristics of a good dictionary example are:

a) Typicality and naturalness. Examples should show typical behaviour of a word in terms of context, syntax, phraseology and collocation. In addition, every attempt should be made to maintain a consistent register, and to offer the right amount of context.
b) Informativeness. Examples must actually contribute something to the entry. In electronic dictionaries, there is a temptation to offer more examples, as there are, seemingly fewer space constraints. Yet, a computer screen is limited in size and more examples will mean more scrolling is needed to see the other parts of the entry. In addition, a lot of examples can act as a distraction to the dictionary users, who may not want or need to read great amounts of texts.
c) Full-sentence form. Examples should be offered as complete sentences, because examples in the form of a short phrase (authentic or invented), which are more typical of dictionaries for NSs, seem abstract and unnatural (Williams, 1996).
d) Appropriate length. Examples should not be too long so that readers are spared long readings, and can focus on the headword and immediate co-text of the example. On the other hand, examples should not be too short, as they may not provide enough context for encoding, or even decoding purposes.
e) Domain representativeness. Examples should reflect any domain preferences of a sense, subsense, sense pattern, or multi-word item. For example, if a sense is mainly found in Sciences, most examples should be taken from the Sciences subcorpora.

6.3.3.4.2 Selecting dictionary examples

Dictionary examples should be taken from the corpus, so they represent authentic language use. Dictionary examples should be selected from the examples already in the database, whenever possible, as the examples recorded during meaning analysis represent common syntactic patterns and collocates and have already undergone a significant amount of automatic and manual selection processes.

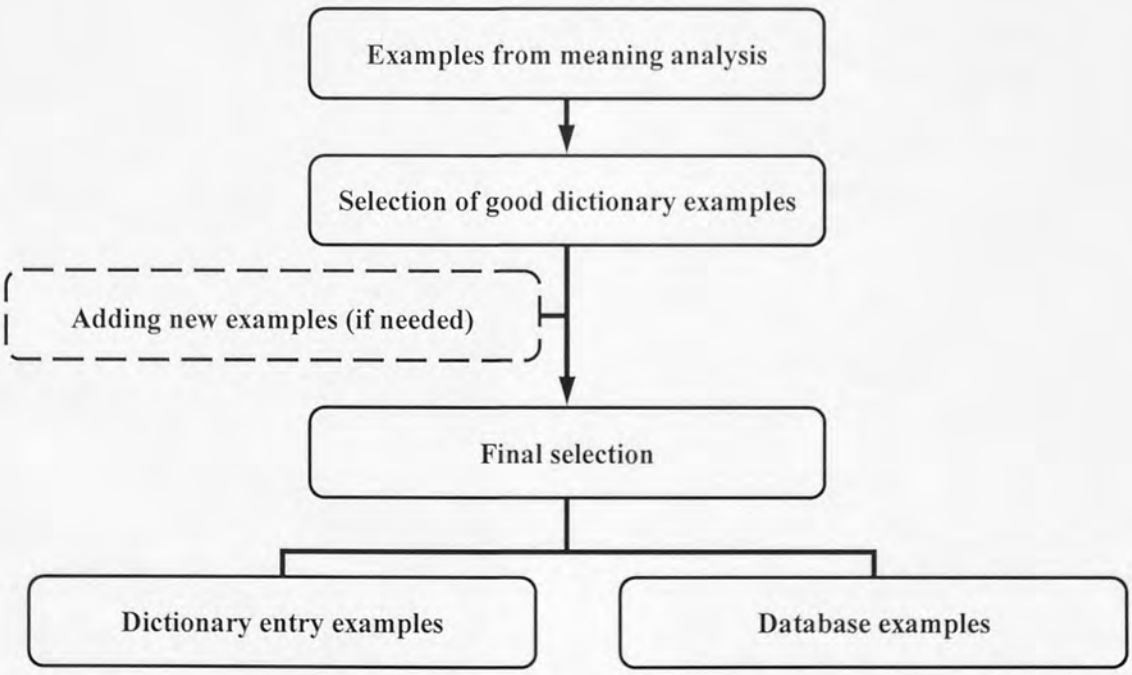
The creation of sample entries, however, has shown that often only some or none of the examples in the database can be considered to be 'good' dictionary examples. Consequently, new examples had to be searched in CAJA (with the specific entry context in mind), and

recorded in the dictionary entry. This proved as a very time-consuming task, not only because examples and relevant information had to be manually copied and pasted into the database, but also because good dictionary examples in academic English were difficult to find.

Examples from the meaning analysis and newly added examples that did not end up in the final entry were not discarded. These examples were saved under the element 'Database.examples'[115], so they can be used by lexicographers for reference, or offered as an additional feature in the dictionary.

The number of dictionary examples depends on the frequency/importance of the headword and/or sense, and the domain distribution of the headword. The number of multi-word items within a sense plays a role because each multi-word item needs to be accompanied by at least one example. More frequent senses, which are likely to be offered first, are often more heavily patterned and should contain more examples than the less frequent ones. Figure 73 shows the process of selection of dictionary examples. The criteria for good dictionary examples (6.3.3.4.1) are followed throughout the process. Domain representativeness must be considered when new corpus examples are added, and during the final selection.

Figure 73. DOAE: Selection of dictionary examples.



[115] The element Database.examples is available for senses, subsenses, and multi-word items.

6.3.3.4.2.1 Sample selection of examples – sense 3 of the verb *attribute*

Sense 3 of the verb *attribute* corresponds to Meaning pattern 3 in the database, whose contents were copied to the dictionary entry. The relevant items of information when selecting for the dictionary entry are pattern definitions, pattern elements, domain labels, and any notes.

Table 68. DOAE database: Relevant infomation for selection of examples for sense 3 of *attribute* (verb).

Meaning	If you **attribute** something, such as a statement or a work of art, **to** a person, you believe that person is its author.
Pattern definitions	[Human 1] **attribute** [Artifact | Event] **to** [Human 2 | Deity] [Artifact | Event] **be attributed to** [Human | Deity] [Human | Deity] **to which/whom** [Artifact | Event] **be attributed**
Pattern elements	Human 1 (rare and none is prevalent): *she, we, scholars* Artifact | Event: *saying, statement, argument, painting, text* Human (2) | Deity: *Michelangelo, Kant, supernatural beings*
Notes	The verb *attribute* is frequently used in the passive. The sense is very rare in Sciences domains - only one example was found in 500 concordance lines (250 Applied Sciences, 250 Life Sciences). The sense is also infrequent in Business Sciences, and Social Sciences (2% of concordance lines; sample: 250 concordance lines).
domain labels	Arts, Humanities

The relevant information for Sense 3 of *attribute* is offered in Table 68 above. An examination of the examples under Meaning pattern 3 showed that none of them contain the third pattern definition (Table 69 and Table 70 below). It was therefore decided to offer examples for the first two pattern definitions only.

Meaning pattern 3 had L2 domain labels 'Arts' and 'Humanities', and all the database examples came from other Arts and Humanities domains. The selection of examples for the dictionary entry began by discarding 10 examples that were considered unsuitable on account of length, complexity, or obscure language (Table 69).

Table 69. DOAE: Discarded database examples of Meaning 3 for Sense 3 of *attribute* (verb)

1	*And to paraphrase a statement attributed to Voltaire, I disapprove of a pastiche but I will defend to the death the right of the individual parts of the pastiche to operate autonomously.* (Art_26_2006_crespy)
2	*A legitimate claim that there is a gap between reality and thought requires knowledge of each, but the argument attributed to Nietzsche renders such knowledge not merely unavailable, but even unthinkable.* (Philos_49_2006_guay)
3	*The music was attributed to James Paisible in the reissue of c.1712.* (Art_24_2006_thorp)
4	*By 1765, when Schürer assembled the Catalogo, between fifty and sixty works attributed to Galuppi had arrived from Iseppo Baldan.* (Music_3_2006_stockigt+talbot)
5	*In addition, KNM-ER 1472 and 1481 are somewhat shorter than specimens with similar estimated mass attributed to H. erectus I used these specimens here to give some idea of how the lower-limb lengths and body masses of these specimens would have influenced their DEE.* (Archaeol_51_2006_steudel-numbers)
6	*The costs of defined episodes are attributed to responsible physicians, and each physician's score is calculated as a function of actual costs and expected costs for attributed episodes.* (Soc_12_2006_thomas)
7	*For it attributes to Kant a claim that does not fit well with his aesthetic theory as a whole.* (Art_47_2007_murray)
8	*To attribute to Moore the view that goodness is reason-providing on the basis of passages as the one just cited would be to jump too hastily to conclusions.* (Philos_84_2006_olson)
9	*Some contemporary scholars have attributed to Descartes the view that sensations of color and the like are qualia on the basis of similar arguments.* (Philos_88_2007_derosa)
10	*Veronique Munoz-Darde, for example, develops an argument, which she attributes to Bertrand Russell, that the family (with the substantial array of parental rights that it involves) is necessary for maintaining the background of diversity against which people can make a wide range of choices about how to live.* (Theol_117_2006_brighouse+swift)

At this point, one new example (Music) had to be added to exemplify the pattern variant '[Human 1] **attribute** [Artifact] **to** [Human 2]'. This made a total of 11 candidate examples for the sense.

In the final selection, three examples were selected for the sense, while eight were saved under database examples (Table 70 below). The three selected dictionary examples display many different syntactic patterns and collocates of *attribute* in sense 3:

- Each example comes from a different domain.
- Each example contains a different Artifact or Event.

- Two examples feature salient collocates identified by Word Sketch (*argument, Michelangelo, Kant).*
- In two examples, *attribute* is in the passive voice, and in one example *attribute* is in the active voice. This reflects the fact that *attribute* in Sense 3 is predominantly found in the passive voice.

Table 70. DOAE: Dictionary and database examples for sense 3 of *attribute* (verb).

Dictionary examples

1 Arts and Humanities: Humanities: History *This painting is very likely the large painting of St Francis listed in the 1613 inventory, attributed to Michelangelo.*

2 Arts and Humanities: Arts: Music *Emilio Bigi, Giuseppe Corsi and most modern literary scholars now attribute the text to Castellani.*

3 Arts and Humanities: Arts: Arts and Art History *The paper distinguishes between two different senses of 'genius' found in Kant's Critique of Judgement, and criticizes an argument commonly attributed to Kant.*

Database examples

1 Arts and Humanities: Humanities: Religion *It might be therefore argued that the saying attributed to Rabbi Yehuda in the Mekhilta Deuteronomy and in the Sifre is almost identical to the one attributed to Rabbi Yose in the Sifra.*

2 Arts and Humanities: Arts: Music *It is tempting tentatively to attribute the anonymous hymn settings, items 9 and 10, to Boluda, as did Robert Stevenson, and as indeed they may plausibly be.*

3 Arts and Humanities: Humanities: Philosophy *Molly Bloom says (attributing the remark to Leopold Bloom), "... the sun shines for you ..." Joel Feinberg's sun shone for all of us in philosophy.*

4 Arts and Humanities: Arts: Music *Violin no.3366 has no label, but is also attributed to Andrea Amati and dated before 1577.*

5 Arts and Humanities: Humanities: Archaeology *Yet the relationship between rock art forms and the behavioural contexts in which each form was produced remains unclear, with custodians attributing the origin of assemblages to supernatural beings, the Dreamtime Ancestors (e. g. Edwards, 1966, p. 36).*

6 Arts and Humanities: Business Sciences: Economics *This type of figure, attributed to Mussa (1974), is often used to analyse the effects of factor mobility, both between countries and (in specific-factors models) between sectors.*

7 Arts and Humanities: Humanities: Archaeology *Those objects most highly valued by museums and collectors, and those which have received a great deal of scholarly interest, are often attributed to a particular artist or school.*

8 Arts and Humanities: Social Sciences: Anthropology *The second part of the Vatican manuscript enumerates several miracles attributed to the saint in Crete, such as helping in the discovery of lost animals and objects, and in the curing of sick people and animals.*

6.3.3.4.2.2 Usefulness of GDEX when selecting examples

The GDEX function in Sketch Engine (see 3.3.1.2.1) was tested extensively when building sample entries. The findings showed that GDEX is of little help when selecting good dictionary examples of academic language. Table 71 offers comments on the most problematic aspects of GDEX heuristics that affect the quality of examples provided.

One of the main problems is that the short examples of academic writing preferred by GDEX often lack necessary context, making them unsuitable for dictionary entries. Another issue is that language of Arts and Humanities domains seems to be less complex, and is thus favoured by GDEX heuristics, so examples from Arts and Humanities are preferred over examples from Sciences.

Table 71. DOAE: Comments on some of the heuristics of GDEX.

GDEX heuristics	comments (based on selecting examples from CAJA)
preferred sentence length: 10-25 words (Note: punctuation is counted as words)	most sentences in academic writing are longer than 25 words[116]
sentences containing words which are not among the commonest 17,000 words are penalized, and additionally penalized for rare words	academic writing is full of terminology, so even sentences containing common words often contain technical terms and other infrequent words
sentences containing pronouns and anaphors like *this, that, it*, or *one* are penalized	these items, especially anaphors, are very frequent in academic writing – *that* is found in almost every 3rd third sentence, and *this* in every 7th sentence[117]
sentences with the target collocation in the main clause are preferred	many sentences have a complex structure, which is likely to cause more errors in the identification of the main clause

Currently, GDEX is not suitable for selecting dictionary examples of academic writing. This is not surprising, considering that GDEX was designed for, and first used on, learners' dictionaries, which focus on more frequent vocabulary. To make GDEX useful for DOAE, its heuristics would need to be adapted to the characteristics of academic language.

[116] According to WordSmith Tools 4 (Scott, 2007) (Sketch Engine does not have this function), the average sentence length in the CAJA corpus is 26 words. However, Sketch Engine's calculations include punctuation as tokens, so GDEX will actually prefer examples that will be shorter than 25 words.

[117] Frequency of *that* = 1,038,027; frequency of *this* = 493,758; number of sentences in CAJA = 3,338,781 (from WordSmith Tools 4).

6.3.3.4.3 Modifying examples

Dictionary practice shows that sometimes examples taken from a corpus need to be modified due to their length or complexity (Krishnamurthy, 1987; Landau, 2001; Atkins & Rundell, 2008). Similar findings can be reported for examples of academic writing. Different types of changes to the examples selected for the dictionary were made; the changes were mainly small but few were quite significant. In general, every attempt was made to avoid oversimplifying corpus examples, particularly for headwords which are normally surrounded by infrequent vocabulary and occur in complex sentences, because as Fox (1987) suggests, such action may significantly reduce the usefulness of the examples for encoding purposes.

Types of change made to examples are presented in Table 132 in Appendix 9. The most frequent type of change was omitting extra context, sometimes quite a considerable amount. The main aim was to remove any distracting or irrelevant text from the example, but still keeping its decoding and encoding value.

In rare cases, major revisions were made to examples. For instance, the example for *state-of-the-art* (below) was shortened because the noun *facility* was repeated too often for a non-headword.

Corpus example:
In addition to its launch pad, the space center will be equipped with state-of-the-art facilities such as storage and supply facilities for liquid propellants, an assembly complex, tracking and controlling facilities, and a ground test facility, etc.

Dictionary example:
The space center will be equipped with state-of-the-art facilities such as an assembly complex and a ground test facility.

6.3.3.5 Collocational information

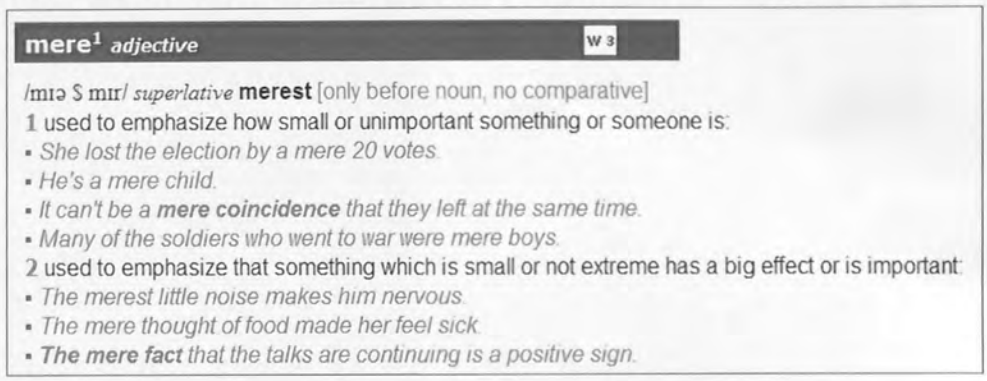
The notion of collocation forms an essential part of many features of an entry, such as definitions, examples, and multi-word units. Collocational information contributes significantly to the encoding value of a dictionary. But collocational information also helps lexicographers in compiling entries; for example, collocations can be used to make distinctions between senses (see 3.4, page 148)

From the user perspective, collocational information can be presented in the entry in three different ways:

a) Indirectly. Collocation is an inherent part of the entry feature (e.g. definition, example, or multi-word unit).

b) Semi-directly. Collocation is a central part of the entry feature, and the collocational pattern is highlighted (e.g. in bold). This approach is often used in examples, especially by most dictionaries for foreign learners (see Figure 74), and some dictionaries for NSs (e.g. NODE CD-ROM). Full-sentence definitions in COBUILD CD-ROM, but not in other dictionaries, also contain highlighted collocational patterns.

Figure 74. LED CD-ROM: Highlighted collocation patterns in the adjective entry *mere*.



c) Directly. Collocations are listed in a separate box at the end of the sense (e.g. MED Online - Figure 75), or at the end of the entry (e.g. e-OALD - Figure 76). One of the shortcomings of the MED approach is that, by listing collocates separately from the headword, it does not provide any information on the most typical position(s) of a collocate in relation to the headword. The method used by e-OALD is slightly more informative as it provides short patterns, but may still confuse the user by offering multiple collocates with multiple headwords.

Figure 75. e-MED: 'Collocations' box at sense 3 in the noun entry *result*.

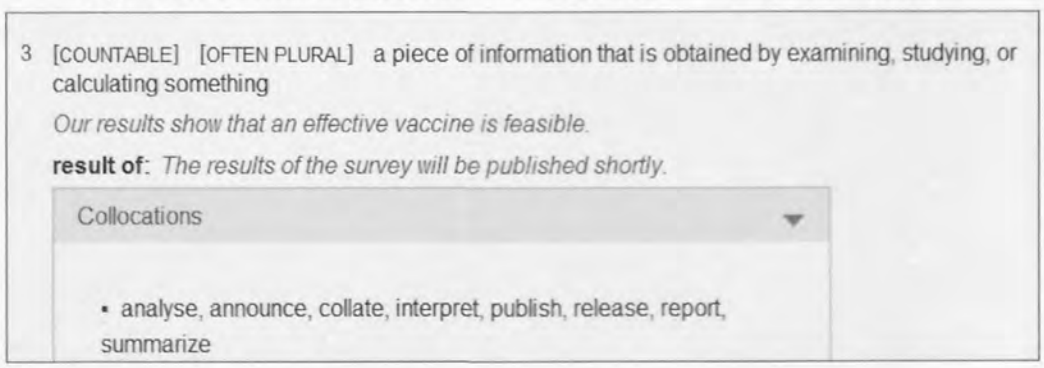
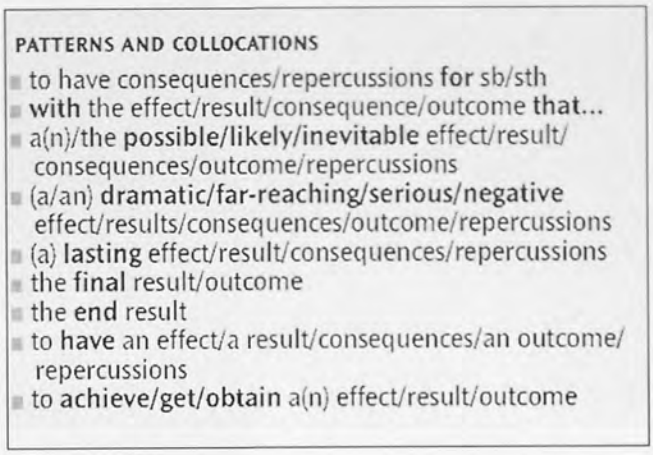


Figure 76. e-OALD: 'Pattern and Collocations' box at the end of the entry *effect*.



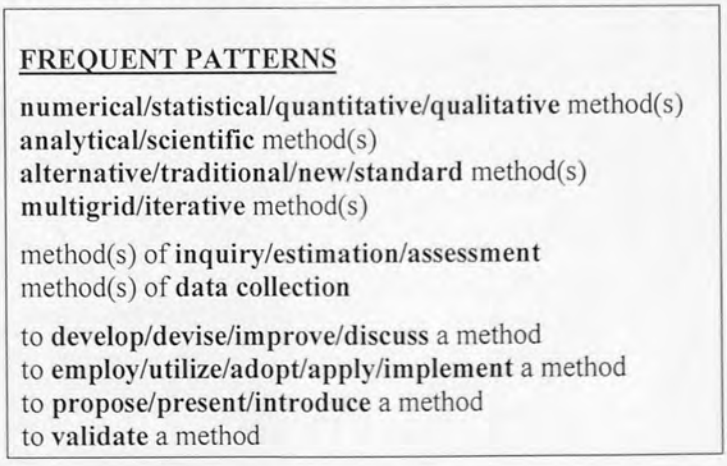
PATTERNS AND COLLOCATIONS

- to have consequences/repercussions **for** sb/sth
- **with** the effect/result/consequence/outcome **that...**
- a(n)/the **possible/likely/inevitable** effect/result/ consequences/outcome/repercussions
- (a/an) **dramatic/far-reaching/serious/negative** effect/results/consequences/outcome/repercussions
- (a) **lasting** effect/result/consequences/repercussions
- the **final** result/outcome
- the **end** result
- to **have** an effect/a result/consequences/an outcome/ repercussions
- to **achieve/get/obtain** a(n) effect/result/outcome

The indirect and semi-direct approach as to presenting collocational information are both used in DOAE, and have been discussed in earlier sections (6.3.3.2, 6.3.3.3, and 6.3.3.4). Conveying collocational information in the definition, examples, and multi-word items is a very user-friendly technique, considering that these parts of entry are consulted most frequently by students.

However, many headwords are so heavily patterned that the indirect approach and the semi-direct approach can provide coverage of only the most salient collocates. To avoid leaving out many collocates (and thus important encoding information for students), the direct approach sometimes needs to be used. The entry feature containing collocational information is called 'Frequent patterns' (Figure 77), and uses the e-OALD method of presentation, with one important difference; only the entry headword is focussed on (related headwords are not provided).

Figure 77. DOAE: 'Frequent patterns' in the entry *method*.



FREQUENT PATTERNS

numerical/statistical/quantitative/qualitative method(s)
analytical/scientific method(s)
alternative/traditional/new/standard method(s)
multigrid/iterative method(s)

method(s) of **inquiry/estimation/assessment**
method(s) of **data collection**

to **develop/devise/improve/discuss** a method
to **employ/utilize/adopt/apply/implement** a method
to **propose/present/introduce** a method
to **validate** a method

Many collocational patterns may be domain-specific and therefore not necessarily useful to all the students (Hyland & Tse, 2007; Durrant, 2009). It is therefore user-friendly to show the students only collocational patterns relevant to them. Such customizability of output is possible in an electronic dictionary, and necessitates recording information on the domain specificity of collocates (an integral part of the meaning analysis in this dictionary Model).

Only level 2 and level 1 domain labels are used when recording collocates in 'Frequent patterns' (see Figure 78); the label 'general' (for collocates that are not domain specific) is added to the existing set, making it nine labels altogether. Each label is represented by a DTD element, so that the output can be manipulated to exclude elements less relevant for particular type of user.

Figure 78. DOAE database: entry *method* - Frequent patterns with labels.

COLLOCATE (general): ***numerical***
COLLOCATE (general): ***statistical***
COLLOCATE (general): ***quantitative***
COLLOCATE (general): ***qualitative***
HEADWORD FOLLOWING: *method(s)*

COLLOCATE (general): ***analytical***
COLLOCATE (general): ***scientific***
HEADWORD FOLLOWING: *method(s)*

COLLOCATE (general): ***alternative***
COLLOCATE (general): ***traditional***
COLLOCATE (general): ***new***
COLLOCATE (general): ***standard***
HEADWORD FOLLOWING: *method(s)*

COLLOCATE (AppliedSciencesL2): ***multigrid***
COLLOCATE (AppliedSciencesL2): ***iterative***
HEADWORD FOLLOWING: *method(s)*

HEADWORD PRECEDING: *method(s) of*
COLLOCATE (general): ***inquiry***
COLLOCATE (general): ***estimation***
COLLOCATE (general): ***assessment***

HEADWORD PRECEDING: *method(s) of*
COLLOCATE (ArtsHumanitiesL1): ***data collection***

HEADWORD PRECEDING: *to*
COLLOCATE (general): ***develop***
COLLOCATE (general): ***devise***
COLLOCATE (general): ***improve***
COLLOCATE (general): ***discuss***
HEADWORD FOLLOWING: *a method*

HEADWORD PRECEDING: *to*
COLLOCATE (general): ***employ***
COLLOCATE (general): ***utilize***
COLLOCATE (ArtsHumanitiesL1): ***/adopt***
COLLOCATE (SciencesL1): ***/apply***
COLLOCATE (SciencesL1): ***implement***
HEADWORD FOLLOWING: *a method*

HEADWORD PRECEDING: *to*
COLLOCATE (SciencesL1): ***propose***
COLLOCATE (SciencesL1): ***present***
COLLOCATE (SciencesL1): ***introduce***
HEADWORD FOLLOWING: *a method*

HEADWORD PRECEDING: *to*
COLLOCATE (LifeSciencesL2): ***validate***
HEADWORD FOLLOWING: *a method*

6.3.3.6 Labels

Labels are used to alert the users to a particular vocabulary type. Labels are important for both language reception and production, as they can provide information on how the word is used, and when (in what context) can be used. Atkins and Rundell (2008:226), referring to

Ogden and Richards' 'meaning triangle' (1923; cited in Atkins and Rundell, 2008), remind us that it is important for lexicographers to know "that only an expression (word or phrase) can be labelled, not a concept (broadly, what you think of when you hear or use the expression) and certainly not a referent (a person in the real world)"[118].

In this Model, Atkins and Rundell's caveat is especially relevant for the use of domain labels. The decisions on the inclusion of domain labels in the entry rely heavily on the domain labels recorded during meaning analysis. It is important to make a distinction between domain labels and database domain labels: domain labels are used to label meaning, whereas database domain labels are used to label domain distribution of the lexical item.

Labels used in DOAE are discussed in 6.2.1 and 6.2.2. Most labels should be recorded during the meaning analysis (see e.g. 6.3.2.1.6 for discussion on domain labelling), and their inclusion is only confirmed or rejected when compiling the entry. Some labels, such as regional labels, may be added after consulting other dictionaries.

6.3.3.7 Assigning database domain labels to dictionary senses for customised sense ordering

Students using the dictionary will be studying different subjects, and will encounter or need to produce certain senses of the words more frequently than others. In existing dictionaries this need is addressed by providing domain labels, and by ordering senses by frequency/importance. The problem is that domain labels can be used to refer only to the meaning of the word or sense, and not to its use. Another problem is that even when domain labels are used, the sense order remains static, and since domain senses are often listed towards the end of the entry, the students may face a long navigation through the entry to find them, or even miss them completely.

A very user-friendly answer to the domain-related needs of students is to tailor the order of senses to each individual user (see 7.2.3 for more). This approach can be used in an electronic dictionary, and while the approach has been advocated by lexicographers such as Lew (2009), it has so far not been exploited by dictionary-makers.

In order to be able to manipulate the order of senses, senses need to be labelled in the database. The set of 51 DTD elements representing labels for senses is presented in Table 72. It is based on database domain labels, presented in 6.2.1 (Table 44), with two important additions. The element Main.sense was created to mark general senses, i.e. senses common to all or most

[118] Explanations in brackets were taken from Atkins and Rundell (2008:225).

domains. Secondly, as only one DTD element representing labels could be assigned to each sense, elements that represent combinations of L2 domain labels were created (e.g. Arts.BusinessSciences.L2.sense) to cover for instances when a sense was found in two L2 domains[119].

Table 72. DOAE database: DTD elements representing sense labels.

Main.sense	Architecture.L3.sense
	ArtsArtHistory.L3.sense
ArtsHumanities.L1.sense	Linguistics.L3.sense
Sciences.L1.sense	Music.L3.sense
	Archaeology.L3.sense
Arts.L2.sense	History.L3.sense
Humanities.L2.sense	Philosophy.L3.sense
BusinessSciences.L2.sense	TheologyReligion.L3.sense
SocialSciences.L2.sense	BusinessManagement.L3.sense
AppliedSciences.L2.sense	Economics.L3.sense
LifeSciences.L2.sense	Finance.L3.sense
	Law.L3.sense
Arts.Humanities.L2.sense	Anthropology.L3.sense
Arts.BusinessSciences.L2.sense	Education.L3.sense
Arts.SocialSciences.L2.sense	PoliticsInternationalRelations.L3.sense
Arts.AppliedSciences.L2.sense	Psychology.L3.sense
Arts.LifeSciences.L2.sense	Sociology.L3.sense
Humanities.BusinessSciences.L2.sense	ComputerScience.L3.sense
Humanities.SocialSciences.L2.sense	Engineering.L3.sense
Humanities.AppliedSciences.L2.sense	Mathematics.L3.sense
Humanities.LifeSciences.L2.sense	Physics.L3.sense
BusinessSciences.SocialSciences.L2.sense	Biochemistry.L3.sense
BusinessSciences.AppliedSciences.L2.sense	Biology.L3.sense
BusinessSciences.LifeSciences.L2.sense	Chemistry.L3.sense
SocialSciences.AppliedSciences.L2.sense	Geography.L3.sense
SocialSciences.LifeSciences.L2.sense	Medicine.L3.sense
	Sports.L3.sense
	VeterinaryScience.L3.sense

[119] Combinations of L1 labels were not created; L2 domain labels were used for senses limited to two or more different domains (e.g. if a sense is found only in Music and Law, label Arts.BusinessSciences.L2.sense is used). Also, L2 combination of Applied Sciences and Life Sciences is not used as these two categories combine into an existing L1 label Sciences.

Each sense is allocated a sense label, and the information on domain distribution recorded during meaning analysis serves as a basis for assigning sense labels. A sense label needs to be provided even if a sense already has a domain label, as sense labels are elements in the database, whereas domain labels are attributes.

It is essential that a copy of the entry is made before senses are saved under sense elements, in order to keep the record of the original sense order devised by the lexicographer. The original sense order in the entry is used as the default unless the option to tailor sense order by domain is utilized by the user. As soon as senses are put into sense elements, the original sense order is likely to be lost because senses with the same sense label are grouped together, and ordered according to the specified hierarchy.

6.3.3.8 Synonyms

Synonyms are one of the microstructural features most frequently consulted by students. Synonyms receive by far the greatest attention in thesauri, but there are two problems. First, none of the existing thesauri is very useful to students, as they do not focus on academic language. Second, the survey findings have shown that thesauri and similar resources are rarely used by students, which means that students expect to find the information on synonyms in dictionaries. Hence, DOAE needs to include information about synonyms.

The examination of the treatment of synonyms in the dictionaries frequently used by students has shown that the dictionaries use one or more of the following four approaches:

a) A synonym or a list of synonyms is used as a method of defining. This approach is mainly found in dictionaries for NSs. A variation is definition by antonym, which uses the pattern *not* + synonym, and is often combined with definition by synonym:

> **few** *adj.* 4b. not abundant; scarce
> **ill** *adj.* 1. not in good health; sick
> (CED CD-ROM)

b) A synonym is offered as an addition to the definition (Figure 79).

Figure 79. LED CD-ROM: Example of a synonym at the end of definition (the verb entry *show*).

12 ART/PICTURES [transitive] to put a group of paintings or other works of art in one place so that people can come and see them SYNONYM **exhibit**

c) Lists of (near)-synonyms are provided in a separate section, with sense numbers connecting them to the senses in the entry (Figure 80).

Figure 80. Dictionary.com: Lists of (near)-synonyms (the entry *fast*).

Synonyms:
1, 2. fleet, speedy. See QUICK. ***5.*** dissipated, dissolute, profligate, immoral; wild, prodigal. ***8.*** secure, tight, immovable, firm. ***9.*** inextricable. ***13.*** faithful, steadfast. ***14.*** enduring. ***20.*** securely, fixedly, tenaciously. ***22.*** recklessly, wildly, prodigally.

d) A list of (near)-synonyms is given, followed by an explanation of differences in usage (e.g. collocation, register, style), and examples. This feature has different names in different dictionaries, and is accessed in different manners. In LED CD-ROM, it is called 'Thesaurus box', and the user has to open it by clicking on "Thesaurus" on the right-hand menu, provided that the option is available. In MWCD CD-ROM, it is called 'Synonymy paragraph', and is offered at the end of the entry. In e-OALD, it is found as 'Synonyms' box at the end of the entry (see Figure 81). In e-MED, it is called Thesaurus and is accessed by clicking the icon T at the end of the sense.

Figure 81. e-OALD: 'Synonyms' box (the entry *fast*)

SYNONYMS

fast · quick · rapid

These adjectives are frequently used with the following nouns:

fast ~	quick ~	rapid ~
car	glance	change
train	look	growth
bowler	reply	increase
pace	decision	decline
lane	way	progress

- **Fast** is used especially to describe a person or thing that moves or is able to move at great speed.
- **Quick** is more often used to describe something that is done in a short time or without delay.
- **Rapid**, **swift** and **speedy** are more formal words.
- **Rapid** is most commonly used to describe the speed at which something changes. It is not used to describe the speed at which something moves or is done: ~~*a rapid train*~~ ◇ ~~*We had a rapid coffee.*~~
- **Swift** usually describes something that happens or is done quickly and immediately: *a swift decision* ◇ *The government took swift action.*
- **Speedy** has a similar meaning: *a speedy recovery.* It is used less often to talk about the speed at which something moves: ~~*a speedy car.*~~
- For the use of **fast** and **quick** as adverbs, see the usage note at **QUICK**.

The first three approaches share the problem that the differences between the synonyms are not explained. This is especially problematic with long lists of (near)-synonyms, where each additional (near)-synonym is often less synonymous with the headword. Also, defining by synonym is not a good option for students, as it assumes a great deal of linguistic knowledge without really providing the explanation of the meaning.

Approach d is the best of the four approaches because it focuses on the most relevant synonyms, and provides a great deal of information on synonyms to help the users understand the differences between them. Approach d can be combined with approach b to establish a quick link between the synonyms, and to provide a quick reference for the students who know the meaning they want, and simply need to be reminded of words they already know.

Figure 82. e-MED: Thesaurus box (linked with sense 1 in the adjectival entry *fast*).

There are some problems with existing dictionaries as regards the use of these two approaches. There are occasional inconsistencies in providing links to synonyms (e.g. in LED CD-ROM, *display* is offered as a synonym under the second sense of the verb entry *exhibit*, but *exhibit* is not offered as a synonym in the verb entry *display*). Synonym boxes are not provided systematically throughout the dictionary – for example, learner's dictionaries tend to provide

synonym information at the more frequent headwords only. Also, sometimes the treatment may be more confusing than helpful for the user; e-MED (Figure 82 above), for example, offers only short definitions for synonymous words, but the wording of the definition often includes the synonyms, making it very difficult for the user to distinguish between them. All these problems should be avoided in DOAE.

Adding synonyms to the definition, and creating Synonym boxes can only begin after all the dictionary entries have been compiled. At that point, common senses/meanings can be identified, and differences in the collocational preferences of (near-)synonyms pointed out.

Synonym boxes could not be created for the purposes of this thesis as only a limited number of sample entries have been built. Nevertheless, the process of how to identify synonyms, and, to some extent, synonymous patterns, could still be tested. The aim was also to also the Thesaurus and Sketch Difference functions within Sketch Engine (presented in 3.3.1.2.3).

The Thesaurus function is useful because it eliminates the role of intuition in suggesting candidate synonyms. Once the list of synonyms is produced, Sketch Difference can be used to compare each synonym with the headword. This approach was tested on the verb *attribute*, and Sketch Difference proved useful in identifying similarities and differences between the grammatical relations of synonyms (see Table 133 in Appendix 9). Having said that, the analysis required a considerable amount of manual inspection of concordances, something which would not have been necessary if the final entries for the synonyms had been available.

The Thesaurus function should not be used only for compiling synonym-related parts of an entry. The list of synonym candidates is very useful during meaning analysis for getting an idea of the meanings of the word, and/or to provide explanation of the meaning by using synonym(s). The use of Thesaurus can be combined with consulting the grammatical relation 'and_or' in the word sketch, which often contains synonyms and antonyms of the word (see also 6.3.2.1.5. and 6.3.3.2.5.1). Synonym-related information recorded during meaning analysis can then present a useful point of departure when the synonym-related parts of the entry are compiled.

6.3.3.9 Other parts of the entry

Dictionary entries also contain other features that assist the user in navigating through the entry, or provide additional information about the headword or related entries. These features are discussed in this section.

6.3.3.9.1 Menus

Menus are used to help the users to navigate through longer entries. Menus are offered at the beginning of the entry and contain "a brief set of mnemonics for the lexical units in the entry" (Atkins & Rundell, 2008:204). As this Model is for an online dictionary, menu mnemonics also contain hyperlinks to the relevant parts of the entry, which means the users can jump to the part of the entry they are interested in with a single mouse click.

The quickest way to create mnemonics in a menu is to use quick definitions as a point of departure, as was done for the noun *authority* (Table 73). If a sense has a label, especially a domain label, the label needs to be included in the menu, as it is an important element for identifying the relevant sense. Once the users find the sense they are looking for, they can either read the main definition, or, if they have already gained a sufficient understanding of the meaning, consult the examples and other information.

Table 73. DOAE: Menu for *authority* (noun).

MENU
1. the power to control people or activities
2. a government organization
3. **(the authorities)** organizations in charge of a country
4. expert
5. important written work
6. person with power
7. official permission
8. personal quality
9. *Computing* a type of internet page

When quick definitions are not available, the mnemonics need to be created from scratch, as was the case with the menu for *fact* (Table 74). Sometimes a mnemonic can use a part of the main definition, but in other instances it is more efficient to offer the frequent phrase of the sense, especially if the sense refers to the phrase and not only to the headword. For example, the mnemonic **in fact / as a matter of fact,** indicating the prevalent phrases of Sense 1, was created because the main definition, which explains the function rather than the meaning of the phrases, could not be shortened to meet the brevity demands of a mnemonic.

Table 74. DOAE: Menu for *fact* (noun).

MENU

1. **in fact / as a matter of fact**
2. **the fact that**
3. a piece of true information
4. an actual event
5. **the fact of the matter is / the fact is**
6. **a fact of life**
7. **stylized fact**
8. *Law* a criminal act
9. *Philosophy* a situation
10. *Computing* a clause in logic programming
11. *Biochemistry* **FACT**

Menus are also useful at headwords with more than one word class because they can help the user identify the word class as well the sense simultaneously. In order to help the user to identify the word class, the mnemonics of a sense need to reflect the word class of the sense. Such a menu was created for the entry *attribute* (Table 75).

Table 75. DOAE: Menu for *ATTRIBUTE*.

MENU

verb

1. assign a finding to something
2. assign a characteristic to something
3. assign authorship to someone
4. **attribute blame/responsibility**
5. assign object to a particular period

noun

1. a characteristic or feature
2. a positive characteristic
3. *Computing* a data item with a certain value
4. *Grammar* attributive adjective or noun
5. *Logic* property of a subject in proposition

6.3.3.9.2 Cross-references

Cross-references are used to inform the user that more information related to an entry can be found in other entries. Cross-references can be used in virtually any part of the entry. Cross-references in this electronic dictionary Model are also hyperlinks, so the user can go directly to the relevant entry or part of the entry by clicking on the cross-referenced item.

Although the full list of different types of cross-reference can only be provided once all the entries are compiled, several different types were identified when building the sample entries for this dictionary Model (Table 76):

a) Equivalence: cross-references to headwords or senses that have the same meaning(s) and similar, but not the same, word form.
b) Word form – lemma relation: the entry with the cross-reference is an irregular word-form of the referenced entry.
c) Synonymy or antonymy: cross-references to the headwords or senses with a similar or opposite meaning.
d) Related entries: cross-references to entries that contain the headword (may not necessarily be related in meaning, e.g. may be related in form, etymology, etc.).

Table 76. DOAE sample entries: Types of cross-reference.

Equivalence: *-entry to entry* *-sense to entry*	**ribonucleic acid** *(noun)* **see RNA** **potential** *(noun)* **2** the quantity determining the energy of charge in an electric field or of mass in a gravitational field ***Also*: electric potential**
Word form - lemma relation	**took** *(verb)* *the past tense of* **take**
Synonymy or antonymy	**therefore** *(adverb)* used to introduce a conclusion based on information that has been mentioned earlier **SYNONYM thus :1**
Related entries	**significant** *(adjective)* RELATED ENTRIES: **significant other**

There is another type of cross-reference which is different from the types of cross-reference mentioned so far, in that it refers to another part of the same entry. This type of cross-reference is represented by hyperlinks in menus (see 6.3.3.9.1) and 'Relation to Sense' cross-references (see 6.3.3.1).

Cross-references should be as specific as possible, i.e. they should immediately direct the user to the relevant part of the entry. For example, if a cross-reference is meant to refer to a

particular word sense, the sense number should be provided in addition to the headword[120] (see example of *thus* under *therefore* in Table 76).

Cross-referencing should not be overdone; if something can be explained in the entry, the user should not be forced to seek information in another entry. Cross-references can still be used as an addition, to establish the link between the entries (e.g. see *potential* in Table 76).

6.3.3.9.3 Illustrations

Although Ilson (1987:193) claims that "illustration is one of the five basic explanatory techniques in dictionaries", the role of illustration in a monolingual dictionary is mainly to assist the definition rather than replace it. Illustration is perhaps better regarded as a form of example, namely a graphical one, because it represents only one version of the item defined (Landau, 2001).

Illustrations are used mainly at concrete nouns, however they can also be effectively used at adjectives, prepositions, and verbs (Ilson, 1987; Svensén, 1993). Landau (2001:144), agreeing with Zgusta (1971), says that illustrations should primarily be used to depict unusual or unfamiliar things. The problem is that it is not that easy to establish what may be unusual or unfamiliar to the average user.

The value of illustrations for students should not be underestimated. Some students may be more visual learners, i.e. may prefer a visual representation of the item rather than the textual definition. In addition, an illustration may activate a memory of something seen before (Svensén, 1993). Illustrations are mainly used for decoding, but they can have an encoding value as well. For example, if students are interested in using a certain item, the illustration may help them confirm that the item they are planning to use is the correct one.

DOAE should therefore contain illustrations. Because academic language is full of abstract words, it is expected that illustrations will be used mainly for technical terms, specifically concrete nouns. Illustrations should be added once the entries are finalized. The online dictionary format suggested by the Model offers an additional potential, namely it can contain links to approved images on the Web instead of, or in addition to, the illustrations offered within the dictionary entries.

[120] Unfortunately, some existing dictionaries used by students fail to follow this practice.

6.3.3.9.4 Usage notes

Usage notes include any additional information about the use of the headword that may be useful for the students. Usage notes should provide any information related to the use of the word in academic writing and/or speech, or information about relevant academic conventions such as referencing (see Figure 83 for an example), or point out any important differences between the use of the word in academic English and general English.

Figure 83. DOAE: Usage note at the entry *et al.*

> **USAGE NOTE:** The abbreviation **et al.** is not usually used in the section 'References' or 'Bibliography' at the end of an article or a book as all authors need to be named in full.

Usage notes are one of a parts of the dictionary entry where lexicographers need to be especially careful to avoid being prescriptive. Hanks' (2005:262) recommendation to advise and warn rather than prescribe is appropriate here. One exception may be the notes on academic conventions (e.g. Figure 83) which may need to be prescriptive.

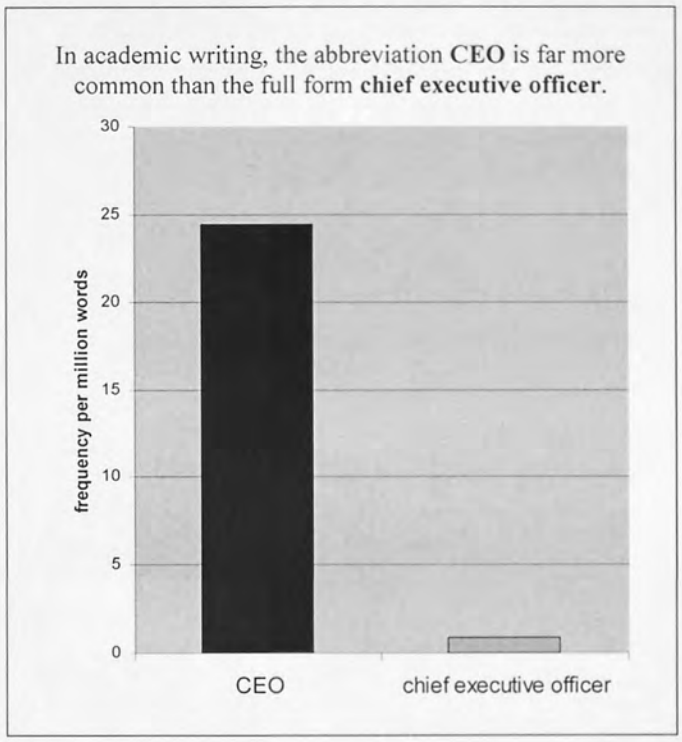
The main survey showed that students, especially NSs, rarely consult usage notes. Every attempt should therefore be made to present usage information within the entry through definitions and examples, i.e. the more frequently consulted parts of the entry. But because usage notes are a separate feature, and can thus be given a prominent position in the entry, they can sometimes be a better solution. Usage notes can also be supported with more visual information, such as frequency graphs.

6.3.3.9.5 Frequency graphs

Frequency graphs are used to compare the relative frequency of entries, multi-word patterns, synonyms, etc. A frequency graph can replace lengthy, discursive comparisons. One of the drawbacks of using frequency graphs is that the user can sometimes be put off using an appropriate item due to its low frequency. In addition, frequency graphs can sometimes distract the user from more important information, for example explanations of the subtle differences between synonyms.

Despite some shortcomings, frequency graphs should be used in DOAE, mainly as graphic support to the explanations. The frequency graph in Figure 84 can also act as a substitute for a usage note about the frequency of two near-synonymous entries (e.g. acronym and full form).

Figure 84. DOAE: Frequency graph at the entry *CEO*.

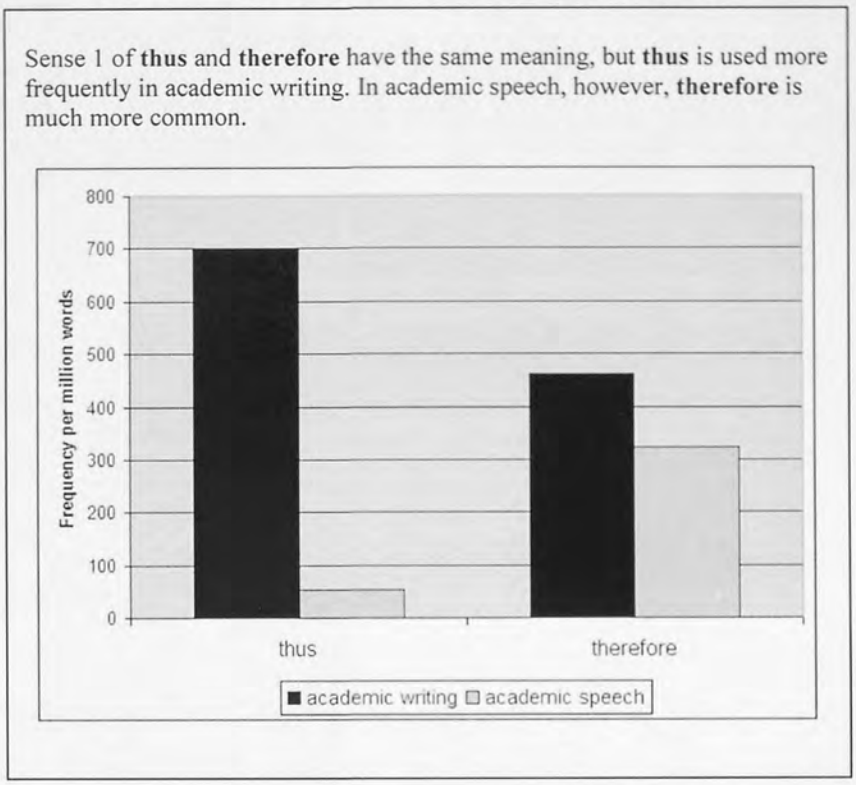


Frequency graphs can contain more than one dimension of frequency information. Especially valuable are comparisons between a word's use in academic writing and in academic speech. For example, the graph in Figure 85, showing a frequency comparison of frequency of two synonyms[121], indicates two important pieces of information; it shows that *thus* is used much more frequently in academic writing than in academic speech, and that *thus* is more frequent in academic writing, but *therefore* is more frequent in academic speech.

Although frequency graphs greatly reduce the amount of textual information needed, they may still need to be accompanied by a comment (i.e. usage note) so that the users know what information is being presented, and why (as in Figure 85).

[121] Frequency information for academic writing was obtained from the CAJA corpus, and frequency information for academic speech was obtained from the BASE corpus.

Figure 85. DOAE: Frequency graph at the entry *thus*.



6.3.3.10 Finalizing dictionary entries - consulting other dictionaries and corpora

Until now, the existing dictionaries that students use have been mainly used in this thesis as a source of ideas on how to present different parts of the entry, i.e. to obtain an understanding of good and bad dictionary practices. In order to avoid any influence on the building of entries, the actual contents (senses, definitions, multi-word items, etc.) were not consulted. Corpora of academic language (other than CAJA) played a greater role in shaping the entry contents, as they were consulted when developing L2 and L1 domain labels, and when creating frequency graphs.

However, existing dictionaries can be used in the final stages to identify any additional meanings and patterns (syntactic patterns, phrases, collocates, etc.) from academic speech or general language that may have been missed during the analysis[122]. The existence of these meanings and patterns needs to be confirmed by consulting corpora.

Table 136 in Appendix 9 lists the meanings and patterns that were absent from the sample entries. These missed meanings and patterns can be divided into general and technical

[122] Only meanings, phrases, collocations, etc. mentioned by at least two different dictionaries were considered.

ones. While missed general meanings were normally found in all the dictionaries that were consulted, missed general patterns were often found only in learners' dictionaries. On the other hand, missed technical meanings and patterns were often found only in NS dictionaries. These findings reflect the differences in coverage and target audience of the dictionaries.

Some meanings and patterns had to be added to the entries because the corpus evidence suggested that they are found in academic language. It was discovered that many missed general meanings and patterns were often already recorded in the database, but only in the form of collocates under grammatical relations. For example, the meaning 'a part of your face' of the noun *feature* was missed because the concordance lines indicating this meaning contained the phrase *facial feature(s)* where *feature* fitted the definition of sense 1 ('an important characteristic or part of something').

Meanings and patterns were also missed because they were very infrequent in CAJA. This is true not only of all technical meanings, but also of some general meanings and patterns, especially the ones that are found in a limited number of domains. An example of this are the two missed senses of the noun *feature* ('newspaper or TV report', and 'a film of standard length') which are not very frequent in academic language, and are found mainly in Arts and Humanities.

Most meanings and patterns that were added to sample entries were given the same status as they had in the existing dictionaries consulted. The exceptions are certain transparent patterns from learners' dictionaries that were provided by a definition, and sometimes also by an example. Considering that users of DOAE are expected to be more proficient in English than advanced learners, they do not require an explanation of such patterns. Nonetheless, if the pattern represents a frequent collocation, it still needs to be pointed out, normally as a construction under the relevant sense. This solution was used with the pattern *the fact remains*.

As a rule, no action was taken for the meanings and patterns, especially general ones, that had no examples in the academic corpora. If a few examples of the meaning or pattern were found, but the decision was made not to include the meaning or pattern in the dictionary entry, a new meaning pattern or meaning pattern definition was still created in the database and examples saved.

Low frequency was however less of a factor when considering the inclusion of some missed (semi-)technical meanings[123]. The Printing meaning of *justify* ('to adjust the spaces

[123] My personal knowledge and experience had to play a significant role in making these decisions, pointing to the need for the involvement of technical experts in the process of dictionary design and final editing.

between the words') was very rare but was included as it was believed that it will be useful to students because *justify* often features in the instructions on the format of essays, and in editor software programs on most computers. Similarly, the Grammar meaning of the noun *attribute* was rare but was included because it is a commonly used term in the area of Linguistics.

A special case is the phrase *in actual fact* which was included by replacing another pattern, namely the phrase variant *in point of fact* under sense 1 of fact. This decision was made after it was established that *in point of fact* is the least frequent variant, and is found only in CAJA (Table 77). It was also decided not to combine the phrases *in fact* and *in actual fact* into *in (actual) fact*, as dictionaries such as e-MED, LED CD-ROM, e-CALD and e-LDOCE do, but to offer them separately to give some indication of the difference in frequency.

Table 77. Corpus frequencies (per million words) of synonymous phrases under sense 1 of *fact*.

	CAJA	BAWE	BASE	MICASE
in actual fact	0.4	2.2	10.4	0.6
in fact	148	144	463	271
as a matter of fact	2.8	1.6	0.8	14
in point of fact	0.3	0	0	0

In addition to providing information on missed meanings and patterns, existing dictionaries can also be helpful in identifying compound entries. This can be straightforward when using the electronic version of the dictionary (as in this case) because the search for a single-word headword produces a list of entries (single-word and multi-word entries) containing the headword. Corpora again help us decide whether the compound is a candidate headword or not (a sample analysis of two entries is offered in Table 78 below).

Some of the compound entries identified by this process (e.g. *potential difference*) were flagged by Word Sketch during the analysis, but were considered to be a cases of transparent collocation and thus offered as constructions or merely featured in examples. *Local authority*, for example, was initially only a construction under sense 2 (albeit a prominent one), but was given headword status once other dictionaries were consulted. The relation between the two entries was kept by providing a 'Related entry' cross-reference to *local authority* under sense 2 of *authority*.

Significant changes to the sample entry *chief executive officer* were needed after a related entry *chief executive* was found in the dictionaries. *Chief executive officer* originally contained only a cross-reference to *CEO*, a much more frequent headword, but because it also

represents a longer form of one of the two senses of *chief executive*, a definition and a frequency graph were added to avoid sending the user from *chief executive* via *chief executive officer* to *CEO* with multiple cross-references.

Table 78. DOAE: Compound entries of *method* and *potential* (based on entries in other dictionaries and corpus evidence).

	compound entry	corpus evidence	action
method	*barrier method*	19 examples in CAJA 2 examples in BAWE	recorded as candidate headword
	direct method	found in CAJA, BAWE, and BASE but not in the technical meaning given in dictionaries (from 1 text)	NONE
	critical-path method	no examples found	NONE
	rhythm method	1 example in CAJA	NONE
	Suzuki method	no examples found	NONE
potential	*potential difference*	55 examples in CAJA 47 examples in BAWE	recorded as candidate headword
	potential energy	429 examples in CAJA 77 examples in BAWE 2 examples in BASE 1 example in MICASE	recorded as candidate headword
	potential divider	4 examples in BAWE	NONE
	potential well	66 examples in CAJA 7 examples in BAWE	recorded as candidate headword
	electric potential	69 examples in CAJA 1 example in BAWE	already a headword
	action potential	328 examples in CAJA 25 examples in BAWE 3 examples in MICASE	already identified as candidate headword during the analysis

Other modifications made to sample entries after consulting existing dictionaries included giving more prominence to certain frequent patterns that were initially only exemplified. This was required with the phrases *realize the full potential of* (sense 1 of the noun potential), *authority figure* (sense 1 of the noun *authority*), and *obtain through* (sense 1 of the verb *obtain*). The first two phrases were put in bold in the examples, whereas the third phrase was turned into a construction '*something is obtained through something*'.

6.4 Making an American English version of DOAE

As mentioned in 5.1.4.4 (page 180), British English has been selected as the main spelling for this dictionary Model. Nonetheless, American English variant spellings have been discussed, and sometimes recorded, in several parts of the database, for example for headwords, collocates, and multi-word items. In addition, phonemic spelling was recorded for database from an American dictionary. Some of the information needed for the American English version of the Model is therefore already in the database.

Nonetheless, the information recorded for the British English version needs to be modified to make it suitable for the American English version. The modifications are required at both macrostructural and microstructural level, and include:

- Changing treatment of headwords. Headwords with American English spellings become fully treated entries, whereas their British English variants are cross-referenced to them.
- Changing preference of displaying information. American English variants become the default setting (e.g. American English pronunciation, American English inflected forms).
- Changing regional labels within entries. *American English* label for variant spellings of different types of entry information (e.g. inflected forms, constructions) is omitted. *British English* label is introduced for British English variants.
- Changing British English spelling to American English spelling in definitions, usage notes, etc.

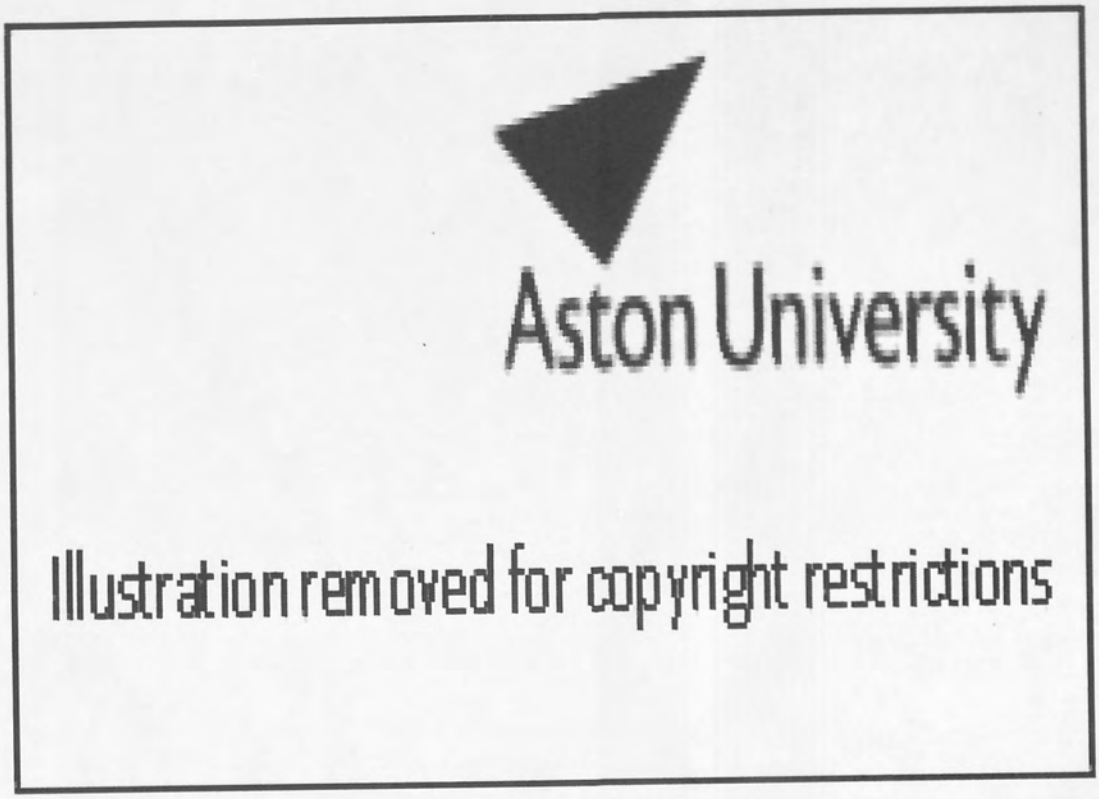
While some of these modifications are relatively easy to make, others require additional lexicographic work. Additional analysis is for instance needed for senses, collocates, multi-word items, and examples of the headword, to establish whether they are equally part of the headword's use in American English. It may therefore occur that parts of the entry need to be omitted, or given a less prominent role in the American English version of DOAE.

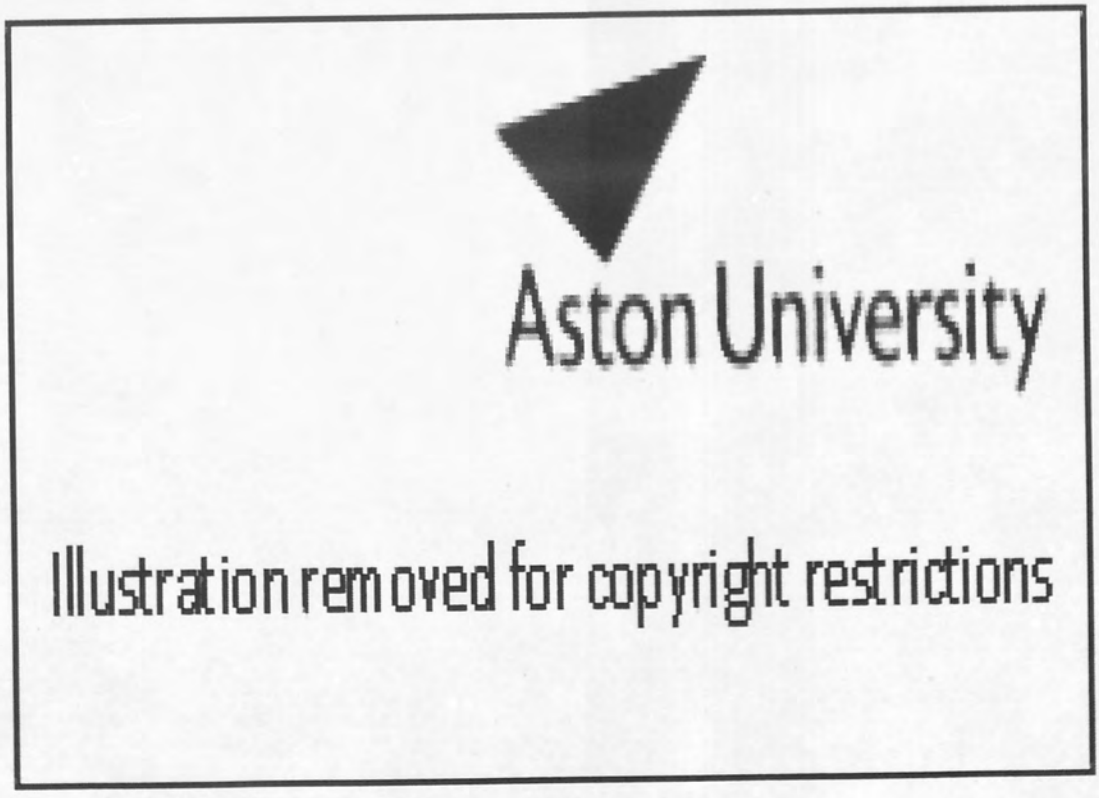
Two approaches to recording the information for the American English version in the database have been considered. The first approach (Table 79), involves recording under the same element the information for both variant spellings. Thus, the information for different spelling versions is contained in the same entry. The benefit of this approach is that the information that is shared by both variant spellings does not have to be duplicated. However, certain problems with the approach have been identified, especially when the British English entry and the American English entry are significantly different in contents (amount of information, order of the information, etc) – for example, compare the British English version

(Figure 86 - left) and the American English version of the entry (Figure 86 - right) for the noun *lift* in e-MED. In addition, the approach is rather impractical for lexicographers, who may be overwhelmed by too much duplicated information on the screen, as the distinctions between the two varieties would only be made clear in the output.

Table 79. DOAE: The first approach to recording information of different variant spellings.

HEADWORD
frequency rank frequency word class grammar label pronunciation: **British English** American English inflected forms: **British English (with label *British English*, to be used when British English is the default setting)** American English (with label *American English*, to be used when British English is the default setting) menu: **British English version** American English version Sense 1: definition: **British English version** American English version example(s): **British English version** American English version construction(s): **British English version** American English version usage note(s): **British English version** American English version etc. Sense 2: definition: **British English version** American English version example(s): **British English version** American English version construction(s): **British English version** American English version usage note(s): **British English version** American English version etc. Sense 3: etc.


Aston University
Illustration removed for copyright restrictions


Aston University
Illustration removed for copyright restrictions

The second approach (Table 80), considered to be more efficient and clear, involves making a copy of entire entries (but only the ones that need modifications) and modifying the contents to reflect the American English spelling. There are still some similarities with the first approach, as it needs to be clearly defined which information should be used for the British English version, and which for the American English version. Nonetheless, in this approach, this distinction is made by labelling the entire entries with the appropriate label (so only one label would be required), as opposed to the first approach where each item of variety-specific information needs to be labelled.

Table 80. DOAE: The second approach to recording information of different variant spellings

HEADWORD – British English version	**HEADWORD – American English version**
frequency rank frequency word class grammar label pronunciation inflected forms menu Sense 1: definition example(s) construction(s) usage note(s) Sense 2: definition example(s) construction(s) usage note(s) Sense 3 etc.	frequency rank frequency word class grammar label pronunciation inflected forms menu Sense 1: definition example(s) construction(s) usage note(s) Sense 2: definition example(s) construction(s) usage note(s) Sense 3 etc.

At this stage, all the information about the entries has been recorded into the database. This information can be split into two types. First, there is information that is relevant only to lexicographers, such as the meaning analysis section of the entry. The other information is information that has been created for the dictionary, i.e. has been created with the dictionary users in mind (all sample DOAE entries, using the default settings described in 7.3, can be found in Appendix 12). The latter type of information still needs to be selected and manipulated to create various versions of DOAE for different types of target users. This is discussed in the next chapter.

7. DOAE: DICTIONARY OUTPUT

Whereas the previous chapter looked at which information is recorded, and how it is recorded, this chapter focuses on how (and which) information is presented to users, namely in terms of DOAE output. The needs of the dictionary users, i.e. students, which have been central in shaping the dictionary information, are considered again. Finally, how the proposed dictionary Model would cater for different types of student is discussed.

The database dictionary entries contain the full range of information that has been assembled about the headword. However, not all information is relevant to the students, so the relevant parts of the information now needs to be selected to compile the dictionary entry. The full list of the different types of information available is as follows:

word class
grammar labels
pronunciation (IPA, phonemic, nonphonemic)
inflected forms
variant forms
variant spellings
menus
domain/style/usage/frequency/etc. labels
quick definitions
main definitions
examples
constructions
synonyms
frequent patterns (collocations)
usage notes
cross-references
frequency graphs
etymology

Any specific dictionary will differ in the output of database information. In TshwaneLex, such outputs are created with the function called 'style sets'; each style set contains the settings that control information that will be displayed, in which order, and what style/formatting should be used for each piece of information.

The main advantage of an online dictionary is that it can be dynamic – it can customise its contents according to the type of user that accesses the dictionary. It is tempting to provide the users with all the information available, and then let them customise the output (types of information, styles, formatting), but they usually do not want to be bothered with settings, especially if they are not regular users. This rationale is reflected in existing online dictionaries, where most of the decisions are made by the publisher, and only a minimal manipulation of contents by the user is possible. Some further limited manipulation is possible only by changing browser settings (e.g. increasing the font).

The problem of pre-determining all the settings is that the same dictionary then has to cater for all the different types of users. If the dictionary is to be maximally useful to each of its users, the users need to be able to change the settings themselves. The need for user involvement will always be required with settings that rely on personal preferences (e.g. colours). There are however, dictionary settings that can be informed by the external characteristics of users, for example their native language, which are obtained when creating the user profile. Different users will possess a different combination of these characteristics, so the aim is to design a different dictionary output (style set) for each type of user.

7.1 Style and formatting settings

Dictionary entries will provide the students with many different types of information, and a crucial part of ensuring the user-friendliness of the information is presentation. Different styles (e.g. font(s), font sizes, colours) and formatting (e.g. positioning of text) should be used to distinguish between, and give prominence to, different parts of entry.

There are many different fonts in existence, so the options seem endless. Nonetheless, an average computer user is likely to have only a limited set of fonts installed; Table 81 and Table 82 below show the 15 most common fonts on Windows and Mac systems respectively. Thus, the font(s) for the dictionary should be selected from these two sets, and should preferably include fonts which appear in both tables.

Font size is also very important. The font should not be too small as it may be difficult to read. Also, it should be borne in mind that computer screens require a bigger font size than paper because they have lower resolution, i.e. are less readable (Anderson, 2007). The font sizes of different fonts do not correspond – for example, all the fonts in Table 81 have font size 11 but exhibit significant differences in sizes (e.g. compare Times New Roman and Verdana).

Table 81. The most common fonts on Windows systems to 9 January 2010.

Font name	Installed (%)
Microsoft Sans Serif	99.68
Verdana	99.40
Courier New	99.30
Tahoma	99.30
Arial	99.15
Trebuchet MS	99.00
Comic Sans MS	98.95
Lucida Console	98.85
Impact	98.75
Times New Roman	98.60
Arial Black	98.55
Georgia	98.55
Lucida Sans Unicode	98.25
Palatino Linotype	98.04
Franklin Gothic Medium	97.89

Source: http://www.codestyle.org/css/font-family/sampler-WindowsResults.shtml.

Table 82. The most common fonts on Mac systems to 9 January 2010.

Font name	Installed (%)
Helvetica	99.71
Lucida Grande	99.13
Monaco	99.13
Geneva	98.84
Courier	98.55
Arial	97.10
Verdana	96.81
Times	96.23
Helvetica Neue	94.74
Georgia	94.20
Trebuchet MS	94.20
Arial Black	93.62
Times New Roman	93.62
Gill Sans	91.52
Impact	91.30

Source: http://www.codestyle.org/css/font-family/sampler-MacResults.shtml.

Font style is another feature that can be used to differentiate between entry features, or highlight them. TshwaneLex offers eight styles: normal, bold, italics, underline, superscript, subscript, small caps, and all caps. Bold text is particularly useful for highlighting individual parts of the text, such as the headword in definitions, or collocational patterns. Italics are often reserved for examples and certain grammar information.

Colours can be used in combination with, or instead of, different fonts, font sizes and styles. Modern dictionaries, especially learner dictionaries, provide substantial evidence of this practice. Blue is used very frequently, probably because studies show that blue text is easier to read for people with poor eyesight[124]. Red is also found in many dictionaries, although it features prominently only in e-MED. Nonetheless, black still dominates the dictionary text in the majority of online dictionaries.

In addition to carefully selecting fonts, font sizes, font styles, and colours, it should not be forgotten that not all users may not like the default settings, or may find them difficult to use

[124] This information was obtained from the editorial staff at HarperCollins during the professional attachment in 2005.

(e.g. because of special needs). With this in mind, four different options of styles and formatting have been designed for DOAE: normal (default), black & white, medium-size, and large-size[125]. The four settings are exemplified by the entry *justify* in Figure 102, Figure 103, Figure 104, and Figure 105 in Appendix 10.

7.1.1 Default style setting

The default style setting (see Figure 102 and Table 137 in Appendix 10) uses five fonts which are common to both Windows and Mac systems[126]. All but one font comes from the sans-serif family, which are easier to read on a computer screen. Arial and Verdana are used for the most important microstructural features (e.g. headwords, definitions, examples). The other three fonts are used to distinguish certain features or parts of the entry; for example, Georgia is used for menus (a change in font preferred to a change in font style or colour).

Most of the entry is in font size 10 or 11. Font size 10 is used mainly for the text in Arial, and font size 11 for the text in Verdana. Arial 10 and Verdana 11 are very similar in size (Table 83 below), and give a uniform look to the entry. Font size 12 is used only for headword variants, sub-entry lemma signs, and pronunciation, while font size 13 is reserved for headwords.

Table 83. Comparison of Arial 10 (top sentence) and Verdana 10 (bottom sentence).

This is a comparison of the two fonts. This is a comparison of the two fonts.

Four font styles are used: normal, italics, bold, and small capitals. Normal style is used for the longer sequences of text, such as definitions and usage notes, and some shorter features such as pronunciation. The only longer pieces of text that do not use normal font style are examples, which are offered in italics. Italics are also used for few other features (e.g. word class, regional labels), but mainly in combination with bold face. Bold face is used primarily for highlighting; it is used for headwords (as a headword sign or when used in the definition), sub-entry lemma signs, and phrases. Small capitals are used only for internal hyperlinked cross-references, and they are used in combination with bold face.

The colour palette consists of black, blue (various shades), red (various shades), green, purple, yellow, and white. Black is used for longer chunks of text (e.g. definition, examples)

[125] All these settings have been devised using a screen resolution of 1400 by 1050 pixels. Other resolutions would require different settings in order to maintain the same layout.
[126] Comic Sans MS does not feature in Table 82 because it is the 22nd most common font on Mac system, and the table shows only the 15 most common fonts.

and visually dominates an entry. Blue is used to highlight headwords, word classes, and labels. Blue is also used along with red, green, and purple for various types of label. Another important function of blue, which is shared with yellow, is that of a background colour. Especially important is the use of a blue background for definitions, which not only highlights the definition but also marks the beginning of every sense. White also has the default function of a background colour, but unlike blue and yellow, it does not highlight any part of the entry. White is nonetheless very important as the right amount of white space between various parts of the entry vastly improves the presentation of information.

Other helpful formatting functions employed were indentation (e.g. different indentation settings for definitions, examples, and constructions), boxes (e.g. for usage notes, word forms, etymology), and bullet points (for examples).

7.1.2 Black & white style setting

The black & white setting (Figure 103 in Appendix 10) is an alternative setting, designed for users who have impaired colour vision, or for users who prefer a more traditional dictionary appearance. The black & white setting retains the majority of fonts, font sizes, and font styles of the default setting (Table 138 in Appendix 10), but has a more limited colour palette (black, grey and white). Blue, red and green have been replaced by black or grey, whereas the blue background is replaced by light grey.

The change of font colour to black necessitated a change to font, font size, and/or font style for certain microstructural features, as the colour was the only characteristic that distinguished them from other parts of the entry. An example of this are domain and subdomain labels, and sense patterns (phrases given the status of sense) – they are all offered in Arial 10 bold in the default setting, but labels are provided in red and sense patterns in dark blue. For the black & white setting, the fonts of both types of feature have been changed (to Georgia for labels, to Arial Unicode MS for sense patterns). Font size has been changed only for sense patterns (to 12pt), and font style only for labels (bold to italics). Sense 8 of *FACT* below can be used to demonstrate the difference between the output using the setting in black & white settings (upper line), and the output that would have been obtained if only the colours had been changed to black (lower line):

8. after/before the fact *Law* after/before a criminal act
8. **after/before the fact Law** after/before a criminal act

7.1.3 Medium-size and large-size style setting

The medium-size setting (Figure 104 in Appendix 10) and large-size setting (Figure 105) offer font sizes larger that the ones used for the default setting. Font sizes in the medium-size settings have been increased by 1pt, and font sizes in the large-size settings by 2pt, compared to the default setting. Fonts, font styles, and colours of the default setting have been retained. The medium size setting and large-size setting are primarily intended for users with poor eye-sight, but are also expected to be a useful option for users who prefer larger text.

But larger font size is not the most important feature of these two settings. After all, internet browsers offer the option of increasing the font size. However, what internet browsers do not do is adapt the layout to the new settings. So, the width of the entries, indentation, size of gaps between different parts of the entry, and size of boxes (e.g. usage notes) remain unchanged. Thus, the settings of these features had to be changed to conform to larger fonts.

7.2 Settings customised to external characteristics of students

The main external characteristics of students that shape different settings (style sets) of this dictionary Model are place of study, native language, and subject of study. Information in the dictionary is often differentiated using these criteria, so it can be customised for specific groups of students.

7.2.1 Language variety settings (based on place of study)

Place of study can be used to determine the preferred target variety of the student. The selection of language variety is important because it determines which database will be used for extracting entry information. Thus, the language variety setting affects all parts of the entry rather than just individual parts (which is the characteristic of most other settings).

The information about the place of study can be obtained automatically by locating the IP (Internet Protocol) address of the student's computer. The IP address provides the information about the city and country (but not university) of the user. Language variety can then be automatically assigned. For example, for students accessing the dictionary from the UK, the British English language variety is automatically selected.

The location may not always allow automatic selection of language variety. For example, it is difficult to determine the preference of language variety for students from countries where English is not an official language. Also, NSs of English studying in another English-speaking

country (e.g. Americans studying in the UK), may still prefer to use their language variety. For this reason, the students should be also offered an option of changing the language variety at any stage of dictionary consultation.

One dictionary feature that is closely connected with language variety is audio pronunciation. Each of the two language varieties discussed in this thesis would have a separate audio pronunciation, which would be selected automatically with the language variety.

7.2.2 Settings based on native language

Native language is an important characteristic that distinguishes students in the way they use dictionaries. NSs and NNSs differ in the entry features they consult, and in the frequency they consult them. Each of the two groups of students requires a different set of information, its own default output (style set). Style sets are developed by using user profile information, and findings of past research on dictionary use.

Table 84 shows the default settings for the two groups of students. Most of the information is displayed by default to both groups. On the other hand, quick definitions and etymology are not displayed to either of the groups. Regular inflected forms are also not displayed, whereas irregular inflected forms are.

Table 84. DOAE style sets: Default settings for NS students and NNS students.

Part of entry	NS students	NNS students
word class	✓	✓
grammar labels		✓
pronunciation		✓
inflected forms (regular)		
inflected forms (irregular)	✓	✓
variant forms	✓	✓
variant spellings	✓	✓
menus	✓	✓
domain/style/usage/frequency/etc. labels	✓	✓
quick definitions		
definitions	✓	✓
examples	✓	✓
constructions	✓	✓
synonyms	✓	✓
frequent patterns (collocations)	✓	✓
usage notes	✓	✓
cross-references	✓	✓
frequency graphs	✓	✓
etymology		

The main differences between the two style sets lies in pronunciation and grammar labels (e.g. *transitive*), the features which are offered to NNSs only (because NSs rarely consult them). Also, grammar information may act as a distraction to NSs who often lack the knowledge of how to interpret it.

It is worth noting that pronunciation, frequent patterns and usage notes are displayed to NSs, even though the survey has shown that NSs rarely consult these types of entry information. The decision to include these entry features by default was made because their contents differ from the ones in existing dictionaries, namely they focus on academic language. This makes these features more relevant for students, and thus increases the possibility that they will be consulted.

7.2.3 Settings based on the subject of study

The subject of study influences the output of three entry features: sub-entry order, sense order and frequent patterns. These settings do not clash with the style sets for NSs and NNSs, as they merely determine the order in which information is displayed (sub-entry order, sense order), or which part of the information is displayed (e.g. domain-specific frequent patterns).

Sub-entries of certain headwords with more than one word class need to be presented in a different order for different domains due to different frequency distribution of word classes. As discussed in 6.3.1.1, different variants of sub-entry order are created, and labelled with relevant domain labels. Settings for *attribute*, based on domain distribution information (Table 46), are provided as an example (Table 85 below)[127].

[127] These settings are only for illustration purposes. Customisable sub-entry ordering could not be included in the DTD of this dictionary model because it requires programming.

Table 85. DOAE style sets: Sub-entry order settings for the entry *attribute*.

sub-entry order 1	domain label	sub-entry order 2	domain label
1. *verb*	Architecture	1. *noun*	✓ Architecture
2. *noun*	✓ Arts and Art History	2. *verb*	Arts and Art History
	✓ Linguistics		Linguistics
	✓ Music		Music
	✓ Archaeology		Archaeology
	✓ History		History
	✓ Philosophy		Philosophy
	✓ Religion		Religion
	Business		✓ Business
	Economics		✓ Economics
	Finance		✓ Finance
	✓ Law		Law
	✓ Anthropology		Anthropology
	✓ Education		Education
	✓ Politics		Politics
	Psychology		✓ Psychology
	✓ Sociology		Sociology
	Computing		✓ Computing
	✓ Engineering		Engineering
	Mathematics		✓ Mathematics
	✓ Physics		Physics
	✓ Biochemistry		Biochemistry
	✓ Biology		Biology
	✓ Chemistry		Chemistry
	✓ Geography		Geography
	✓ Medicine		Medicine
	✓ Sports		Sports
	✓ Veterinary Science		Veterinary Science

Whereas the setting for sub-entry order can be shared by several different subjects, the setting for sense order is specific to each subject. Senses are prioritized according to the relation to the subject using sense labels assigned during the analysis (see 6.3.3.7). Table 86 shows the sense order settings for Chemistry. Priority is always given to main senses which apply to all the domains (1). These are followed by senses that apply to a wider range of related domains – for Chemistry, these are first Sciences senses, followed by any senses that include Life Sciences (L2 domain category to which Chemistry belongs) (2). Next are Chemistry senses (3), followed by senses from other Life Sciences domains (4). The next group in the sense order is represented by senses of Applied Sciences (5), which represent the category of domains neighbouring Chemistry, followed by senses of individual domains in that category (6). At the bottom of the sense order are all Arts and Humanities senses, starting with broader senses (7) and eventually concluding with domain-specific ones (8).

Table 86. DOAE style sets: Sense order settings for Chemistry.

1	Main.sense
2	Sciences.L1.sense
	LifeSciences.L2.sense
	Arts.LifeSciences.L2.sense
	Humanities.LifeSciences.L2.sense
	BusinessSciences.LifeSciences.L2.sense
	SocialSciences.LifeSciences.L2.sense
3	Chemistry.L3.sense
4	Biochemistry.L3.sense
	Biology.L3.sense
	Geography.L3.sense
	Medicine.L3.sense
	Sports.L3.sense
	VeterinaryScience.L3.sense
5	AppliedSciences.L2.sense
	Arts.AppliedSciences.L2.sense
	Humanities.AppliedSciences.L2.sense
	SocialSciences.AppliedSciences.L2.sense
	BusinessSciences.AppliedSciences.L2.sense
6	Physics.L3.sense
	ComputerScience.L3.sense
	Engineering.L3.sense
	Mathematics.L3.sense
7	ArtsHumanities.L1.sense
	Arts.Humanities.L2.sense
	Arts.BusinessSciences.L2.sense
	Arts.SocialSciences.L2.sense
	Humanities.BusinessSciences.L2.sense
	Humanities.SocialSciences.L2.sense
	BusinessSciences.SocialSciences.L2.sense
8	Arts.L2.sense
	Humanities.L2.sense
	BusinessSciences.L2.sense
	SocialSciences.L2.sense
	Architecture.L3.sense
	ArtsArtHistory.L3.sense
	Linguistics.L3.sense
	Music.L3.sense
	Archaeology.L3.sense
	History.L3.sense
	Philosophy.L3.sense
	TheologyReligion.L3.sense
	BusinessManagement.L3.sense
	Economics.L3.sense
	Finance.L3.sense
	Law.L3.sense
	Anthropology.L3.sense
	Education.L3.sense
	PoliticsInternationalRelations.L3.sense
	Psychology.L3.sense
	Sociology.L3.sense

Frequent patterns settings also use labels, albeit fewer of them (only level 1 and 2), but here the settings focus on which information will be displayed. Thus, only information relevant for a particular domain is selected. For Chemistry, for example, four types of information are selected: general frequent patterns (common to all or most domains), Sciences patterns, and frequent patterns typical of Applied Sciences, and Life Sciences (Table 87).

Table 87. DOAE style sets: Frequent patterns settings for Chemistry.

Collocate.box.pattern.general	✓
Collocate.box.pattern.ArtsHumanitiesL1	
Collocate.box.pattern.SciencesL1	✓
Collocate.box.pattern.ArtsL2	
Collocate.box.pattern.HumanitiesL2	
Collocate.box.pattern.BusinessSciencesL2	
Collocate.box.pattern.SocialSciencesL2	
Collocate.box.pattern.AppliedSciencesL2	✓
Collocate.box.pattern.LifeSciencesL2	✓

Settings for sense order and frequent patterns can also be tailored to the needs of Combined Honours students who are sometimes studying two seemingly unrelated courses. The settings for students of English language and Psychology (a Combined Honours combination available at Aston University) are presented in Table 88 and Table 89.

Table 88. DOAE style sets: Sense order settings for Combined Honours: Linguistics and Psychology.

1	Main.sense
2	ArtsHumanities.L1.sense
	Arts.SocialSciences.L2.sense
	Arts.Humanities.L2.sense
	Arts.BusinessSciences.L2.sense
	Humanities.SocialSciences.L2.sense
	BusinessSciences.SocialSciences.L2.sense
	Arts.AppliedSciences.L2.sense
	SocialSciences.AppliedSciences.L2.sense
	Arts.LifeSciences.L2.sense
	SocialSciences.LifeSciences.L2.sense
3	Linguistics.L3.sense
	Psychology.L3.sense
4	Architecture.L3.sense
	ArtsArtHistory.L3.sense
	Music.L3.sense
	Anthropology.L3.sense
	Education.L3.sense
	PoliticsInternationalRelations.L3.sense
	Sociology.L3.sense
5	Humanities.BusinessSciences.L2.sense
	Arts.L2.sense
	Humanities.L2.sense
	BusinessSciences.L2.sense
	SocialSciences.L2.sense
6	Humanities.LifeSciences.L2.sense
	BusinessSciences.LifeSciences.L2.sense
	Humanities.AppliedSciences.L2.sense
	BusinessSciences.AppliedSciences.L2.sense
7	Archaeology.L3.sense
	History.L3.sense
	Philosophy.L3.sense
	TheologyReligion.L3.sense
	BusinessManagement.L3.sense
	Economics.L3.sense
	Finance.L3.sense
	Law.L3.sense
8	Sciences.L1.sense
	LifeSciences.L2.sense
	AppliedSciences.L2.sense
9	Chemistry.L3.sense
	Biochemistry.L3.sense
	Biology.L3.sense
	Geography.L3.sense
	Medicine.L3.sense
	Sports.L3.sense
	VeterinaryScience.L3.sense
	Physics.L3.sense
	ComputerScience.L3.sense
	Engineering.L3.sense
	Mathematics.L3.sense

The approach to sense ordering and frequent pattern selection resembles the approach for students on a single course (in that general and broader senses and patterns and are given priority over more domain-specific ones). The main difference between the approaches is that in this case, two domains, Linguistics[128] and Psychology (rather than one) are considered, so two

[128] Linguistics is selected because in the absence of sense label for English language, it is considered the most likely domain of the 28 available to be selected by students of English language.

sets of L2 sense combinations (see sense group 2 in Table 88) and neighbouring domains (see sense group 4 in Table 88) need to be included.

Table 89. DOAE style sets: Frequent patterns settings for Combined Honours: Linguistics and Psychology.

Collocate.box.pattern.general	✓
Collocate.box.pattern.ArtsHumanitiesL1	✓
Collocate.box.pattern.SciencesL1	
Collocate.box.pattern.ArtsL2	✓
Collocate.box.pattern.HumanitiesL2	✓
Collocate.box.pattern.BusinessSciencesL2	✓
Collocate.box.pattern.SocialSciencesL2	✓
Collocate.box.pattern.AppliedSciencesL2	
Collocate.box.pattern.LifeSciencesL2	

So how do different settings affect the output? Figure 106 and Figure 107 in Appendix 10 show the senses and frequency patterns of the entry *fact* using Chemistry and Combined Honours (Linguistics and Psychology) settings respectively[129]. The order of senses 6-11 in the two entries is different; the Chemistry entry offers the Biochemistry sense and the Computing sense before senses from Arts and Humanities, while the opposite is the case in the Combined Honours entry. Both entries offer frequent patterns, but differ in collocates provided in line 2.

The subject of study could be also used as a further criterion, in combination with NS status, for determining which entry features should be displayed. This would mean that different style sets could be used for NS students and NNS students of the same subject. For this Model, however, the user profile developed has not shown that the subject of study could be used to further distinguish between NSs and NNSs, so this was not incorporated in the Model.

One subject, however, that does influence the output is Linguistics. Students of Linguistics are studying language in addition to using it for functional purposes. They therefore require more metalinguistic information than students of other subjects. The style set for Linguistics uses the same settings as the style set for NNSs; therefore, the changes affect only NSs who are provided with pronunciation and grammar labels – the reason being that they are expected to need this information, and possess the knowledge to interpret it.

[129] The focus is on senses and frequent patterns; thus, other features such as examples, constructions, inflected forms, and etymology were omitted from the output.

7.3 449 style sets – 449 dictionaries

So far, four different types of setting have been discussed, and since each of them has a certain number of options, many different variations are possible. There are 448 variations (see below) and each requires its own style set. The 449th variation is the one where no variables are selected – it uses the default setting (default style display, NNS output, British English), original database ordering, and displays all the frequent pattern information available. The default setting has been used to produce the output of all sample DOAE entries, offered in Appendix 12.

Number of style sets = 4 style and formatting settings
x
2 language variety settings
x
2 native-language settings
x
28 subject of study settings = **448**

Because each style set caters for a different type of user, it can be argued that each style set represents a separate dictionary. All the different dictionaries are forms of DOAE, and the users of these dictionaries are students, but each dictionary answers a different set of needs of a specific type of student user.

So how do these style sets translate into practice, i.e. dictionary entries? This is illustrated by the entry *attribute*, using style sets for two different types of student – one for a NS student of Engineering, and the other one for a NNS student of Business and Management. Default style and formatting, and the British English version have been used as default for both. Although the entries for the two types of student (Figure 108 and Figure 109 in Appendix 10) share most of the contents, they do differ. The most significant differences are in the order of sub-entries, and in the order of senses. Other differences are found in frequent patterns (Business and Management students are offered an additional group of collocates), and in pronunciation and grammar labels, which are omitted from the entry for NS students (in this case, students of Engineering).

The differences between the entries produced by using different style sets, as was shown in the example of the entry *attribute,* clearly cater for the differences in the needs of different types of student. The tailored order of information means that the relevant information can be found more quickly. Similarly, the tailored output in terms of the types of information offered

means that the user can focus on the information that is relevant, and likely to be frequently consulted.

In order to be able to use the developed style sets, the four types of information discussed so far would need to be collected from the student. This should be done once only, when the students uses the dictionary for the first time. Internet browsers allow this by saving cookies on one's computer where the information on preferences for a particular website is stored. However, cookies may be deleted after a certain period. The better alternative is to have users sign up for the dictionary use, and create a profile (where the necessary information is provided). Then, the style set is automatically assigned to the user profile, and activated every time the user signs in. Furthermore, any changes to the profile are automatically reflected in the style set.

7.4 Customisable options

Pre-determined settings have been developed according to the way in which students predominantly consult dictionaries. This has resulted in some of the features being omitted, and students should be able to activate these features while using the dictionary. These features are part of the customisable options. One of the customisable features has already been discussed, namely different styles and formatting settings. Other customisable features are listed in Table 90, along with various options available to students.

Table 90. DOAE: Additional customisable options available to the user.

Part of entry	Option	Extra Options
word class		full / abbreviated
grammar labels	Display/Hide	
pronunciation	Display/Hide	IPA / phonemic / nonphonemic
inflected forms (regular)	Display/Hide	
quick definitions (if available)	Display/Hide	without main definitions / with main definitions
extra examples	Display/Hide	
etymology	Display/Hide	
language variety		British English / American English
subject of study		Archaeology / Chemistry / etc.

It can be seen that the students are not given full power over dictionary contents – only some microstructural features can be manipulated. These are mainly features that have been omitted from the default output, and/or offer additional options (e.g. three types of

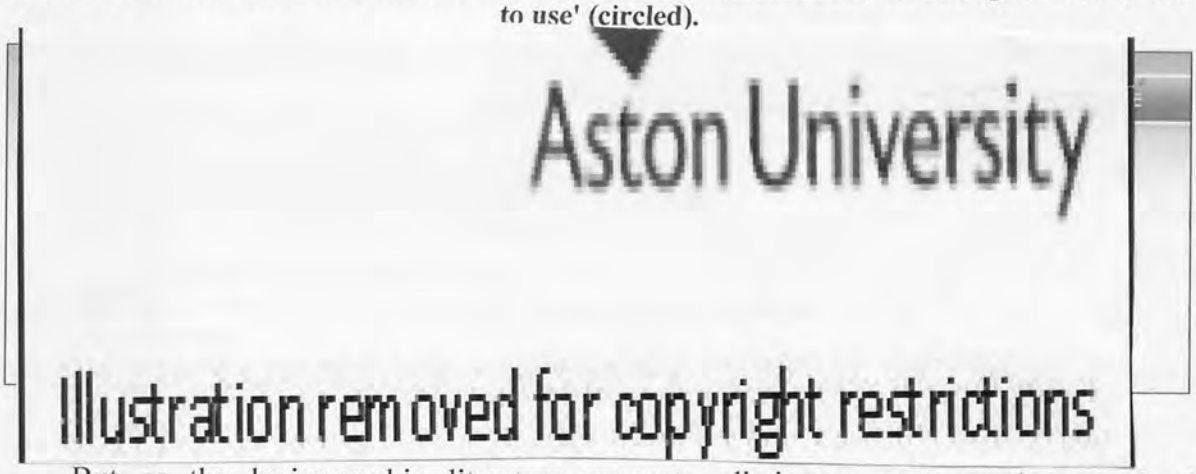
pronunciation). The language variety and subject of study are part of user profile but should also be possible to change.

Students should be also offered a pre-selected set of options which would reduce/increase the amount of information on screen. A similar approach is found in e-MED and uses Show More/Show Less buttons. The activation of Show More button displays all the information that is part of the chosen default setting, whereas the activation of Show Less button reduces the amount of information on the screen (e.g. removes menus and pronunciation, reduces the number of examples), thus giving more prominence to definitions. This feature is especially useful at longer entries.

7.5 Tips and hints on how to use the dictionary

Students need to be provided with basic instructions on how to use the dictionary. The instructions can be offered on a separate page, and the link to that page should be available at any point of dictionary consultation. A similar approach is offered by e-LDOCE which has 'How to use' link in the main toolbar (Figure 87).

Figure 87. e-LDOCE: Toolbar with search window, and various links, including the link to 'How to use' (circled).



But as the lexicographic literature suggests, dictionary users rarely consult the instructions on how to use the dictionary. Other means should thus be used, for example, boxes with hints and tips. MED, for example, has links to tips in red underlined text (Figure 88). Figure 89 shows an example of a window when a link is clicked on. Also available are windows explaining certain entry features, which appear when a mouse cursor is positioned over the feature (see Figure 88 for window explaining pronunciation).

Figure 88. e-MED: Explanation window of pronunciation and tips (adjective *good*).

It is often better to show the students how to do something rather than using a lengthy description. DOAE could offer short video tutorials for most common problems. These video tutorials would show a recorded screen of someone performing a search and using particular functions. Some dictionaries on CD-ROM already offer such content but not as an integral part of the dictionary.

Instructions, tips and hints should be used moderately and subtly. They will unavoidably need space on the screen, but they must not take any prominence away from the entries. Otherwise, these forms of help may defeat their own purpose, focussing user's attention on the explanation of the features rather than their contents.

7.6 Living dictionary: users as shapers of dictionary contents

No matter how many customisable options are available, it is unavoidable that the students will at some point encounter problems. This may be due to their unfamiliarity with the dictionary, the poor user-friendliness of certain dictionary contents, or the lack of the relevant

contents (e.g. entries). But the main point is, as stated by the authors of the Macmillan English Dictionary (Rundell, 2002), is that if students have problems finding and using the information they need, it is not their fault but the dictionary's fault.

The dictionary must therefore be constantly improved, and one advantage of online dictionaries is that individual improvements can be made available as soon as they are implemented – there is no need to wait for the next edition containing all the improvements, as is the case with print dictionaries. The improvements are based on students' experience with the dictionary. Three levels of student involvement in shaping the dictionary contents are suggested:

1. Passive. Students' dictionary use is monitored, and new entries and features (e.g. help with common misspellings, etc.) are added, based on unsuccessful searches or difficulties resulting in long searches. Any observed changes in patterns of dictionary consultation can be used to modify default style sets, or develop new ones.
2. Semi-active. Students are invited to submit suggestions for new entries and features. The suggestions are then reviewed by lexicographers, and the entries and/or features created if found relevant.
3. Active. Students can submit new entries with definitions and other information, and/or submit senses, definitions, etc. to existing entries. The suggested contents are reviewed by lexicographers, and offered to other users and experts for comments, before they are made an integral part of the dictionary. This approach is already used by dictionaries such as MED and Merriam Webster (these two dictionaries use the feature 'Open Dictionary').

Encouraging students to participate in the shaping of the dictionary will not only improve the dictionary and make it more user-friendly, but it is also likely to prompt students to use the dictionary more frequently.

One aspect that should not be overlooked is that students should be invited to suggest improvements not only to the dictionary contents, but also to the look of the dictionary and the website. There is however a balance to be struck here; modifications to the website must not significantly affect the authoritative look of the dictionary otherwise it may deter students from using it.

8. DISCUSSIONS

This chapter reviews the proposed Model for DOAE, and the methods and software used to develop it. The potential benefits as well as shortcomings of both the Model and the methodology are evaluated, and suggestions for further improvements are made.

The second half of the chapter turns to the implications of this research, both in terms of lexicography and pedagogy. The chapter concludes by making recommendations for future research, based on some of the findings of this research, and considering some of the solutions proposed by this dictionary Model.

8.1 Reviewing the Model for DOAE

So far, all the design and implementation aspects of this dictionary Model have been discussed. It is now time to reflect on how the proposed Model meets the needs of students, and consider how the Model would be turned into an actual product. A comparison with existing dictionaries and PDEV is also made. Lastly, a few potential enhancements to the Model are proposed.

8.1.1 Advantages of the Model for DOAE

The proposed Model has several advantages, for both students and lexicographers. The main beneficiaries of the Model are of course students, because the Model addresses the current lack of a suitable dictionary resource for this type of user. There are also implications for other users, as many features of the Model can be applied to other dictionaries.

The main advantage of the Model, distinguishing it from existing dictionaries, is the focus on academic language. Based on a corpus of academic language, the proposed dictionary Model offers meanings, collocational patterns, examples, usage notes and other features that reflect how words are used in academic register. It describes the language that students need help with, as opposed to general language they are already familiar with. Even if this was the only thing that made the Model different from any existing dictionary, it would justify turning the Model into an actual dictionary.

The next advantage is the focus on the target users, i.e. students. Decisions on which entry features to include or exclude have been informed by research on the dictionary habits of students. The same is true of the contents or form of the features – more frequently consulted

features (e.g. definitions, examples) are given more prominence, and assigned more space in the entry. A very good example of this is full-sentence definitions; the definitions contain important phraseological information which, if provided separately, is rarely consulted by students.

The dictionary Model is student-friendly for an additional reason, because it caters for different types of student, rather than students as a single group. Students of different subjects, different NS status, and even those with different visual needs are each provided with their own customised dictionary. Furthermore, the Model offers the students additional customisable features which ensure that the dictionary can be tailored to their needs even further.

Students are also offered the opportunity to shape the contents after the dictionary is made. This means that the dictionary can be constantly improved. Asking for students' contributions could motivate the students to use the dictionary more frequently, or to choose it over its competitors.

Many of the user-friendly features of the proposed Model are possible only because the Model is offered in the online format. The online format has been selected not only because of its many advantages over other dictionary formats, but also because it is the format that is preferred by the highest percentage of students, especially when doing academic work (4.1.1.2). In addition, the online format represents the current trend in lexicography; lexicographers have been devising more and more methods of utilizing the full potential of the format and, it seems, will continue to do so.

Another advantage is that the online format is used as the basis for designing the dictionary database. As a result, the dictionary contents and features are designed with the online format in mind – a large amount of information for the entry is recorded, without having to think about any space restrictions. This makes the Model different from most existing online dictionaries which, although they do exploit some features of the online format, show signs that the contents were originally designed for another format.

Quality and consistency in recording the database contents is ensured by the use of a modern corpus analysis tool (Sketch Engine) and a modern dictionary-writing system (TshwaneLex). The two resources have been created mostly or solely for lexicographic purposes, and help lexicographers to reduce the time spent on he analysis. Additionally helpful is the compatibility of Sketch Engine and TshwaneLex, so that saving information in the database can be semi-automatic.

The user-friendliness of TshwaneLex also allows lexicographers and editors greater input into the design of the dictionary, from the database to the entry features. Some

computational knowledge is required, however this is a characteristic expected of a modern lexicographer. Increasing the lexicographers' and editors' role in shaping dictionary contents constitutes a much needed reversal of a practice that is still used by many publishing houses, namely entrusting the design of electronic formats to external parties or people with little interest in language (e.g. software companies). Lexicographers and editors can thus influence not only which information is presented but also how it is presented, input that they have so far had mainly in compiling print dictionaries.

8.1.2 The publication potential of DOAE

In order to be accepted as realistic, the proposed Model for DOAE needs to be discussed in terms of the realities of dictionary publishing. This section will outline how the dictionary proposed by the Model would be published.

First and foremost, the dictionary should be available for free. This follows the current trend of online publishing. It is important that all the dictionary contents are made available, as opposed to only part of it (as is the case with dictionaries such as e-MED and e-LDOCE). Another reason for making the dictionary freely available is that the survey has shown that students prefer to access freely available dictionaries; it is the price (or rather lack of it) than is given priority over quality.

Offering the dictionary for free also makes sense because one rarely has to pay for internet content these days. In fact, only the content that is unique, for example music, films and academic articles, is payable. From the students' perspective, dictionaries are not unique – they define the same words, albeit in different ways. Students are unlikely to be prepared to pay to get the answers to their linguistic problems – this trend is clearly evidenced by the frequent use of Dictionary.com.

One important problem with making the dictionary available for free is that it does not make it a cost-effective proposition for dictionary publishers. A dictionary is not created overnight, and the people involved in the process of compiling it need to be paid. There are three potential solutions to this issue:

a) The dictionary is a commercial project, and the costs are covered by offering advertising on the dictionary website. This is the practice seemingly followed by most existing online dictionaries (e.g. Dictionary.com, e-MED, e-CALD).

b) The dictionary is a commercial project, and the costs are mainly covered by charging a subscription fee to institutions, i.e. universities, rather than individual students. This approach is currently used by the Oxford English Dictionary (OED). The advantage of this approach is that universities then promote the dictionary to their students (and as shown by survey data in the case of Aston University, this seems to have been quite efficient in getting students to use it),[130] even though the OED is far from suitable for students.

c) The dictionary is a non-profit project, which makes the dictionary open to public funding. This solution is appealing because it allows the possibility to turn the dictionary into an academic project, giving the academic community an opportunity to shape the resource that is intended for them. Another advantage of the dictionary being a non-profit (academic) project is that it may result in easing the process of obtaining copyright permission for the contents of the corpus of academic language.

All three solutions make the dictionary contents freely available to students, but the first solution has the disadvantage in that it needs to assign valuable screen space to advertising. This problem is not shared by the second and the third solution, which are thus considered more appropriate for the publishing of this dictionary Model.

8.1.3 Other formats

The online dictionary format has been selected as the format for the proposed Model, and since this is the most complete dictionary format currently available, the creation of other formats of DOAE would not require any additional analysis. A new style set would need to be designed for each separate format. The majority of work would go into deciding on the best way to present the entry information, and which types of information to omit.

A special case is mobile phones with internet capability with which users would be able to access the (full) online version of the dictionary. One thing to bear in mind, though, is that screens of mobile phones are much smaller than computer screens – consequently, the amount of information on the screen would have to be reduced to make the navigation more manageable and user-friendly.

[130] The OED is a historical dictionary and thus not suitable for most students.

The problem of small screen size applies also to pocket electronic dictionaries (PEDs). In addition, the size of PED's storage is more limited than that of an online dictionary so some features may need to be omitted, or reduced. PEDs do however benefit from the fact that they have the dictionary already installed, thus the user does not require an internet connection to use the dictionary.

CD-ROMs are even more limited than PEDs in the amount of data they can store, but they do benefit from the larger computer screen. CD-ROM format therefore loses more in terms of contents than in terms of functionality.

Print format, on the other hand, would deprive users of many key components of the contents and functionality. The contents would need to be reduced by using various space-saving techniques, such as omitting the less frequent entries/senses, using abbreviations, and shortening examples or reducing the number of examples. Each space-saving technique takes something away from the entry as it was envisaged by lexicographers, thus making it less user-friendly and potentially less effective. Even more important is the loss of functionality – contents would have to be predetermined (by lexicographers), so the dictionary could not cater for different types of user, and the benefits of electronic format such as quick searches would be lost.

The question is not whether other formats of the Model can be created, but whether they should be created. Pocket electronic formats and CD-ROM format are preferred by only a small minority of students. Print format, on the other hand, is preferred by 40% of the students, but it is not used frequently for academic work. The popularity of these dictionary formats among students is continuously decreasing, mainly on account of their increasing use of online dictionaries[131].

It thus seems more sensible to focus on providing the use of the online format to students, not only because this dictionary format can cater for different types of students, but also because it is the preferred format of students when doing academic work. In addition, future generations of students are likely to be even more inclined to use online dictionaries because they will encounter, and learn how to use, online dictionary format very early in their lives (without doubt earlier than current students). In fact, the online format may become the first dictionary format that future generations of students will encounter.

[131] This is also indicated by the differences in results obtained by Hartmann (1999), and the results of the survey conducted for the purposes of this thesis (see 4.1.1.2).

8.1.4 Comparison of DOAE entries with existing dictionaries and PDEV

A comparison of the sample DOAE entries with corresponding entries in existing dictionaries is a good way of identifying some of the reasons why students should choose to use DOAE. Certain aspects such as coverage and entry selection cannot be compared because this research produced only a model rather than a complete dictionary.

The first aspect to be compared is sense order. As sense order in the proposed dictionary is not fixed (it can be customised to different types of student), the sense order in the database entries (default style set) was used for comparison. There were, as expected, many differences, partly because there are already differences among the dictionaries themselves. One thing that distinguished several sample entries from the corresponding entries in dictionaries was that the first sense in the sample entry was often not among the top senses in dictionary entries, or was not even a sense. This was the case at entries *significant* (Table 91; for complete entries, see Table 139 in Appendix 11), *obtain, fact,* and *argue* (Table 95).

Table 91. Comparison of sense order of *significant* in sample DOAE entry and in existing dictionaries.

DOAE sample entry	CED CD-ROM	NODE CD-ROM	MWCD CD-ROM	e-LDOCE, LED CD-ROM	COBUILD CD-ROM
1.	4.	3.	2b	example under 2	/
2.	/	1.	2a	2.	1.
3.	3.	2.	2a	1.	2.

DOAE sample entry	e-MED	e-CALD	e-OALD	Dictionary.com
1.	/	/	pattern example under 1	3.
2.	1.	**significant** (important)	1.	/
3.	2.	**significant** (important)	1.	1.

Closely related to differences in sense order are differences in treatment. The senses in sample entries that do not feature in existing dictionaries are often found to be reduced to the role of a construction/phrase with a definition and/or example(s), a marked collocational pattern in an example, or just an example. For example, the phrases *in fact*, *the fact (of the matter) is*, and *a fact of life*, which have sense status in the sample entry *fact*, are offered in the Phrases section at the end of the entry in both NODE CD-ROM and e-MED. The low prominence given

to these phrases in dictionaries is not user-friendly, especially considering that *in fact* occurs in 35% of the concordance lines of FACT in CAJA.

The differences discussed so far have been mainly connected with the positioning and prominence of senses and patterns. In other words, similar information can be found in different dictionaries, it is just a question of where in the entry. Much more important are differences in the information contents, and this applies to definitions and examples in particular.

Differences between the wording of definitions in the sample DOAE entries and existing dictionaries[132] are a good indicator of how 'academic' the sense or the word is. Senses in the sample entries that exhibit significant differences in wording are often very specific to academic language, or are used more frequently in academic language than in general language. For example, sense 1 of the sample entry for *obtain* has the following definition:

> 1. If you **obtain** a result or data, you get it by doing research or conducting an experiment.

In other dictionaries, this meaning is covered by the definition of the main sense that is synonymous to *get* (offered as sense 2 in the sample entry). Here are definitions from a selection of dictionaries:

> 1. to gain possession of; acquire; get (CED CD-ROM)
> 1. get, acquire, or secure (something) (NODE CD-ROM)
> 1. to get something that you want, especially through your own effort, skill, or work (LED CD-ROM)
> 1. To **obtain** something means to get it or achieve it. (COBUILD CD-ROM)
> 1. to get something that you want or need, especially by going through a process that is difficult (e-MED)

The definition in the sample entry avoids the use of a non-specific general superordinate *something*, used in some of the dictionaries, by introducing more specific superordinates for an object (which are also salient collocates) that reflect typical usage in academic language. In fact, superordinates are often the distinguishing element between the definitions of academic senses in sample entries and in existing dictionaries. The superordinates in the definitions of the sample DOAE entries are more specific, and more typical of academic use (see Table 92, and also *obtain* above). In either case, the high relevance of such superordinates for students increases the user-friendliness of the definition.

---

[132] There are also some differences in definition form, but that is to be expected as the Model uses more than one type of definition whereas the dictionaries tend to stick to one only. The differences tend to be greatest in NS dictionaries (with NODE often being an exception); on the other hand, and least in learners' dictionaries.

Table 92. DOAE definition containing more specific superordinates than definitions in existing dictionaries (superordinates are offered in bold) – a sense in the entry for *argue*.

DOAE sample entry	1. If someone argues a **view** or an **idea** in an article or book, they present the idea and support it with evidence.
LED CD-ROM, e-LDOCE	2. to state, giving clear reasons, that **something** is true, should be done etc
e-MED	2. to give reasons why you believe that **something** is right or true
e-OALD	2. to give reasons why you think that **sth** is right/wrong, true/not true, etc., especially to persuade people that you are right
COBUILD CD-ROM	4. If you argue that **something** is true, you state it and give reasons why you think it is true.

Similarities in the wording of definitions are found in senses of words that are not register-specific, for example grammatical words. This merely suggests that the meaning of these words or senses in academic language is similar to or the same as in general language. The usage, however, is different – and this is evidenced by examples.

Examples in the sample entries are representative of academic language, and of the language of academic articles in particular, so it is not surprising that their contents are quite different from the examples in existing dictionaries, which reflect general language. The differences are evident in collocational behaviour and in surrounding co-text. A good instance of this are the examples under the main sense of *method* (Table 93 below), with the examples in the sample DOAE entry pointing to collocates such as *qualitative, quantitative, use, analysis,* and the examples in existing dictionaries pointing to collocates such as *traditional, teaching, new*, and *effective.* Only a few collocates are shared, for example *of, for, control* and *alternative.*

Examples in the sample DOAE entries also differ from examples in existing dictionaries in complexity, length, and form. Examples in the sample DOAE entries often contain subordinate clause(s), or consist of more than one clause (see DOAE examples in Table 93). In some cases, examples even consist of more than one sentence (see the example for *fact* below). Consequently, examples in sample DOAE entries also tend to be (much) longer than examples offered in existing dictionaries. All examples are complete sentences, as opposed to existing dictionaries which sometimes provide only phrases (see Table 93).

explained/supported/complicated/etc. by the fact that

Roten and Mullineaux (2002) find that commercial banks charge lower fees than investment banks. This may be explained by the fact that in their sample commercial banks have only just entered the market.

(the DOAE entry for *fact*, sense 2)

Table 93. Examples under the main sense of *method* in DOAE and 5 dictionaries.

DOAE sample entry	*They **used** both qualitative and quantitative **methods** to collect and analyse their data.* *We next describe our data source and **method of analysis**, after which we present our statistical findings.* *Over much of the last four decades, fertility control methods have been generally available for wealthier women through private health clinics (Barroso 1984).* ***method for doing something*** *Categorization tools normally have a method for ranking the documents in order of which documents have the most content on a particular topic.* ***method of doing something*** *This facilitated an opportunity for other pupils to argue an alternative method of solving the equation.*
e-LDOCE, LED CD-ROM	*traditional teaching methods* *I think we should try again using a different method.* ***method of/for (doing) something*** *Today's methods of birth control make it possible for a couple to choose whether or not to have a child.* *effective methods for the storage and retrieval of information*
e-MED	*It was a handmade rug produced by traditional methods.* *Farming methods haven't changed here for decades.* ***method of:*** *We are trying to develop new methods of pollution control.* ***method of doing something:*** *They have adopted an alternative method of financing the scheme.* ***method for doing something:*** *Vaccination is one of the most effective methods for preventing disease.*
COBUILD CD-ROM	*The pill is the most efficient method of birth control.* *...new teaching methods...* *The usual method of getting through the Amsterdam traffic is to cycle to your local railway station and take the train.*
NODE CD-ROM	*a method for software maintenance* *labour-intensive production methods*

The examples in the sample entries may be criticized as less user-friendly than examples in existing dictionaries due to their more linguistically complex structural features, but the fact of the matter is that they reflect the authentic complexity of academic language. In addition, the examples contain some features that are very specific to academic language, for example academic conventions such as referencing (see the example for *fact* above, and the third example for *method* in Table 93). Considerable shortening and simplifying of the examples would make them unnatural and would strip them of their productive value.

Collocational information is found throughout the entries, and some comparisons have already been made while discussing other microstructural features. Here, the focus is more on

instances when collocations are made explicit, e.g. as constructions or bold items in examples. In terms of the amount of collocational information in the proposed Model, there is more similarity with learners' dictionaries than with NS dictionaries; NS dictionaries provide almost no information of this kind[133]. But none of the existing dictionaries contains collocational patterns frequently found in *academic language*; in the proposed Model such patterns are focussed on in definitions, constructions and examples, and in frequent patterns.

Overall, the proposed dictionary Model is more similar to learners' dictionaries in terms of depth of treatment and amount of phraseological information. On the other hand, the Model is similar to NS dictionaries in terms of coverage of technical senses. Content-wise, however, the Model is quite different from existing dictionaries, especially in terms of definitions and examples, two of the microstructural features most frequently consulted by students. The differences in these features confirm that existing dictionaries offer little help to students with dealing with language problems relating to academic English.

8.1.4.1 Comparison with PDEV

It is also useful to compare the sample entries with the PDEV entries. An approach similar to CPA (used by PDEV) has been used in the Model for the meaning analysis of sample entries – this ensures comparability. However, the PDEV database offers information on the percentages of individual patterns, and since the information is based on *general* language data, it offers the opportunity to compare pattern distribution in general language and in academic language.

At the time of consultation, the verb *argue* was the only verb in the sample entries that had an entry in PDEV. The comparison shows that there are great similarities between the patterns identified in PDEV, and the senses compiled for the sample entry (Table 94). In fact, every single PDEV pattern of *argue* corresponds to a sense in DOAE.

Another similarity between the PDEV entry and the sample DOAE entry *argue* is the frequency distribution of senses/patterns. There is one sense/pattern that is significantly more frequent than the others. Furthermore, all the other senses/patterns are similar in their low rates of frequency, although the percentage of sense 2 of *argue* is slightly higher than the percentage of corresponding PDEV pattern 3, and the percentage of sense 6 is slightly lower than the percentage of corresponding PDEV pattern 5.

[133] NODE is often an exception, but this is not a typical NS dictionary.

Table 94. Senses and sense percentages of the sample DOAE entry *argue* and corresponding PDEV patterns and percentages.

DOAE			PDEV		
sense no.	**definition**	**%**	**pattern no.**	**pattern (line 1) + definition (line 2)**	**%**
1.	If you **argue** a view or an idea in an article or book, you present the idea and support it by evidence. Note: **argue** is very often followed by a ***that***-clause.	90	1.	**[[Human | Institution | Document]] argue [NO OBJ] {[that-CLAUSE | QUOTE]}** [[Human | Institution | Document]] states reasons for believing [CLAUSE]	88
			2.	**[[Human]] argue {{Propositioncase} | [[Proposition]]} {[that-CLAUSE | QUOTE]}** [[Human]] states reasons for believing [[Proposition]]	2
2.	If you **argue for** or you **argue in favour of** an idea or theory, you agree with it and provide evidence that supports it.	6	3.	**[[Human | Institution | Document]] argue [NO OBJ] {for | in favour of [[Action]]}** [[Human | Institution | Document]] states reasons in favour of doing [[Action]]}	3
3.	If you **argue against** an idea or theory, you provide evidence that opposes it.	2	4.	**[[Human | Institution | Document]] argue [NO OBJ] {against [[Action]]}** [[Human | Institution | Document]] states reasons in favour of not doing [[Action]]}	2
4.	If you **argue** with someone about/over something, you discuss it because you have different opinions.	<1	6.	**[[Human 1]] argue [NO OBJ] {with [[Human 2]]} (over| about [[Proposition]] | over| about [[Anything]])** [[Human 1]] and [[Human 2]] express contrary opinions about [[Proposition | Anything]], typically in a heated or angry way	1
5.	If you **argue with** someone or someone's view, you disagree with it.	<1	7.	**[[Human]] argue [NO OBJ] {with [[Proposition]]}** [[Human]] denies the truth of [[Proposition]]	<1
6.	If people **argue**, they talk angrily to each other because they disagree.	<0.5	5.	**[[Human = PLURAL]] argue [NO OBJ] (over| about [[Proposition]] | (over| about [[Anything]]))** [[Human = PLURAL]] express contrary opinions about [[Proposition | Anything]], typically in a heated or angry way	3

Because PDEV uses a 50-million version of the BNC for its analysis and DOAE is based on CAJA (94.3 million words), the comparison effectively compares the behaviour of *argue* in general language with its behaviour in academic language. The results show there are hardly any differences in frequency and order of patterning[134]. It is therefore surprising that existing dictionaries, which are based on general language (some even on the BNC), rarely offer the first pattern (by far the most frequent pattern) as the first sense (Table 95; for complete entries, see Table 140 in Appendix 11). Furthermore, many dictionaries position the least frequent patterns at the beginning of the entry.

Table 95. Comparison of sense order in the sample DOAE entry *argue*, with the sense order in the existing dictionaries.

DOAE sample entry	CED CD-ROM	NODE CD-ROM	MWCD CD-ROM	e-LDOCE, LED CD-ROM	COBUILD CD-ROM
1.	3.	1.	3. (tr.)	2.	4., 6.
2.	2.	/	1. (intr.)	2. (pattern)	5.
3.	2.	/	1. (intr.)	2. (pattern)	5.
4.	/	/	2. (tr.)	1.	3.
5.	/	2.	/	1.	3.
6.	1.	2.	/	1.	1., 2.

DOAE sample entry	e-MED	e-CALD	e-OALD	Dictionary.com
1.	2.	argue (give reasons)	/	4.
2.	2. (pattern)	argue (give reasons)	2.	1.
3.	2. (pattern)	argue (give reasons)	2.	1.
4.	1a	argue (disagree)	1.	2.
5.	1a	argue (disagree)	1.	2.
6.	1.	argue (disagree)	/	/

8.1.5 Potential enhancements to the proposed Model

There are several enhancements that could be made to the proposed Model. For example, the recoding of information in the database, improving the output, or adding new information to the entry contents. This section briefly discusses enhancements related to adding new information to the entry contents.

One potential enhancement is adding information on common errors, which would be obtained by analysing student writing and speech. A similar feature can already be found in one existing dictionary for advanced learners (see Figure 90). Errors would be grouped according to

[134] It is worth pointing out that PDEV does not always order patterns by frequency (see for example Figure 11).

native language and/or subject of study, which would make the feature compatible with the customisability of the existing Model[135].

However, there is currently a lack of suitable corpora that would enable the analysis of student errors. More attention is paid to 'good' examples of student academic language (found in corpora such as BAWE and MICUSP). The inclusion of this enhancement therefore presupposes the creation/availability of a suitable corpus.

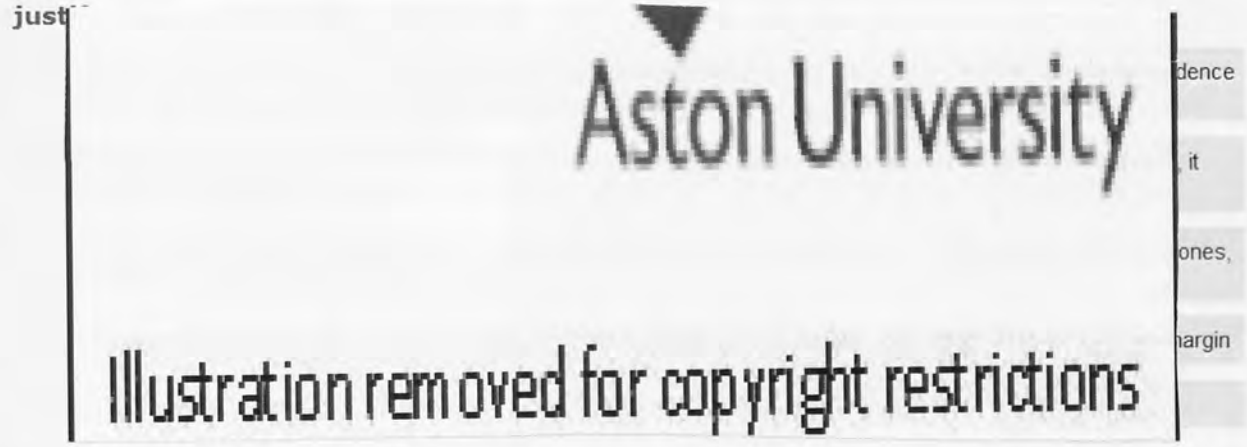
Figure 90. MEDAL2: 'Get it right' box[136].

[135] It should be stressed that the academic writing/speech of both NSs and NNSs would be analysed, as it has been established that both groups of students are learners of academic language (see 2.1.4).
[136] The image was obtained from http://www.macmillandictionaries.com/about/MED2/keyfeatures.htm#getitright.

Another potential enhancement would be to make the dictionary bilingualised, or multilingualised, by adding translations to each entry or sense. The database would contain translations into many different languages, but only the relevant language would be displayed to the user. Information on the student native language would have to be obtained beforehand, but this is already part of the proposed Model.

Translations could be very useful in helping NNS students to distinguish between different senses, and to find the relevant sense. Translations could be offered before the main definition, which would give them a role similar to signposts in advanced learners' dictionaries. This enhancement is exemplified in Figure 91 below, using Slovenian translations (provided in brackets) [137] which were produced by the researcher with the help of the Oxford-DZS Comprehensive English-Slovene dictionary (Krek, 2005; 2006) and the FidaPLUS corpus of Slovene (http://www.fidaplus.net/).

Figure 91. DOAE: The entry *justify* with Slovenian translations.



Making these enhancements would require collaboration with researchers and translators, which is yet another reason for choosing the non-profit option of publishing the dictionary (see 8.1.2).

8.2 Review of methodology

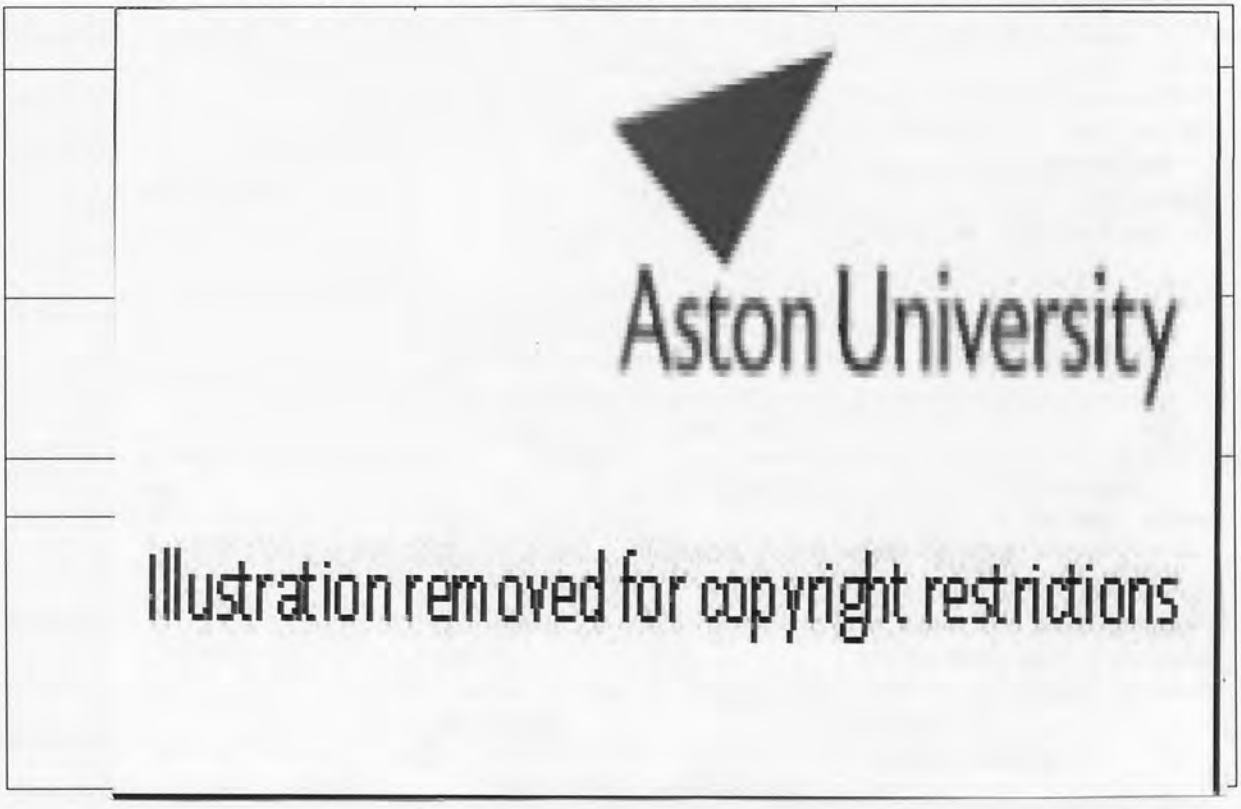
This section revisits the methodology used in designing the proposed Model, and comments on its advantages and disadvantages.

[137] Non-essential features of the entry have been omitted in this illustration.

8.2.1 Researching dictionary use

The methods most often used for researching dictionary use are questionnaires (or questionnaire-based interviews), tests, and observation. Some advantages and disadvantages of each method can be found in Table 96. Because each method has its shortcomings, it has been suggested that the best procedure is to use several methods, and then collate the results (Nesi, 2000; Lew, 2002). The multi-method approach is very difficult to adopt as it requires spending a great amount of time in preparation, implementation, and analysis. There is also a problem of achieving the comparability of results; subjects' perceptions of the way they use dictionaries, reported in a questionnaire, are likely to differ from the way they will use dictionaries in tests or observed tasks – with their normal behaviour probably being somewhere in between. Ultimately, all the methods face the same problem: they are intrusive by default as they are used to investigate an act that is by nature very private (Nesi, 2000; Nesi & Haill, 2002).

Table 96. Advantages and disadvantages of methods of researching dictionary use[138].



Dictionary-use research can use only one method and still be reliable. It is probably more important to select the most suitable method(s) for the questions that the research is trying

[138] The majority of the information in this table is based on Nesi and Haill (2002), Nesi (2000) and Hatherall (1984).

to answer. For the purposes of this research, the questionnaire was identified as the most suitable method for the following reasons:

- the research was trying to address a large group of students;
- the majority of the questions were asking for factual information, and not for the students' perceptions;
- there is already quite a lot of test-based and observation-based research, focussed on various aspects of dictionary use, with students as subjects.

Having a large group of students proved to be beneficial later for the dictionary design as it enabled the creation of many user-friendly features. This would not have been possible with any other method.

One thing that was discovered during the administration of the questionnaire was that an online questionnaire can be regarded as a completely separate research method. The subjects can answer questions more quickly and can thus answer more questions; for the researcher, the analysis is semi-automatic and thus less time-consuming. As far as user-friendliness is concerned, an online survey is the form of questionnaire that students are likely to be most familiar with. An additional benefit of administering an online questionnaire was that it provided valuable experience in designing user-friendly online content, which was then utilized for creating the dictionary Model.

8.2.2 Corpus of Academic Journal Articles (CAJA) compared to other corpora

CAJA was built because existing corpora of academic English were inadequate for DOAE purposes. This section discusses the advantages of CAJA, especially over existing academic corpora, and considers some of its potential shortcomings.

CAJA's main advantages are size and balance. With 83.5 million words (or 93.3 in Sketch Engine) it is the largest corpus of academic language known to this researcher. In addition, CAJA is without a doubt the most balanced of the academic corpora. Carefully designed domain classification and fairly equal subcorpus sizes make the corpus unique and, as demonstrated by the Model, highly useful for lexicographic purposes. Domain classification is a perfect example how a well-designed corpus can be used to develop very user-friendly dictionary features (e.g. domain-specific sense ordering).

Representativeness can be viewed as one of the CAJA's shortcomings. CAJA contents are limited to expert academic writing, and academic articles in particular. The corpus is

missing several other genres that students are likely to encounter during their studies, such as books, textbooks, course packs, syllabuses, etc. Also, the corpus contains no academic speech.

The focus on the single genre of journal articles was needed due to the time-consuming nature of collecting some of the other types of data, especially spoken data. Yet, without the focus on a single genre, it would have been difficult to do this research within the limited timescale, and to create such a large corpus with a very detailed subcorpus classification that contains carefully selected articles from leading academic journals. The lack of other genres, especially spoken data, was compensated for by consulting existing corpora of academic English.

CAJA was built primarily for lexicographic purposes, but also has some pedagogic potential (see 8.3.2.2). CAJA does however share the issue of non-availability with other corpora of academic English – there was simply no time to obtain copyright permissions for all 13,116 texts, and it is also unlikely that it would have been given. As a matter of fact, obtaining copyright permissions on academic texts will possibly be one of the main problems that the makers of DOAE will face. A potential solution is to use the Directory of Open Access Journals (DOAJ), which currently offers free access to 4,926 scientific and scholarly journals (nearly 4,000 are available in English) and 384,124 articles[139]. One of the potential issues of using DOAJ as a resource of corpus data is the probable absence of the most highly ranked journals.

8.2.3 The corpus-driven approach and the role of intuition

The comparison of the sample DOAE entries with corresponding entries in existing dictionaries has shown that in academic language words often have their own particular meanings, different frequency distribution of meanings, and different collocational patterns. Consequently, trusting the corpus is necessary since existing lexicographic resources or even intuition are not suited to represent academic language. The adoption of a corpus-driven approach for this dictionary Model has thus proven to be beneficial.

The use of a corpus-driven approach in this Model has also revealed some additional benefits of this approach for modern dictionaries. As has been demonstrated in 7.2.3, a well-devised corpus that reflects the user classification can contribute to the customizability of dictionary output. A corpus therefore also 'drives' the user-friendliness of the dictionary, in addition to providing the basis for lexicographic analysis.

---

[139] Information obtained from http://www.doaj.org/, accessed on 17 April 2010.

There are however certain aspects of the corpus-driven approach that cannot always be followed. Firstly, a narrow interpretation of the corpus-driven approach would suggest the use of raw data, but it is easier, and probably necessary, to do the analysis on lemmatised and POS-tagged data (despite occasional errors) as it saves time. This becomes particularly evident when analysing a large corpus like CAJA. Furthermore, the Word Sketch function in Sketch Engine, which has an important role in the analysis, can only be used on POS-tagged data.

Secondly, meanings and patterns not found in the corpus were sometimes added. This, however, has only been necessary because CAJA focuses on a particular genre of academic language. It is expected that a proper dictionary project would use a larger and more representative corpus of academic language, and would not need to consult any other corpora.

Thirdly, the use of unedited corpus examples is often neither practical nor user-friendly due to their complexity and length, two characteristics which reflect the nature of academic language. Nonetheless, only minor changes to corpus examples are made, which ensures that the integrity of corpus evidence is maintained.

Despite some of its limitations, a corpus-driven approach is most suitable for DOAE, especially bearing in mind that there is no such thing as a NS of academic language. But what about the intuition (of a NS of English) and its role in the corpus-driven approach? John Sinclair has always maintained that intuition, while a part of the corpus-driven linguistic analysis, should not be used until later stages in the process of dictionary compilation (see e.g. 1985). When designing this dictionary Model, it has become clear that intuition is used throughout the dictionary compilation:

a) While building the headword list or adding new headwords:
   - intuition helps to identify both single-word headword candidates (separating actual candidates from tagging errors) and multi-word headword candidates;

b) When recording basic information:
   - intuition triggers searches for potential variant forms;
   - intuition helps to identify potential errors in tagging;

c) During meaning analysis:
   - intuition assists in the interpretation of corpus data, i.e. the description of patterns, collocations, and meanings of the word;
   - intuition helps to identify (common) patterns of the word missed by Word Sketch (see 6.3.2.3.1).

Intuition is thus an important part of a corpus-driven approach, but it should be noted that DOAE is completely reliant on corpus data; even if a lexicographer's intuition prompts the search for a certain missed pattern or meaning, that pattern and meaning is not included in the entry unless it is found in the DOAE corpus.

Intuition should not be relied on when creating the dictionary output(s). Using intuition at that stage would mean using one's own linguistic knowledge to make assumptions on what the users know and do not know about the words. This would make the output subjective, and consequently not user-friendly. The user-friendliness of the dictionary can be addressed only with reference to the user profile, which is informed by research into the dictionary use of its target users.

8.2.4 Secondary resources reviewed

The secondary resources used in designing this dictionary Model were existing corpora, existing dictionaries, and Pattern Dictionary of English Verbs (PDEV). All these resources were used at various stages of both analysis and evaluation.

Existing corpora were consulted when considering the addition of missed meanings and patterns, and when creating frequency graphs. The corpora used, especially corpora of academic language, proved to be of valuable assistance, but they would have been even more valuable if they had matched CAJA in size and subcorpus domain classification. For example, a comparison of the occurrence of a particular pattern in a 93.3-million-word corpus (CAJA) and in a 1.2-million-word corpus (BASE) lends limited reliability to the related frequency graph. The shortcomings of existing corpora are thus carried into the analysis.

Using existing corpora has been a useful makeshift method when designing this Model, but it should not be used in the actual project. The suggested solution is to enhance CAJA with the currently missing (written and spoken) types of text, and to ensure that the enhanced corpus is balanced and representative of academic language. Existing corpora such as BAWE could still be used to create entry features, such as usage notes on frequent student errors (see 8.1.5).

Existing dictionaries have influenced both the contents and presentation of the sample DOAE entries. Dictionaries acted as a reference for identifying meanings and patterns missed during the analysis. While using dictionaries for this purpose may not be considered to be corpus-driven, it is vital to ensure that the dictionary is comprehensive. Even with excellent analytical skills and intuition, lexicographers will have gaps in their knowledge and experience.

It is therefore useful for them to have references to consult – and existing dictionaries are a very important type of reference.

Similarly, the role of existing dictionaries as a source of ideas for presentation of entry features is essential to the Model. Menus, constructions, frequency graphs, Show More and Show Less buttons are just some of the features that were suggested by existing dictionaries. All of these features are found mainly in learners' dictionaries, which is why the presentation of the proposed Model resembles a learner's dictionary much more than a NS dictionary.

PDEV has been fundamental to meaning analysis by providing a framework for identifying meaning patterns of the words. The approach used in this Model differs somewhat from the CPA approach used by PDEV, as it splits patterns into individual pattern definitions. This is done to assist lexicographers in the writing of definitions, and in the selection of constructions and examples.

The comparison of the sample entry *argue* with the PDEV entry pointed to many similarities in identified patterns, and suggested that some verbs that are very common to academic language may be often used in their academic meanings in general language as well; a fact that is not reflected in existing dictionaries (which are supposed to represent general language).

8.2.5 Analytical software

This section reviews the two pieces of software used in the analysis, Sketch Engine and TshwaneLex.

8.2.5.1 Sketch Engine

Sketch Engine proved immensely useful during the analysis. Its numerous functions made searches for patterns and meanings quite easy, although some computational knowledge is required to conduct the more complex searches. An essential part of the analysis was the Word Sketch function which significantly reduced the time required for the analysis, and made large quantities of corpus data much more manageable. A part of Word Sketch is the TickBox Lexicography function, which was devised especially for lexicographers. By allowing quick and simple export of data to TshwaneLex, TickBox Lexicography represents an extremely useful link between the corpus analysis software (Sketch Engine) and the DOAE database.

Other useful functions in Sketch Engine are Thesaurus, which assisted in identifying synonyms and antonyms, and Sketch Difference, which indicated the differences between two synonyms. An additional benefit of Sketch Engine for the purposes of this Model was that is provided access to several other corpora consulted in the analysis, such as BASE, BAWE and the BNC.

Important advantages of Sketch Engine are its speed and customizability. The searches are extremely fast – the data is produced almost instantly. Consequently, one hardly pays any attention to the tool and could completely focus on the analysis of data. Sketch Engine becomes a subtle companion – which is exactly what a lexicographer wants. The user-friendliness of Sketch Engine is further enhanced by the customizability of almost every function. For example, the number of collocates in the Word Sketch output can be increased. Also, concordance font size can be increased using the browser setting. While this is the quality of the browser, it is the advantage of Sketch Engine to have its menus and output adaptable to such changes.

The analysis indicated some (major) disadvantages of Sketch Engine as well. One disadvantage is technical problems associated with tokenisation and POS-tagging. As discussed in 3.3.1, the tokenisation used by Sketch Engine counts punctuation marks as tokens. This affects not only word count, but also functions like multi-word query, query by context, Sort, Word List, Collocation, and TickBox Lexicography output. Lexicographers need to adapt the searches accordingly; for example, the Sketch Engine phrase query for '*an* IDEA' produces 734 hits in CAJA (Figure 92), which do not include concordance lines with a punctuation mark (in this case a double quotation) between *an* and IDEA, so the solution is to make a query in the Corpus Query Language box, using the following command: [word="an"] [word="\""]{0,1} [word="idea"] – which returns 741 hits in total, i.e. with 7 lines missed by the original search (Figure 93).

Figure 92. Sketch Engine: The first 20 concordance lines for the Phrase query '*an* IDEA'.

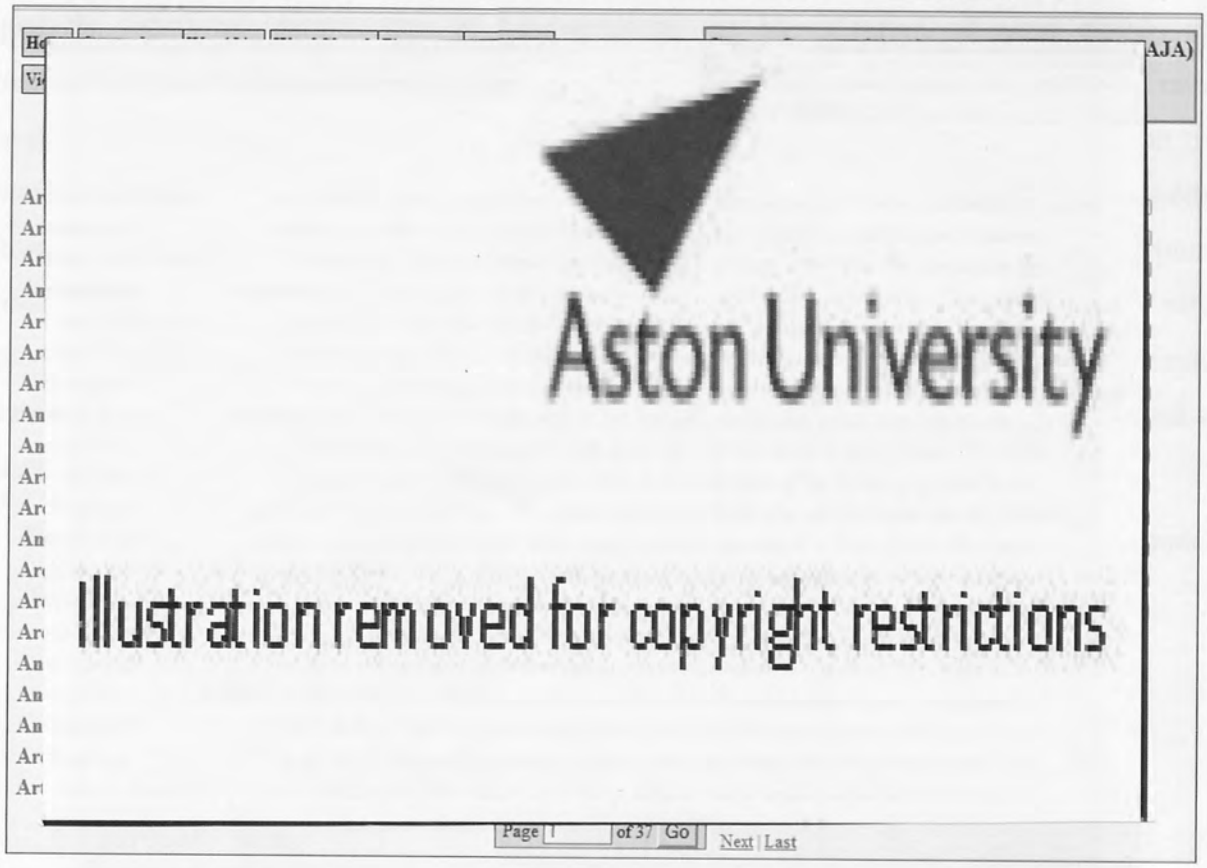
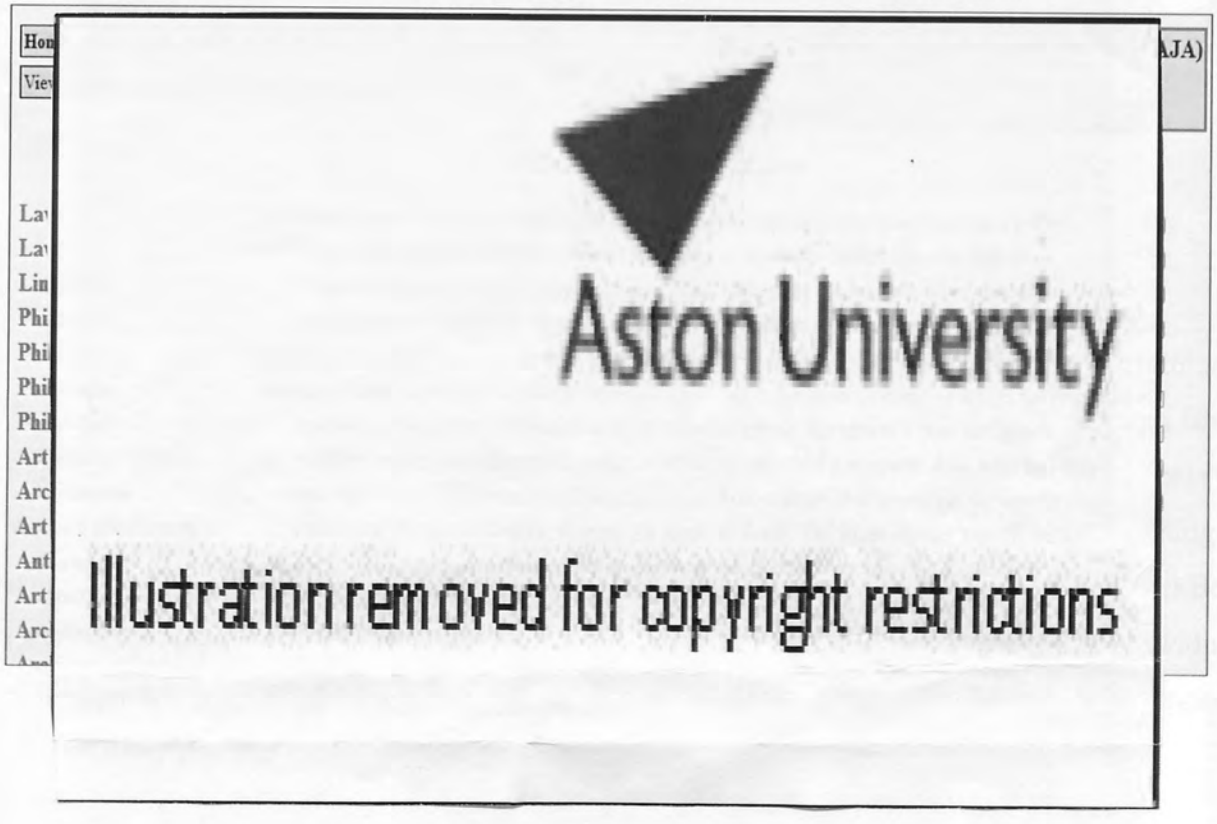


Figure 93. Sketch Engine: Partial concordance for the query *[word="an"] [word="\""]{0,1} [word="idea"]*, sorted by node, showing 7 extra concordance lines.



Tokenisation is closely connected with tagging, and errors in tokenisation can cause errors in lemmatisation, and consequently POS-tagging. Tagging errors affect all stages of the analysis, from the identification of headwords to the identification of word classes and meanings. Some entries can contain a significant percentage of incorrect tags which can result in misleading statistics. As shown by the analysis of four sample entries with more than one word class (Table 97), the most problematic in terms of tagging errors seem to be verbs, especially past participles and 3rd person singular (present), which have the same form as plural nouns and adjectives respectively. The tagging error percentage is also quite high for other present tense verb forms, and adjective forms, which have the same form as the singular noun form of the headword. Lexicographers should therefore not blindly trust the output based on tags (e.g. word sketches) but should approach the information vigilantly and critically.

Table 97. Sketch Engine: Percentage of tagging errors in CAJA for word forms of four sample entries.

	tags used	**percentage of tagging errors (per 100 random concordance lines)**
attributes – verb	VVZ	41%
attributes – noun	NNS	0%
attribute – verb	VV, VVP	10%
attribute – noun	NN	1%
features – verb	VVZ	35%
features – noun	NNS	0%
feature – verb	VV, VVP	19%
feature – noun	NN	0%
justified - verb	VVN	90%
justified - adjective	JJ	2%
potential – adjective	VVZ	10%
potential – noun	NN	2%

The second disadvantage of Sketch Engine is related to its functions. Word Sketch and TickBox Lexicography have some limitations (see 6.3.2.3), and the GDEX heuristics has been found to be unsuitable for academic language (see 6.3.3.4.2.2). The suggested improvements to Word Sketch include offering access to domain distribution of collocates without opening a separate window, and accounting for text- or domain-specificity of collocates. TickBox Lexicography could be improved by offering more support when exporting data, and providing more options when selecting examples (e.g. ability to select preceding and following sentences).

GDEX could be made more useful by adapting its heuristics to academic language. Improving GDEX would also benefit other features such as TickBox Lexicography.

Another disadvantage of Sketch Engine is missing features that could benefit the analysis. One such a feature is N-grams which could be used to identify multi-word items of the headword. At the moment, only two parts of the multi-word item can be identified at a time by using Word Sketch or Collocation. The Collocation function is also missing some potentially useful features, such as the positional distribution of the collocates.

Finally, there is a lack of more direct access to statistical information that is relevant to lexicographers. This problem was often encountered during the design of this Model. The frequency distribution of word classes was needed when recording basic information. Also, the percentage of a word's occurrences found in each grammatical relation would save time when recording the Word Sketch data.

The aforementioned disadvantages, apart from the tagging-related ones, can probably be easily resolved. It should not be forgotten that Sketch Engine is serving the needs of many users, and the authors do not know the requirements of each individual. To be able to customise the tool to the needs of the project, it is important to know the requirements of the lexicographers working on the project, and designing models like the one presented in this thesis is the perfect way to identify those requirements. Yet, even with existing shortcomings related to this Model, Sketch Engine is still very much suited to lexicographic purposes. At the moment, there is currently no better corpus analysis tool known to this author.

8.2.5.2 TshwaneLex

TshwaneLex represents the other half of the software used in this thesis, namely the software used for storing dictionary information, and preparing dictionary output(s). As TshwaneLex is one of the dictionary-writing systems widely used by publishers, its use has helped to simulate the working environment of a modern lexicographer.

The useful features of TshwaneLex have already been presented in detail in 3.3.2. While these features were helpful during the analysis and entry design, three of them were found particularly valuable to the design of this Model: the customisable DTD, the customisable style sets, and the option to import data from Sketch Engine. The option to customise the DTD of the Model (without the need for high-level computational skills) during the creation of the sample entries has made the design of the Model a dynamic process – ideas were tested immediately and implemented if found useful. Similarly, style sets allowed extensive testing of different

outputs, but their main value for the Model was in the option of creating a different style set for each type of user. The ability to import data from Sketch Engine is as important, if not more, as the other two features because it saves a great deal of time. The time-saving value of importing grammatical relations, collocates and examples directly into the database became even more obvious when examples had to be saved manually (using copy and paste in Sketch Engine).

Nonetheless, TshwaneLex has a few disadvantages. The most noticeable disadvantage encountered during the analysis was the slowness of the software when opening longer entries, or when moving between different parts of longer entries. The software does not seem to be suited for handling large amounts of data in an entry, which is somewhat surprising considering that the software is supposed to handle entire dictionary databases. The problem of speed may be connected to the fact that TshwaneLex was mainly designed for bilingual dictionaries[140] where the entries contain (much) less information[141].

Another shortcoming, which also affects the speed of accessing information, is the occasional lack of intuitiveness. For example, the selection of another entry takes the user to the beginning of the entry with all the subtrees open. This caused great problems when importing data as imported data is initially saved as a newly-created entry, and needs to be then copied into an existing entry with the same name. It would be much more useful if the software remembered the exact position of the last consultation of the entry.

Also slightly problematic is the Help function in the software. When facing a problem, the user is always referred to either the User Guide, or the software website. The search for relevant information can be time-consuming and frustrating, especially when consulting the 103-page User Guide (even if the Find function in Adobe Reader is used). It needs to be said, however, that continuous use of the software significantly reduces the need for consulting Help.

In summary, TshwaneLex has been a very useful tool in designing this Model due to its customisability and functionality. In addition, the use of the software has resulted in the acquisition of new computational and lexicographic skills by the author. There are concerns that TshwaneLex may not be well-suited to handling large databases, indicated by the software's slow performance when working with long entries and when switching between entries. One of

[140] The focus on bilingual dictionaries is evident from the examples provided in the instructions document and on the software website, and even in the template DTD (a translation element is among the elements offered).
[141] The computer on which the analysis was conducted was discarded as the potential cause of the slowness of TshwaneLex since it had no problems with running other programs that deal with even larger amounts of data than TshwaneLex (e.g. dictionaries installed on the computer, often several running simultaneously; WordSmith Tools – tested with the CAJA corpus).

the potential solutions could be to create an online version of TshwaneLex[142] which could be used by lexicographers working on the DOAE project. The standalone version of TshwaneLex, as exists now, could be used by editors in the preparatory stages (when developing the dictionary DTD based on sample entries), and the final stages of the project (when designing style sets).

8.3 Implications of the research

The research conducted for the purposes of this thesis has many implications, not only for lexicography, but also for other related disciplines such as pedagogy. These implications are discussed in this section.

8.3.1 Implications for lexicography

The main finding of this research for lexicographers is that there is a gap in the dictionary market for a dictionary of academic English. The thesis has also proposed a solution on how to address this gap, and it is now in the hands of lexicographic community to produce the actual dictionary.

Once such a dictionary is produced, it could provide the impetus to a number of dictionaries and other lexicographic products for students, such as thesauri, collocation dictionaries, and bilingual dictionaries (of academic language). In addition, the dictionary Model presented in this thesis could be used as a basis for designing models for dictionaries of other academic languages (e.g. academic German, academic French).

The research has also indicated that most students do not know, or do not care, whether the dictionary they are using is actually targeted at them. Quick access, fast searches and similar features are much more important than the contents. The lexicographic and academic communities should therefore ensure that DOAE, if produced, is sufficiently and appropriately promoted among student population (e.g. universities could promote it by providing a link to the dictionary on their webpages).

Some of the findings of this research are likely to have implications for other online dictionaries. Currently, most online dictionaries are not very different from their print versions. But, as has been shown, the online dictionary format allows the introduction of many new and valuable user-friendly features, especially user-customisability. Dictionary contents no longer

[142] Online tools are normally much faster due to their capacity to store much larger amounts of data (Sketch Engine is a good example).

have to be static – they can be adapted to the individual user. It is therefore time for dictionary makers to start basing their online dictionaries on a 'one user, one dictionary' principle, rather than 'one dictionary suits all'.

The research also has implications for the process of dictionary-making. The electronic dictionary, more specifically the online dictionary, should replace the print dictionary as the point of departure for dictionary design. This is necessary because the online dictionary represents the most complete format; other versions of the dictionary, including a print version, can always be extracted from it.

Another implication of the research is that the role of lexicographers and editors in the dictionary-design process should be given more importance. The creation of this Model has clearly shown that some useful features of the dictionary can be designed only during or after the analysis of data. In addition to conducting the analysis and building entries, lexicographers should be able to influence decisions on how the information is recorded in the database, and later presented to the user. Consequently, a modern lexicographer needs to be equipped with certain computational skills (e.g. the basics of writing a DTD), and needs to be aware of principles of how to present textual information on a computer screen in an efficient and user-friendly manner.

8.3.2 Implications for pedagogy

The proposed DOAE would not only be a lexicographic product but also a pedagogic tool, because it would help students to understand and produce academic language. This would indirectly benefit EAP teachers who would be able to focus teaching or classroom activities on more problematic words and phrases, using the dictionary as an aid. There are, however, other benefits that the dictionary, or rather the dictionary database, would bring to pedagogy.

8.3.2.1 Using the dictionary database as a resource for teaching materials

The database of the proposed dictionary would be a valuable resource for EAP/ESP teachers, due to its comprehensive description of meanings and patterns of words in academic English. The information from the dictionary database could be used to develop teaching materials such as textbooks, workbooks, and other classroom materials.

The findings of this research have also confirmed some of the criticisms directed at academic wordlists, a highly popular approach to teaching academic vocabulary. First of all, it has been shown that words have different meanings in academic language, some general and

some domain-specific, which wordlists cannot possibly account for. Similarly, wordlists do not reflect the fact that phraseology such as collocations and multi-word items represent a significant part of (academic) language. Moreover, because wordlists simply list lemmas, they cannot show any differences between domains in frequency distribution of word classes of the same lemma.

Nonetheless, despite their many shortcomings, wordlists, as opposed to dictionaries, provide a more focused and limited list of words, so they are still likely to be used by teachers. The aim should therefore be to use DOAE as a resource to not only improve wordlists (both general academic wordlists, and discipline-specific wordlists) by identifying the relevant words, but also to enhance them by adding the relevant meanings and phraseologies to each word on the list.

In fact, the dictionary contents could be manipulated to act as a wordlist. Style sets could be developed for or by the teachers, however these 'teacher' style sets would differ from the ones discussed in Chapter 7. The main stress would be on showing only the information relevant for the user (i.e. teacher), particularly senses, phrases, and frequent patterns. Sense ordering would still be customized, but the teachers would also specify which senses they want to see, depending on the level of homogeneity or heterogeneity of their students. The majority of the other entry features (e.g. etymology, frequency graphs, etc.) would probably not be of interest to the teacher so they would not be displayed. Examples would be hidden, but could always be made visible (as with any other feature) if needed.

For example, an EAP teacher of a more heterogeneous group of students in terms of their subject of study would develop a style set focused primarily on general senses and frequent patterns, and on the senses and frequent patterns found across the majority of domains.

On the other hand, an ESP teacher of a group of students of Computing Science would develop a style set focused on senses and patterns in Sciences, Life Sciences, Applied Sciences, and those specific to Computing.

The sample outputs for the entries *authority* and *justify* using the two style sets described above are presented in Figure 94 and Figure 95 below. It is immediately clear that not all the senses of the two entries are offered to neither the EAP teacher (6 out of 9 for *authority*, 4 out of 5 for *justify*) nor the ESP teacher (1 out of 9 for *authority*, 2 out of 5 for *justify*). Also, the entries have much fewer senses that are of interest to the ESP teacher than to the EAP teacher. For example, all but one sense of *authority* are (more) typical of Arts and Humanities, and therefore not offered to the teacher teaching Computing Science students.

Figure 94. DOAE style sets: The entries *authority* and *justify* for an EAP teacher of a heterogeneous group of students

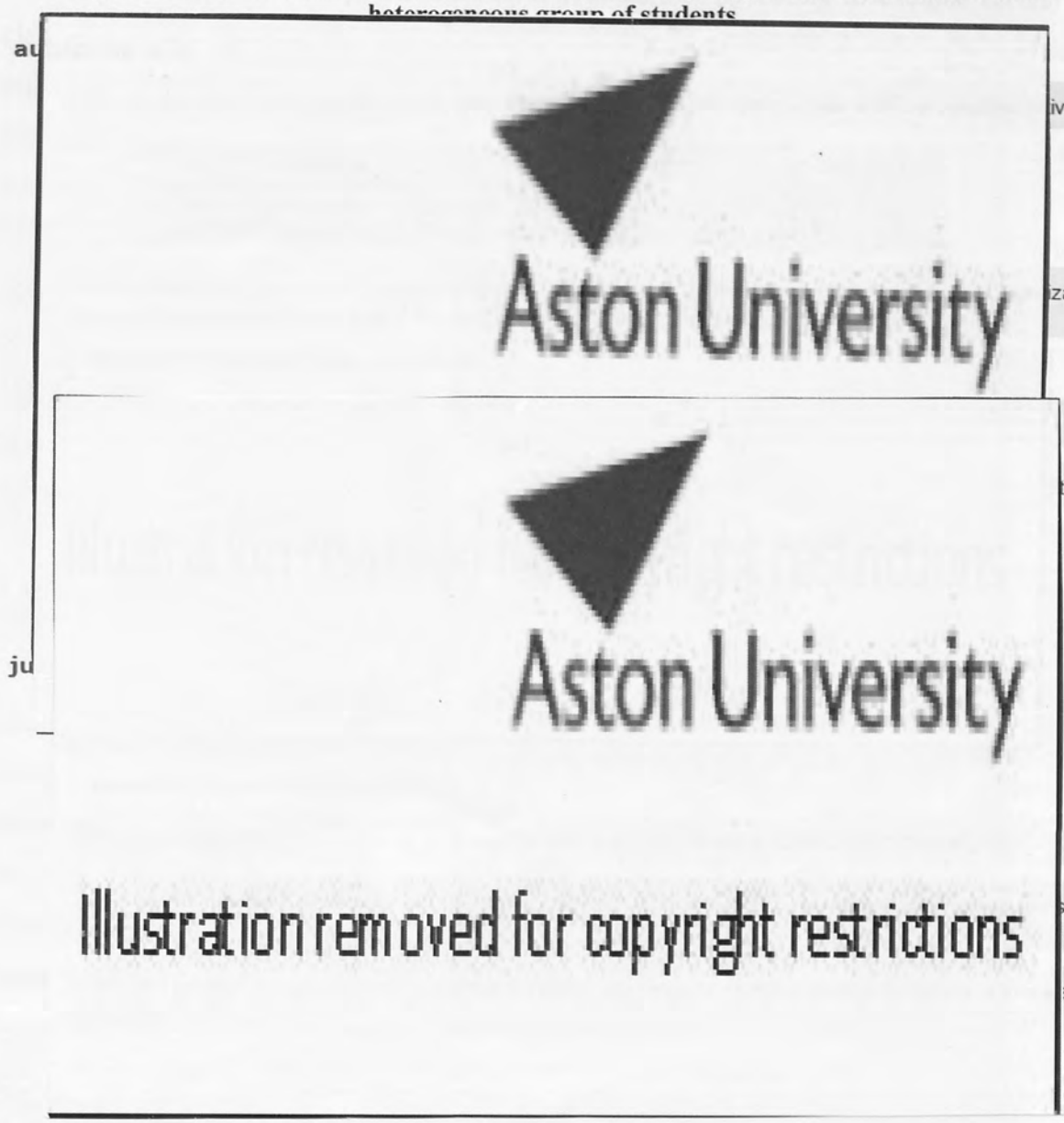
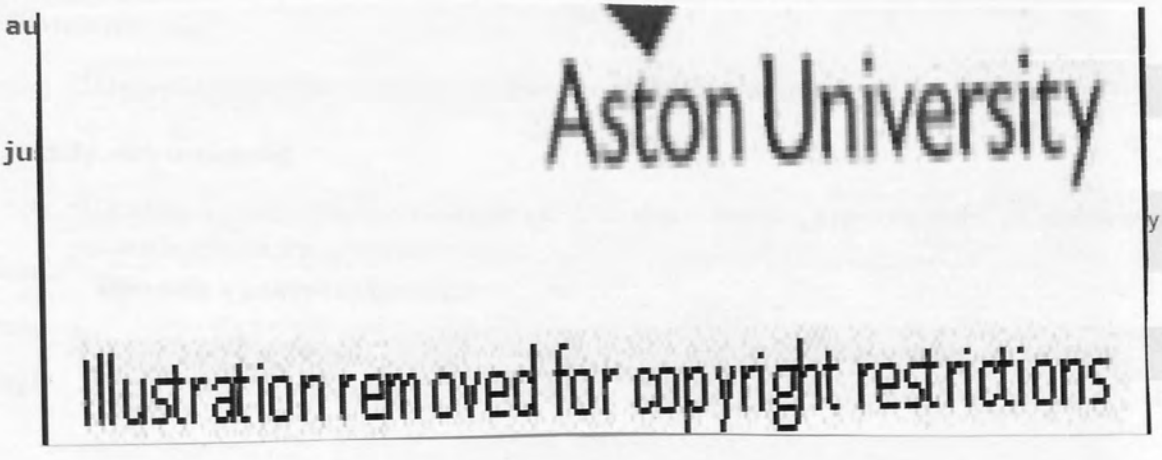


Figure 95. DOAE style sets: The entries *authority* and *justify* for an ESP teacher of a group of Computing Science students.



Teachers could manipulate the output even further by opting to exclude certain word classes, or to exclude words according to their frequency (rank frequency or frequency per million words). For example, a teacher may decide to initially focus on only the top 1,500 most frequent words – in the case of this Model, entries such as *etc.*, *justify*, and *subsequently* would not be part of the output.

Another additional potential of the dictionary database for pedagogy lies in the abundance of collocational information stored under the Meaning analysis element. There, collocates of salient grammatical relations are provided, and labelled if they are domain-specific. Making such information available for extraction and manipulation would be of particular use to ESP teachers, who could search for grammatical relations and/or collocates specific to their subject, and then incorporate this information in their classroom materials.

8.3.2.2 Implications for CALL software

DOAE could be integrated into various types of CALL software. Writing tools, websites offering language help to students, thesauri in word processing programs and similar software could offer links to the dictionary, or have their contents based on the dictionary. For example, the website UEfAP (Using English for Academic Purposes) (Gillett, 2009), which amongst other things contains many exercises related to academic vocabulary and its use, could have the dictionary as an integral part of it so the users accessing the website could consult the dictionary when doing the exercises.

Corpus data used to build DOAE could become a very useful addition to corpus tools (e.g. concordancers), especially because of the scarcity of well-balanced and accessible corpora of academic language. The greater availability of authentic data could then lead to the production of more corpus-based materials, and encourage teachers to use corpora more frequently in the classroom.

8.3.2.3 Wider pedagogic implications of the dictionary

It has been established that there are differences between the meanings and phraseologies of words in academic language and in general language. Many academic meanings of words are thus likely to be new to first year students, whether they are NSs of English or not. Considering that for most students academic language is a tool and not a subject

of study, it is helpful if they are familiar with general academic words and their meanings before the beginning of their study.

Better student proficiency in academic language at the beginning of the study could be achieved by increasing the amount of contents on academic language in curricula at a pre-university level. School-final teaching materials could be informed by DOAE and related EAP/ESP pedagogic materials.

8.4 Recommendations for further research

According to some of the findings of this thesis, certain recommendations regarding future research can be made.

8.4.1 Language problems of different types of student

It has been shown in this thesis that DOAE can be tailored to the needs of different types of student. Nonetheless, tailoring features such as entry order, sense order, and frequent patterns involves merely rearranging or omitting information. Currently missing are features that would provide additional helpful information for a particular type of student.

One characteristic that creates differences among students in terms of their academic language proficiency is native language. It would be therefore useful if the dictionary were to contain language-specific usage notes on frequent errors. This calls for a study of the language problems of students from different language backgrounds. The findings of studies using ICLE could be of some use, but it has been pointed out in 2.4.2 that ICLE is not very representative of student academic writing.

The study suggested here would need to include NSs of English who are currently underrepresented in the research into academic English. Many still do not consider NSs as learners of academic language (including the developers of ICLE), so there is a scarcity of studies into the language problems of this type of student.

Another characteristic that contributes to differentiation among students in terms of their language difficulties is the subject of study. Students can experience language problems due to the domain-specific characteristics of academic language, and a dictionary feature that would contain useful information on addressing such problems would be a very valuable addition.

8.4.2 Studies into the dictionary use of students

It is important to continue conducting studies into the dictionary use of students because, as discussed in 2.3, there are still many unknowns about how students actually use dictionaries. This is especially the case with the dictionary habits of NSs of English, a group of students often overlooked by studies.

Other than investigating which dictionary features are used, how they are used, and how successfully, researchers should also aim to obtain more insight into what motivates the students to pick a particular dictionary. Getting answers to such questions is important to ensure that the dictionaries devised for students are actually used by students; for example, as demonstrated by the main survey (see 4.1.1), many students do not select their dictionary based on contents but on features such as availability and the name of the publisher.

Further studies are also needed on students' dictionary habits specific to different dictionary formats. There is plenty of research into the use of print dictionaries, and also electronic dictionaries such as CD-ROM, but not a lot is known about how students use PEDs or online dictionaries.

Online dictionaries have presented researchers with an opportunity to use a new method of researching dictionary use, namely by logging files of students' searches. By analysing log files, researchers can for the first time monitor and analyse dictionary use in natural conditions – when students are using the dictionaries in private, and for their own purposes. This highly unobtrusive method of monitoring dictionary use can reveal crucial patterns and issues in dictionary use, which existing methods have not been able to identify. But since log files are the property of the dictionary publishers, it is essential that the publishers make them available to researchers. Furthermore, in order for researchers to be able to analyse dictionary habits of individual types of student, the students would have to log in before using the dictionary, and provide certain personal details (which is suggested in the proposed Model; see 7.3).

8.4.3 Research into online dictionaries

One of the main goals of modern lexicography should be a detailed study of the online dictionary format and its potential for modern dictionaries. This goal goes hand in hand with studying dictionary use (mentioned in the previous section), but there is one thing that must be borne in mind here: many existing online dictionaries are merely print dictionaries offered

online, so studies of their use will never identify all the possibilities, and problems, of the online format.

Thus, in order to explore and exploit the full potential of the online format, the existing practice of simply transferring dictionary contents online is not sufficient. Online dictionaries should be designed from scratch. As evidenced by this Model, this practice can result in features such as customisable sense ordering, which are unlikely to be suggested by dictionary users.

The creators of online dictionaries will also need to redefine the collaboration between lexicographers and editors, and computer experts (e.g. software companies), which is currently sequential and autonomous (Atkins & Rundell, 2008). On the one hand, lexicographers and editors should get a bigger say in how dictionary contents are presented to the user. On the other hand, computer experts, while having less autonomy in creating an online version of a dictionary, should be involved in the process of creating a dictionary from the start in order to discuss the ideas of lexicographers andeditors, and suggest their own.

9. CONCLUSIONS

University students encounter difficulties with academic English because of its vocabulary, phraseology, and variability, and also because academic English differs in many respects from general English, the language which they have experienced before starting their university studies. So a dictionary is a welcome aid for students when dealing with any language-related problems, not least because they use language for functional purposes and want to focus on the content of their subject of study.

This research has identified a significant gap in the dictionary market: the lack of a suitable dictionary for students, i.e. a dictionary that focuses on academic English. Many existing dictionaries claim to fill this gap, for example so-called 'dictionaries for students', large NS dictionaries, and advanced learners' dictionaries, but it has been clearly shown that, based as they are on general corpora, these dictionaries are by no means representative of academic English. Moreover, some of these dictionaries have some rather student-unfriendly features (e.g. complex definitions, very few (short) examples), often because they attribute too high a level of language proficiency to the students.

Not only publishers, but also researchers, have overlooked students as a group of dictionary users with their own needs. This is evidenced by the lack of research into dictionary habits of students; studies by Béjoint (1981), Hartmann (1999), and Nesi and Haill, (2002) are rare exceptions that shed some light on which dictionaries students use, and how they use them. However, the rapid progress of technology in recent years, which has had an effect on how dictionaries are used, has already made many of the findings of those studies outdated.

This thesis has made an important first step in attempting to fill this gap in the dictionary market by designing a Model for a Dictionary of Academic English. The proposed Model has delineated each stage of dictionary compilation, from the user profile to the dictionary output(s), and has even suggested how the proposed dictionary could be published. The lack of existing research into student dictionary use and the unsuitability of existing resources (i.e. dictionaries and corpora) meant that the data for each stage of research had to be collected specifically for this Model. This benefited the Model in two ways: a) it made it much more topical and student-informed, and b) it has had many implications for other fields such as pedagogy and corpus linguistics.

The user profile based on the student survey has had a big impact on the design of the Model. First and foremost, it has revealed the popularity of the online dictionary format among students, especially when doing academic work, which has dictated the selection of this format for the Model. Also, the user profile has shown that different types of student (e.g. in terms of native language and/or subject of study) use dictionaries differently, which has resulted in the creation of student-tailored outputs or style sets. Moreover, the survey has confirmed the findings of some of the past studies about the activities that dictionaries are used for, and the microstructural features most frequently consulted, findings which were then considered when designing different features, and outputs.

The Model also demanded the creation of a completely new corpus of academic language (CAJA), as existing corpora proved to be unsuitable for DOAE. The main advantages of CAJA are its large size and balance. The large size benefited both the lexicographic analysis, making patterns easier to identify, and the selection of dictionary examples. Having access to a large corpus of academic language was also essential for the corpus-driven approach to data analysis. A good corpus balance in terms of domains enabled a detailed domain-labelling of senses, patterns, collocates, etc. in the dictionary database, which was then effectively used to tailor the output according to the needs of different types of student.

The analysis of data and the design of dictionary entries have been considerably helped by using state-of-the-art lexicographic software. The corpus tool Sketch Engine, especially its function Word Sketch, was very useful in identifying meanings and patterns, and TshwaneLex provided great flexibility and customisability in recording the information and designing different outputs.

Although existing dictionaries are of limited use to students, they have proved a useful resource for identifying user-friendly features for the Model. This is especially true of advanced learners' dictionaries. The adopted features include menus, full-sentence definitions, full-sentence examples, presentation of constructions, and frequency graphs. In addition, existing dictionaries were useful for identifying any important missed meanings and patterns.

But it is the dissimilarities between the Model and the existing dictionaries that are the testament to the real value of the proposed dictionary to students. The differences were found in senses, sense order, constructions, collocational patterns, and contents of examples, which yet again affirmed the differences between academic English and general English. Furthermore, the difference is also in the format of the dictionary and the way it is designed – the Model proposes an online dictionary that is designed as an online dictionary from scratch.

If the proposed dictionary breaks new ground as far as meeting the needs of students is concerned, it is even more revolutionary in the way it addresses the needs of different types of student. It presents students with a dynamic dictionary that tailors its contents according to the user's native language, subject of study, variant spelling preferences, and/or visual preferences (e.g. black and white).

The Model undoubtedly presents a vision of a dictionary that is much needed by the students, and would be a welcome addition to the dictionary market. Nonetheless, with all the benefits the Model has to offer, it also has certain shortcomings. These are discussed next.

9.1 Shortcomings of the proposed Model

The shortcomings of the Model can be divided into methodological shortcomings, and shortcomings related to the design and the contents of the proposed dictionary.

One of the methodological shortcomings is the use of only one method of investigating dictionary use, namely a questionnaire, to develop the user profile. The inherent weakness of questionnaires is that they ask subjects to self-report on their dictionary use, which often results in over-reporting due to the subjects' desire to conform (Hatherall, 1984; Nesi, 2000). The Model would therefore benefit from other methods, especially ones that measure dictionary use in more natural conditions. One such method is to use the log files of student online dictionary searches, which would be highly appropriate to the online format proposed by the Model. Nonetheless, even findings based on search log files would have to be interpreted carefully, considering that existing online dictionaries differ from the proposed dictionary in functionality.

Another valid criticism of the user profile and related dictionary features designed for the Model is that they are based on the dictionary habits and preferences of students from a single university. A survey of students from different universities in different countries may produce different findings. Even so, the survey conducted in this thesis still has validity, because a) the students from Aston University are representative of the UK student population (see 4.1.1.1), and b) UK universities occupy an important place in the global academic arena.

CAJA can be criticised for its poor genre representativeness. The corpus contains only academic articles, and thus lacks other genres and discourses that students will encounter during their studies, for example books and book reviews, and academic speech. Moreover, scholars would argue that any corpus that wants to be representative of academic language would have to include genres that specifically target students, such as course packs. Other shortcomings of

the corpus are the lack of spoken academic data, and the lack of learner data, i.e. student academic work.

The decision to compile a corpus consisting of academic articles only was a sensible starting point. The academic research article is highly representative of target academic writing in terms of contents, structure, and length. Research articles contain most, if not all, of the genre families and genres identified in the BAWE corpus (Heuboeck et al., 2008). Books, which are absent from CAJA, are long, topic-specific texts, and these characteristics may result in skewing of the data. Furthermore, books are very atypical of the writing students need to produce.

The inclusion of all of the other types of academic texts was simply beyond the scope and timescale of this thesis. The compilation of CAJA took over a year, so collecting textbooks and other material, especially spoken data and learner data, would have been impossible to do. Learner data would be useful only if analysed for errors, which would have been another time-consuming process.

The rather low number of sample entries which were provided in support of the Model can also be regarded as a shortcoming in the research. More sample entries could have led to additional useful elements for the DTD, and indicated further user-friendly features for the output. However, the sample entries were used not only to simulate lexicographic analysis, but also to establish the procedures for recording information in the database, developing the dictionary DTD, and designing the style sets.

As far as the Model's design and contents of the Model are concerned, some lexicographers may question some of the decisions regarding microstructural features, such as using more than one type of definition, providing a limited amount of grammatical information, and using authentic corpus examples that may contain complex language. These decisions have been explained in Chapter 6 (in 6.3.3 in particular), and the main thing to stress is that they are based on the needs of the students.

The Model did not include all the search options that students would need in order to find the relevant information. This was done intentionally as searches need to be tested on the complete dictionary to fully evaluate their value. Also, writing various search options would require the skills of an advanced programmer, something that this researcher was not competent to do, but that would be feasible if the dictionary was going to be produced. Certain aspects of preparing information for different searches were discussed, for example tagging fixed and flexible parts of multi-word phrases.

The flexibility and customisability of the dictionary output has more advantages than disadvantages, but one of the disadvantages that needs to be mentioned is the fact that many of these user-friendly features are possible only if students provide certain personal data, preferably by creating an online user profile. Still, in view of the fact that the required personal information is not sensitive in nature and that a profile has become a regular part of many websites, this should not be a serious problem.

A more contentious issue may be whether the format of the Model makes the dictionary an unrealistic proposition, as an online dictionary may not be very appealing to dictionary publishers. But as has been suggested in 8.1.2, the involvement of publishers may not even be needed – the proposed dictionary could become a joint effort by the academic community.

Possibly the main weakness of the Model is that it has not been tested on students. This important stage of dictionary-making could enhance the user-friendliness of the Model. However, testing on users can be done properly only once the entire dictionary is complete; it would be difficult to simulate actual dictionary use with sample entries alone. Improvements can be expected; after all, the Model proposes a living dictionary – the dictionary whose design and contents are constantly shaped by the users.

9.2 The latest development in EAP lexicography – the Louvain EAP dictionary

After the Model had already been designed, and while the final chapters of the thesis were being written, a new dictionary project emerged that may partly fill the gap in the dictionary market identified in this thesis. The project, called the Louvain EAP Dictionary (LEAD), is currently under way at the Centre for English Corpus Linguistics at the Université catholique de Louvain in Belgium (Granger & Paquot, 2010).

LEAD will share certain characteristics with the Model proposed in this thesis, for example online format and customisability of contents in terms of student's subject of study and/or native language background. Furthermore, LEAD will include some of the proposed enhancements of the Model, such as bilingual equivalents and notes of frequent native-language-specific errors (see 8.1.5). Error notes, as well as frequency information and stylistic notes, will be based on the data provided by a new learner corpus called VESPA (Varieties of English for Specific Purposes dAtabase)[143]. Other useful features of LEAD include

[143] The VESPA corpus will contain non-native speaker texts from "a wide range of disciplines... genres... and degrees of writer expertise in academic settings" (Granger & Paquot, 2010:84). More information is available at http://www.uclouvain.be/en-258647.html.

the option to be used as a semasiological or an onomasiological dictionary, and access to discipline-specific corpora.

However, the methodology that the authors are using to compile LEAD indicates that the dictionary will not represent a complete dictionary solution for students. First of all, LEAD's coverage will be limited as its headword list will be based on Paquot's (2007) production-oriented wordlist of 836 lemmas (see 2.1.2.1, page 44, for more). Also, much weight is attached to the learner corpus used in building LEAD (the VESPA corpus), but very little is known about the corpus of target texts that the dictionary will be based on.

Secondly, LEAD will target NNSs of English only. Furthermore, features such as error notes will not be available to all target users, but only to users with the native language background represented in the VESPA corpus (currently, there are only 5 – French, Norwegian, Polish, Spanish, and Swedish).

Thirdly, while LEAD "takes into account the most recent findings in genre analysis, language teaching and second language acquisition" (Granger & Paquot, 2010:83), it seems to make little use of studies on the dictionary use of students. For example, dictionary customisability is focused on examples and collocations, although these features are consulted by students (much) less frequently than definitions and synonyms (see 2.3.2.1 and 4.1.1.7). Similarly, the poor habits of users (such as the 'choose the first definition' strategy) and their problems in finding the relevant sense are not addressed.

Customising the selection of dictionary examples by discipline may be problematic because subject-specific examples may not always be good dictionary examples, i.e. may not be very user-friendly. For instance, here are two dictionary examples of *issue* (for medicine and business students) provided by Granger and Paquot:

> Medicine example
> *The* ***issue*** *of underreporting of chronic obstructive pulmonary disease (COPD) exacerbations has been addressed by a series of articles, all of which are based on a cohort of patients with COPD living in East London.*
>
> Business example
> *A central* ***issue*** *in the debate is whether foreign direct investment produces positive or negative effects of technology spillovers on domestic firms in host countries.*
>
> (Granger & Paquot, 2010:84)

Both examples are quite long, and contain complex noun phrases (e.g. *chronic obstructive pulmonary disease*), which may cause problems to students. In addition, students

may not be familiar with some of the terminology, even if they are studying a subject in that particular subject. As the LEAD coverage will be limited, the students may not be able to find explanations of precisely the words and phrases that they did not understand. In this particular pair of examples, there is also a question of whether long examples detract from their purpose, which is supposedly to point out the frequent patterns '***the issue of*** something' and '***a central issue in*** something' in medicine and business respectively.

Some of LEAD's features, for example the focus on academic vocabulary and features tailored to student characteristics, do demonstrate some awareness of the needs of students. Online format and customisability show a readiness to implement state-of-the-art lexicography. Nonetheless, with its limited coverage and NNS target audience, LEAD is much more similar to an advanced learners' dictionary than to a comprehensive dictionary of academic English, such as is proposed in this thesis.

9.3 Where does lexicography go from here?

The Model designed in this thesis introduces some very innovative ideas, which are a reflection of the effect that technological progress has had on how dictionaries are used nowadays. Some implications of the Model for lexicography have already been discussed, but it is now time to consider what the future of lexicography holds, regardless of whether the dictionary proposed by the Model gets published or not.

One thing that lexicography will not be able to ignore is the potential, and the popularity, of the online dictionary format. More and more dictionaries will be offered online, so eLexicography (short for electronic lexicography), a new strand in lexicography, will be given an increasingly prominent role. The first signs are already evident – in 2009, the first conference focusing solely on eLexicography (*eLEX2009 – eLexicography in the 21st century*) was held in Belgium.

The online format will become the norm of modern dictionaries. Lexicographers will need to adapt to all the possibilities that this format brings; new ways of presenting and customising dictionary contents will have to be explored. The focus will shift from creating dictionaries for large groups of users to creating dictionaries for individual users (see papers in Bergenholtz, 2009 for more discussion).

Future lexicographers will therefore need to have many different skills in their repertoire, in addition to their ability to analyse data and define meanings. But where will publishers find such lexicographers? Training lexicographers in-house will be expensive, and probably counter-

productive as it may prevent lexicographers from coming up with original ideas. It is therefore envisaged that this situation will revitalise lexicography as a university subject. Language, corpus linguistics and computational topics that would be taught on such a course, combined with work experience at a publishing house, would equip a new generation of lexicographers with the knowledge and skills, useful not only for writing dictionaries, but also for language teaching, developing CALL software, and webpage design.

The increasing domination of the online format will also bring changes to how dictionary use is taught. Currently, print dictionaries still dominate in classrooms, which may be one of the reasons why instruction on dictionary use is almost non-existent among students or teachers. Using online dictionaries for teaching dictionary use is likely to motivate students more, considering the popularity of computers among students.

Despite this shift in trend to online dictionaries, it is not likely that print dictionaries will suddenly become obsolete. The survey has shown that many students still prefer print dictionaries, albeit when using them for non-academic purposes. Many users from older generations will also probably continue to prefer print dictionaries due to their relative lack of familiarity with computer technology. Finally, it should not be forgotten that some dictionary users do not have computer access at all.

The design of the Model has resulted in much more than a suggestion on how to build a dictionary that would meet the needs of university students. The survey conducted for the purposes of the Model has shed more light on how students use dictionaries. Information on the dictionary use of NS students is particularly relevant considering the lack of research in this area. The design of the corpus of academic articles has lead to a new classification of domains. The selection of the online format has revealed new forms of dictionary user-friendliness, and brought to light many new skills that future lexicographers will have to acquire. The implications of the proposed Model are expected to be far-reaching, not only in lexicography, but also in pedagogy and corpus linguistics.

But most importantly, the main aim of the thesis, a model for a corpus-driven Dictionary of Academic English, has been achieved. The aim has actually been exceeded, as the Model proposes not only one but multiple dictionaries, each tailored to a different type of student. Turning the proposed dictionary into a reality is something that should excite lexicographers, and give students something to look forward to.

10. REFERENCES

ABC Amber PDF Converter, version 4.07. http://www.processtext.com/abcpdf.html. ProcessText Group. Downloaded: 24 October 2008.

AHD1: *American Heritage Dictionary, Second College Edition.* (1983). Boston: Houghton Mifflin.

AHD2: *The American Heritage Dictionary of the English Language, 4th edition.* (2000). Boston: Houghton Mifflin.

Altenberg, B. (1998). On the Phraseology of Spoken English: The Evidence of Recurrent Word-Combinations. In Cowie, A.P. (ed.) *Phraseology: Theory, Analysis, and Applications.* Oxford: Oxford University Press. 101-122.

Anderson, S. (2007). What's The Best Font Size To Use In Website Design? (published on 23 December 2007). http://www.hobo-web.co.uk/seo-blog/index.php/best-font-size/. Accessed: 26 January 2010.

Ard, J. (1982). The use of bilingual dictionaries by EFL students while writing. *ITL Review of Applied Linguistics* **58:** 1-27.

Aston Business School (2007). Journal League Tables. http://www.abs.aston.ac.uk/newweb/research/rankings/. Accessed: January 10th 2008.

Aston Business School (2009). Journal League Tables website. http://www.abs.aston.ac.uk/newweb/research/rankings/. Accessed: January 5th 2009.

Aston, G. (1997). Small and Large Corpora in Language Learning. In Lewandowska-Tomaszczyk, B. & Melia, P.J. (eds.) *PALC '97: Practical Applications in Language Corpora.* Łódź: Łódź University Press. 51-62.

Aston University Planning Office (2008a). All students by age group and level of study (Internal Document). www.aston.ac.uk – Accessed January 28th 2009.

Aston University Planning Office (2008b). All students by domicile and level of study (Internal Document). www.aston.ac.uk – Accessed January 28th 2009.

Aston University Planning Office (2008c). All students by gender and level of study (Internal Document). www.aston.ac.uk – Accessed January 28th 2009.

Aston University Planning Office (2008d). Student Numbers by Area of Domicile and Level of Study (Internal Document). www.aston.ac.uk – Accessed March 11th 2007.

Atkins, B.T.S. (2008). Theoretical Lexicography and its relation to dictionary-making. In Fontenelle, T. (ed.) *Practical Lexicography: A Reader.* Oxford: Oxford University Press. 31-50.

Atkins, B.T.S. & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography.* Oxford: Oxford University Press.

Atkins, B.T.S. & Varantola, K. (1998). Language Learners Using Dictionaries: The Final Report on the Euralex/AILA Research Project on Dictionary Use. In Atkins, B.T.S. (ed.) *Using Dictionaries: Studies of Dictionary Use by Language Learners and Translators.* Tübingen: Max Niemeyer Verlag. 21-81.

Axelsson, M.W. (2003). The Uppsala Student English corpus (USE) - Manual. Uppsala University, Department of English. (http://ota.ahds.ac.uk/) Downloaded June 21st 2006.

Ayto, J.R. (1983). On specifying meaning: Semantic analysis and dictionary definitions. In Hartmann, R.R.K. (ed.) *Lexicography: Principles and Practice.* London: Academic Press. 89-98.

Barnbrook, G. (2002). *Defining language: A local grammar of definition sentences.* Amsterdam: John Benjamins.

BASE: *British Academic Spoken English corpus.* http://www2.warwick.ac.uk/fac/soc/celte/research/base. Accessed 2007-2010.

Battenburg, J. (1989). A Study of English Monolingual Learners' Dictionaries and their Users. Purdue University.

Battenburg, J. (1991). *English Monolingual Learners Dictionaries: A User-Oriented Study.* Tübingen: Max Niemeyer.

Baudot, J. & Clas, A. (1984). A Model for a Bilingual Terminology Mini-Bank. *Lebende Sprachen* **2:** 49-54.

Bauer, L. (1993). *Manual of Information to Accompany The Wellington Corpus of Written New Zealand English.* Wellington, New Zealand: Victoria University of Wellington (http://khnt.hit.uib.no/icame/manuals/wellman/index.HTM). Accessed March 23rd 2007.

BAWE: *British Academic Written English corpus.* http://www2.warwick.ac.uk/fac/soc/celte/research/bawe/. Accessed 2007-2010.

Baxter, J. (1980). The dictionary and vocabulary behaviour: a single word or a handful? *TESOL Quarterly* **14**(3)**:** 325-336.

Béjoint, H. (1981). The Foreign Student's Use of Monolingual English Dictionaries: A Study of Language Needs and Reference Skills. *Applied Linguistics* **II**(3)**:** 207-222.

Béjoint, H. (2000). *Modern Lexicography: An Introduction.* Oxford: Oxford University Press.

Benesch, S. (2001). *Critical English for Academic Purposes: Theory, Politics, and Practice.* London: Lawrence Erlbaum Associates.

Benson, M., Benson, E. & Ilson, R. *The BBI Dictionary of English Word Combinations* (1997). John Benjamins BEBC.

Bensoussan, M., Sim, D. & Weiss, R. (1984). The effect of dictionary usage on EFL test performance compared with student and teacher attitudes and expectations. *Reading in a Foreign Language* **2**(2)**:** 262-276.

Bergenholtz, H., Nielsen, S. & Tarp, S. (eds.) (2009). *Lexicography at a Crossroads: Dictionaries and Encyclopedias Today, Lexicographical Tools Tomorrow.* Bern: Peter Lang.

Biber, D. (1988). *Variation across Speech and Writing.* Cambridge: Cambridge University Press.

Biber, D. (2006). *University Language: A corpus-based study of spoken and written registers.* Amsterdam: John Benjamins.

Biber, D. & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes* **26**(3)**:** 263-286.

Biber, D. & Conrad, S. (1999). Lexical bundles in conversation and academic prose. In Hasselgard, H. & Oksefjell, S. (eds.) *Out of Corpora: Studies in Honour of Stig Johansson.* Amsterdam: Rodopi. 181-189.

Biber, D., Conrad, S. & Cortes, V. (2003). Lexical bundles in speech and writing: An initial taxonomy. In Wilson, A., Rayson, P. & McEnery, T. (eds.) *Corpus Linguistics by the Lune: A Festschrift for Geoffrey Leech.* Frankfurt: Peter Lang. 71-92.

Biber, D., Conrad, S. & Cortes, V. (2004a). If you look at...: Lexical bundles in university teaching and textbooks. *Applied Linguistics* **25**(3)**:** 371-405.

Biber, D., Conrad, S., Reppen, R., Byrd, P. & Helt, M. (2002). Speaking and Writing in the University: A Multidimensional Comparison. *TESOL Quarterly* **36**(1)**:** 9-48.

Biber, D., Conrad, S.M., Reppen, R., Byrd, P., Helt, M., Clark, V., Cortes, V., Csomay, E. & Urzua, A. (2004b). *Representing Language Use in the University: Analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus.* Princeton, New Jersey: Educational Testing Service. Available at: http://www.ets.org/Media/Research/pdf/RM-04-03.pdf. Accessed: 19 September 2009.

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999). *The Longman Grammar of Spoken and Written English.* London: Longman.

Bitchener, J. & Basturkmen, H. (2006). Perceptions of the difficulties of postgraduate L2 thesis students writing the discussion section. *Journal of English for Academic Purposes* **5**(1): 4-18.

Black, A. (1986). The effects on comprehension and memory of providing different types of defining information for new vocabulary: a report on two experiments conducted for Longman ELT Dictionaries and Reference Division. Cambridge: MRC Applied Psychology Unit (unpublished internal report).

BNC: *The British National Corpus.* http://www.natcorp.ox.ac.uk/corpus/index.xml. Accessed: 2007-2010.

Bogaards, P. (1996). Dictionaries for Learners of English. *International Journal of Lexicography* **9**(4): 277-320.

Bogaards, P. (1998). What Type of Words do Language Learners Look Up? In Atkins, B.T.S. (ed.) *Using Dictionaries: Studies of Dictionary Use by Language Learners and Translators.* Tübingen: Max Niemeyer Verlag. 151-157.

Bogaards, P. & van der Kloot, W.A. (2001). The Use of Grammatical Information in Learners' Dictionaries. *International Journal of Lexicography* **14**(2): 97-121.

Bogaards, P. & van der Kloot, W.A. (2002). Verb Constructions in Learners' Dictionaries. In Braasch, A. & Poulsen, C. (eds.) *Proceedings of the Tenth EURALEX International Congress.* Copenhagen: Center for Sprogteknologi. 747-757.

Bolinger, D. (1965). The Atomization of Meaning. *Language* **41:** 555-573.

BOS: *Bristol Online Surveys.* http://www.survey.bris.ac.uk/. Accessed: 2007-2009.

Bunton, D. (2005). The structure of PhD conclusion chapters. *Journal of English for Academic Purposes* **4**(3): 207-224.

Cadman, K. (2002). English for Academic Possibilities: the research proposal as a contested site in postgraduate genre pedagogy. *Journal of English for Academic Purposes* **1**(2): 85-104.

Cambridge Academic Content Dictionary. (2009). 1st edition. New York: Cambridge University Press.

Camiciottoli, B.C. (2004). Interactive discourse structuring in L2 guest lectures: some insights from a comparative corpus-based study. *Journal of English for Academic Purposes* **3**(1): 39-54.

Campion, M. & Elley, W. (1971). *An Academic Vocabulary List.* Wellington: New Zealand Council for Educational Research.

Carroll, J.B., Davies, P. & Richman, B. (1971). *The American Heritage Word Frequency Book.* New York: Houghton Mifflin, Boston American Heritage.

CED CD-ROM: *Collins English Dictionary (Desktop Edition), 1st edition.* (2004). Worthing: Harper Collins.

Chambers 21st Century Dictionary, 2nd edition (online). (1999). Edinburgh: Chambers Harrap Publishers Ltd. http://www.chambersharrap.co.uk/chambers/features/chref/chref.py/main.

Charles, M. (2003). 'This mystery...': a corpus-based study of the use of nouns to construct stance in theses from two contrasting disciplines. *Journal of English for Academic Purposes* **2**(4): 313-326.

Chen, Q. & Ge, G.-c. (2007). A corpus-based lexical study on frequency and distribution of Coxhead's AWL word families in medical research articles (RAs). *English for Specific Purposes* **26**(4): 502-514.

Cheng, W., Greaves, C. & Warren, M. (2005). The Creation of a Prosodically Transcribed Intercultural Corpus: The Hong Kong Corpus of Spoken English (prosodic). *ICAME Journal* **29:** 47-68 (http://gandalf.aksis.uib.no/icame/ij29/ij29-page47-68.pdf).

Chung, T.M. & Nation, P. (2003). Technical vocabulary in specialised texts. *Reading in a Foreign Language* **15**(2)**:** 103-116.

Cobb, T. & Horst, M. (2001). Reading academic English: Carrying learners across the academic threshold. In Flowerdew, J. & Peacock, M. (eds.) *Research perspectives on English for academic purposes.* Cambridge: Cambridge University Press. 315-329.

COBUILD1: *Collins Cobuild English Language Dictionary.* (1987). London, Glasgow: Collins.

COBUILD CD-ROM: *Collins Cobuild Dictionary for Advanced Learners, 3rd edition (part of the Collins Cobuild Resource pack).* (2001). Worthing: HarperCollins.

COCA: *Corpus of Contemporary American English.* http://www.americancorpus.org/. Accessed 2007-2010.

CODCE: *Compact Oxford Dictionary of Current English, 3rd edition.* (2005). Oxford: Oxford University Press.

Code Style webpage. http://www.codestyle.org. Accessed: 26 January 2010.

COEDUCS: *Compact Oxford English Dictionary for University and College Students.* (2006). Oxford: Oxford University Press.

Conrad, S. (2001). Variation among disciplinary texts: A comparison of textbooks and journal articles in biology and history. In Conrad, S. & Biber, D. (eds.) *Variation in English: Multi-Dimensional Studies.* London: Longman. 94-107.

Corson, D. (1997). The Learning and Use of Academic English Words. *Language Learning* **47**(4)**:** 671-718.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes* **23**(4)**:** 397-423.

Cowie, A.P. (ed.) (1998). *Phraseology: Theory, Analysis, and Applications.* Oxford: Oxford University Press.

Cowie, A.P. (1999). Phraseology and Corpora: Some Implications for Dictionary-Making. *International Journal of Lexicography* **12**(4)**:** 307-323.

Coxhead, A. (2000). A New Academic Word List. *TESOL Quarterly* **34**(2)**:** 213-238.

Coxhead, A. & Nation, P. (2001). The specialised vocabulary of English for Academic Purposes In Flowerdew, J. & Peacock, M. (eds.) *Research Perspectives on English for Academic Purposes.* Cambridge: Cambridge University Press.

CPA: *Corpus Pattern Analysis*. http://nlp.fi.muni.cz/projekty/cpa/. Masaryk University, Brno. Acessed: 19 December 2008.

Cumming, G., Cropp, S. & Sussex, R. (1994). On-Line Lexical Resources for Language Learners: Assessment of Some Approaches to Word Definition. *System* **22**(3)**:** 369-377.

De Cock, S. (1998). A Recurrent word combination approach to the study of formulae in the speech of native and non-native speakers of English. *International Journal of Corpus Linguistics* **3**(1)**:** 59-80.

De Cock, S. (2006). Getting down to Business: Monolingual Learners' Dictionaries and Business English. In Corino, E., Marello, C. & Onesti, C. (eds.) *Euralex 2006: Proceedings.* Turin: Edizioni dell'Orso. 819-824.

de Schryver, G.-M. (2003). Lexicographers' Dreams in the Electronic-Dictionary Age. *International Journal of Lexicography* **16**(2)**:** 143-199.

de Schryver, G.-M. & Joffe, D. (2004). On How Electronic Dictionaries are Really Used. In Williams, G. & Vessier, S. (eds.) *Proceedings of the Eleventh EURALEX International Congress.* Lorient, France: Universite de Bretagne-Sud. 187-196.

de Schryver, G.-M., Joffe, D., Joffe, P. & Hillewaert, S. (2006). Do Dictionary Users Really Look Up Frequent Words? – On the Overestimation of the Value of Corpus-based Lexicography. *Lexicos* **16:** 67-83.

DeCarrico, J. & Nattinger, J.R. (1988). Lexical phrases for the comprehension of academic lectures. *English for Specific Purposes* **7**(2)**:** 91-102.

Department of Education Science and Training (2005). DEST Selected Higher Education Staff Statistics. Table C.1 Student Characteristics: 1996 - 2003. Available at http://www.universitiesaustralia.edu.au/documents/publications/stats/Students.xls. Accessed: 9. September 2009.

Detagger, version 2.4.0.12. http://www.jafsoft.com/detagger/. JafSoft Limited. Downloaded: 18 October 2008.

DOAJ: Directory of Open Access Journals. http://www.doaj.org/. Accessed: 17 April 2010.

Drury, H. (2001). Short answers in first-year undergraduate science writing. What kind of genres are they? In Hewings, M. (ed.) *Academic Writing in Context : Implications and Applications : Papers in Honour of Tony Dudley-Evans.* Birmingham: University of Birmingham Press. 104-121.

Dudley-Evans, T. & St John, M.J. (1998). *Developments in English for Specific Purposes: A Multi-Disciplinary Approach.* Cambridge: Cambridge University Press.

Durrant, P. (2009). Investigating the viability of a collocation list for students of English for academic purposes. *English for Specific Purposes* **28**(3)**:** 157-169.

Dziemianko, A. (2006). *User-friendliness of Verb Syntax in Pedagogical Dictionaries of English.* Tübingen: Max Niemeyer Verlag.

e-CALD: *Cambridge Advanced Learner's Dictionary, 3rd edition (online).* (2008). Cambridge: Cambridge University Press. http://dictionary.cambridge.org.

e-LDOCE: *Longman Dictionary of Contemporary English, 4th edition (online).* (2006). Harlow: Pearson Longman. http://www.ldoceonline.com.

e-MED: *Macmillan English Dictionary, 2nd edition (online).* (2007). Oxford: Macmillan. http://www.macmillandictionary.com.

e-OALD: *Oxford Advanced Learner's Dictionary, 7th edition (online).* (2005). Oxford: Oxford University Press. http://www.oup.com/elt/catalogue/teachersites/oald7/?cc=global.

Encarta World English Dictionary Online. http://encarta.msn.com/encnet/features/dictionary/dictionaryhome.aspx. Accessed: September 2008.

ERIH (European Reference Index for the Humanities) (2007). Guidelines. http://www.esf.org/research-areas/humanities/research-infrastructures-including-erih/erih-initial-lists.html. ERIH. Accessed: 2006-2009.

ERIH (European Reference Index for the Humanities) (2008). ERIH Expert Panels: process and methodology of selection. http://www.esf.org/research-areas/humanities/research-infrastructures-including-erih/erih-governance-and-panels/erih-expert-panels.html#c23966. ERIH. Accessed: 2006-2009.

FidaPLUS corpus of Slovene. http://www.fidaplus.net/. Accessed: September 2009.

Fillmore, C.J. (1989). Review article of LDOCE2 and COBUILD1. *International Journal of Lexicography* **2**(1)**:** 57-83.

Flowerdew, J. (ed.) (1994). *Academic Listening: Research Perspectives.* New York: Cambridge University Press.

Flowerdew, J. (ed.) (2002a). *Academic Discourse.* Harlow: Pearson.

Flowerdew, J. & Peacock, M. (2001). Issues in EAP: A preliminary perspective. In Flowerdew, J. & Peacock, M. (eds.) *Research Perspectives on English for Academic Purposes.* Cambridge: Cambridge University Press. 8-24.

Flowerdew, L. (2002b). Corpus-Based Analyses in EAP. In Flowerdew, J. (ed.) *Academic Discourse.* Harlow: Pearson. 95-114.
Fox, G. (1987). The Case for Examples. In Sinclair, J. (ed.) *Looking up: An Account of the COBUILD Project in Lexical Computing.* London: Collins. 137-149.
Foxley, E. & Gwei, G.M. (1989). Synonymy and Contextual Disambiguation of Words. *International Journal of Lexicography* **2**(2): 111-134.
FrameNet Project. http://framenet.icsi.berkeley.edu/. Accessed: April 2009.
Francis, W.N. & Kucera, H. (1979). Brown Corpus Manual. http://icame.uib.no/brown/bcm.html. Accessed March 15th 2007.
Fraser, H. (1997). Dictionary Pronunciation Guides for English. *Int J Lexicography* **10**(3): 181-208.
Freddi, M. (2005). Arguing linguistics: corpus investigation of one functional variety of academic discourse. *Journal of English for Academic Purposes* **4**(1): 5-26.
Gabrielatos, C. (2005). Corpora and Language Teaching: Just a fling or wedding bells? *TESL-EJ* **8**(4) http://www-writing.berkeley.edu/TESL-EJ/ej32/a1.html. Accessed: 17 September 2009.
Ghadessy, P. (1979). Frequency Counts, Word Lists, and Materials Preparation: A New Approach. *English Teaching Forum* **17**: 24-27.
Gillett, A. (2009). Using English for Academic Purposes - A Guide for Students in Higher Education. http://www.uefap.com. Accessed: 27 February 2010.
Gledhill, C. (1996). Science as a Collocation: Phraseology in Cancer Research Articles. In Botley, S., Glass, J., McEnery, T. & Wilson, A. (eds.) *Proceedings of Teaching and Language Corpora 1996* (UCREL Technical Papers Volume 9).
Gledhill, C. (2000). The Discourse Function of collocation in Research Article Introductions. *English for Specific Purposes* **19**(2): 115-135.
Graddol, D. (2006). *English Next.* London: British Council.
Granger, S. (ed.) (1998a). *Learner English on computer.* London: Longman.
Granger, S. (1998b). Prefabricated Patterns in Advanced EFL Writing: Collocations and Formulae. In Cowie, A.P. (ed.) *Phraseology: Theory, Analysis, and Applications.* Oxford: Oxford University Press. 145-160.
Granger, S. & Paquot, M. (2010). Customising a general EAP dictionary to meet learner needs. In Granger, S. & Paquot, M. (eds.) *eLexicography in the 21st century: New challenges, new applications. Proceedings of ELEX2009.* Cahiers du CENTAL. Louvain-la-Neuve, Presses universitaires de Louvain. 83-86.
Groom, N. (2005). Pattern and meaning across genres and disciplines: An exploratory study. *Journal of English for Academic Purposes* **4**(3): 257-277.
Hancioglu, N., Neufeld, S. & Eldridge, J. (2008). Through the looking glass and into the land of lexico-grammar. *English for Specific Purposes* **27**(4): 459-479.
Hanks, P. (1987). Definitions and Explanations. In Sinclair, J. (ed.) *Looking up: An Account of the COBUILD Project in Lexical Computing.* London: Collins. 116-136.
Hanks, P. (1994). Linguistic Norms and Pragmatic Exploitations or, Why Lexicographers Need Prototype Theory, and Vice Versa. In Kiefer, F., Kiss, G. & Pajzs, J. (eds.) *Papers in Computational Lexicography: Complex '94.* 89-113.
Hanks, P. (2000). Do Word Meanings Exist. *Computers and the Humanities* **34**(1-2): 205-215.
Hanks, P. (2004). The Syntagmatics of Metaphor and Idiom. *International Journal of Lexicography* **17**(3): 245-274.
Hanks, P. (2005). Johnson and Modern Lexicography. *International Journal of Lexicography* **18**(2): 243-266.

Hanks, P. (2008). Mapping meaning onto use: a Pattern Dictionary of English Verbs. *American Association for Corpus Linguistics (AACL).* Utah: Available at http://nlp.fi.muni.cz/projekty/cpa/Pattern%20Dict%20Utah.ppt. Accessed: November 2008.

Hanks, P. (forthcoming). *Lexical Analysis: Norms and Exploitations.* MIT Press.

Hanks, P. & Ježek, E. (2008). Shimmering Lexical Sets. In Bernal, E. & DeCesaris, J. (eds.) *Proceedings of the Thirteenth EURALEX International Congress.* Barcelona, Spain: Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra. 391-402.

Hartmann, R.R.K. (1987). Four perspectives on dictionary use: A critical review of research methods. In Cowie, A. (ed.) *The Dictionary and the Language Learner.* Tübingen: Niemeyer. 11-28.

Hartmann, R.R.K. (1999). The Exeter University Survey of Dictionary Use. In Hartmann, R.R.K. (ed.) *Dictionaries in Language Learning: Recommendations, National Reports and Thematic Reports from the TNP Sub-Project 9: Dictionaries.* Berlin: Freie Universität. 36-52. Available at http://www.fu-berlin.de/elc/TNPproducts/SP39dossier.doc. [Accessed 15 June 2009].

Harvey, K. & Yuill, D. (1997). A study of the use of a monolingual pedagogical dictionary by learners of English engaged in writing. *Applied Linguistics* **18**(3): 253-278.

Hatherall, G. (1984). Studying dictionary use: some findings and proposals. In Hartmann, R.R.K. (ed.) *LEX'eter '83 Proceedings: Papers from the International Conference on Lexicography at Exeter, 9-12 September 1983.* Tübingen: Niemeyer Verlag. 183-189.

Hausmann, F.J. & Gorbahn, A. (1989). COBUILD and LDOCE II A comparative review. *International Journal of Lexicography* **2**(1): 44-56.

Herbst, T. (1996). On the way to the perfect learners' dictionary: a first comparison of OALD5, LDOCE3, COBUILD2 and CIDE. *International Journal of Lexicography* **9**(4): 321-357.

HESA (2008). All students by institution, mode of study, level of study, gender and domicile (2006/07). http://www.hesa.ac.uk/dox/dataTables/studentsAndQualifiers/download/institution0607.xls?v=1.0. Higher Education Statistics Agency. Date Accessed: January 28th 2008.

HESA (2009). Table 2e - All HE students by level of study, mode of study, subject of study, domicile and gender 2007/08. http://www.hesa.ac.uk/index.php?option=com_datatables&Itemid=121&task=show_category&catdex=3. Higher Education Statistics Agency. Date Accessed: 9. September 2009.

Heuboeck, A., Holmes, J. & Nesi, H. (2008). The BAWE Corpus Manual. http://www.coventry.ac.uk/researchnet/external/content/1/c4/51/60/v1214467490/user/BAWE.documentation.pdf. Accessed: 3 December 2008.

Hewings, A. & Hewings, M. (2001). Anticipatory it' in academic writing: an indicator of disciplinary difference and developing disciplinary knowledge. In Hewings, M. (ed.) *Academic Writing in Context : Implications and Applications : Papers in Honour of Tony Dudley-Evans.* Birmingham: University of Birmingham Press. 199-214.

Hewings, M. & Hewings, A. (2002). "It is interesting to note that...": a comparative study of anticipatory [`]it' in student and published writing. *English for Specific Purposes* **21**(4): 367-383.

Hoey, M. (2005). *Lexical priming.* Oxford: Routledge.

Hollósy, B. (1988). On the Need for a Dictionary of Academic English. In Magay, T. & Zigany, J. (eds.) *BudaLEX'88 Proceedings: Papers from the EURALEX Third International Congress.* Budapest: Akadémiai Kiadó. 535-542.

Howarth, P. (1996). *Phraseology in English Academic Writing: Some Implications for Language Learning and Dictionary Making.* Tübingen: Niemeyer.
Howarth, P. (1998). The Phraseology of Learners' Academic Writing. In Cowie, A.P. (ed.) *Phraseology: Theory, Analysis, and Applications.* Oxford: Oxford University Press. 161–186.
Hunston, S. (1993). Evaluation and ideology in scientific writing. In Ghadessy, M. (ed.) *Register analysis: Theory and practice.* London: Pinter. 57-73.
Hunston, S. (1995). A corpus study of some English verbs of attribution. *Functions of Language* **2:** 133-158.
Hyland, K. (1994). Hedging in academic writing and EAP textbooks. *English for Specific Purposes* **13**(3)**:** 239-256.
Hyland, K. (1996a). Talking to the academy: Forms of hedging in science research articles. *Written Communication* **13:** 251-281.
Hyland, K. (1996b). Writing without conviction? Hedging in science research articles. *Applied Linguistics* **17:** 433-454.
Hyland, K. (1998). *Hedging in Scientific Research Articles.* Philadelphia: John Benjamins.
Hyland, K. (1999). Talking to Students: Metadiscourse in Introductory Coursebooks. *English for Specific Purposes* **18**(1)**:** 3-26.
Hyland, K. (2001). Humble servants of the discipline? Self-mention in research articles. *English for Specific Purposes* **20**(3)**:** 207-226.
Hyland, K. (2002). Activity and Evaluation: Reporting practices in academic writing. In Flowerdew, J. (ed.) *Academic Discourse.* Harlow: Pearson. 115-130.
Hyland, K. (2006). *English for Academic Purposes: An Advanced Resource Book.* London: Routledge.
Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes* **27**(1)**:** 4-21.
Hyland, K. & Tse, P. (2007). Is There an "Academic Vocabulary"? *TESOL Quarterly* **41:** 235-253.
ICLE: *International Corpus of Learner English.* http://cecl.fltr.ucl.ac.be/Cecl-Projects/Icle/icle.htm. Accessed June 25th 2007.
Ilson, R. (1987). Illustrations in dictionaries. In Cowie, A.P. (ed.) *The Dictionary and the Language Learner.* Tübingen: Max Niemeyer Verlag. 193-212.
InfoRapid Search & Replace, version 3.1f. http://www.inforapid.de/html/srdownload.htm. JafSoft Limited. Downloaded: 10 November 2008.
ISI Web of Knowledge, Journal Citation Reports. www.isiknowledge.com. Last accessed: January 2008.
Jackson, H. (1988). *Words and Their Meaning.* London: Longman.
JDEST: Huizhong, Y. (1986). A New Technique for Identifying Scientific/Technical Terms and Describing Science Texts: (An Interim Report). *Literary and Linguistic Computing* **1**(2)**:** 93-103.
Joffe, D. & de Schryver, G.M. (2004). TshwaneLex - A State-of-the-Art Dictionary Compilation Program. In Williams, G. & Vessier, S. (eds.) *Proceedings of the Eleventh EURALEX International Congress.* Lorient, France: Universite de Bretagne-Sud. 99-104.
Joffe, D., MacLeod, M. & de Schryver, G.M. (2008). Software Demonstration: The TshwaneLex Electronic Dictionary System. In Bernal, E. & DeCesaris, J. (eds.) *Proceedings of the Thirteenth EURALEX International Congress.* Barcelona, Spain: Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra. 421-424.
Johansson, S. (1978). *Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English, for Use with Digital Computers.* Oslo: University of Oslo,

Department of English (http://khnt.hit.uib.no/icame/manuals/lob/INDEX.HTM). Accessed March 23rd 2007.
Jordan, R.R. (1997). *English for Academic Purposes: A Guide and Resource Book for Teachers.* Cambridge: Cambridge University Press.
Keselj, V. & Keselj, T. (2006). What do users want from an on-line dictionary: A seven-years usage study of an e-dictionary. *Paper presented at 9th INTEX/NooJ Conference.* http://nooj.matf.bg.ac.yu/pptpdf/06%20Vlado%20Keselj%20-%20nooj.pdf. Accessed: 30 September 2009.
Kilgarriff, A. (2000). Business Models for Dictionaries and NLP. *International Journal of Lexicography* **13**(2): 107-118.
Kilgarriff, A., Husák, M., McAdam, K., Rundell, M. & Rychly, P. (2008). GDEX: Automatically Finding Good Dictionary Examples in a Corpus. In Bernal, E. & DeCesaris, J. (eds.) *Proceedings of the Thirteenth EURALEX International Congress.* Barcelona, Spain: Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra. 425-432.
Kilgarriff, A. & Rychly, P. (2008). Finding the Words Which Are Most X. In Bernal, E. & DeCesaris, J. (eds.) *Proceedings of the Thirteenth EURALEX International Congress.* Barcelona, Spain: Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra. 433-436.
Kilgarriff, A., Rychly, P., Smrz, P. & Tugwell, D. (2004). The Sketch Engine. In Williams, G. & Vessier, S. (eds.) *Proceedings of the Eleventh EURALEX International Congress.* Lorient, France: Universite de Bretagne-Sud. 105-116.
Kirkness, A. (2006). Lexicography. In Davies, A. & Elder, C. (eds.) *The Handbook of Applied Linguistics.* Oxford: Blackwell Publishing. 54-81.
Klotz, M. (2003). Oxford Collocations Dictionary for Students of English. *International Journal of Lexicography* **16**(1): 57-61.
Knowles, F.E. (1988). The Role of Dictionaries in Developing Writing Skills. In Ager, D. (ed.) *Written Skills in the Modern Languages Degree.* Birmingham: AMLC. 95-109.
Koren, S. (1997). Quality versus Convenience: Comparison of Modern Dictionaries from the Researcher's, Teacher's and Learner's Points of View. *TESL-EJ* **2**(3).
Koutsantoni, D. (2004). Attitude, certainty and allusions to common knowledge in scientific research articles. *Journal of English for Academic Purposes* **3**(2): 163-182.
Krek, S. (ed.) (2005). *Veliki angleško-slovenski slovar Oxford-DZS (Oxford-DZS Comprehensive English-Slovene Dictionary), Volume 1 (A-K).* Ljubljana: DZS.
Krek, S. (ed.) (2006). *Veliki angleško-slovenski slovar Oxford-DZS (Oxford-DZS Comprehensive English-Slovene Dictionary), Volume 2 (L-Z).* Ljubljana: DZS.
Krishnamurthy, R. (1987). The Process of Compilation. In Sinclair, J. (ed.) *Looking up: An Account of the COBUILD Project in Lexical Computing.* London: Collins. 62-85.
Krishnamurthy, R. (2001). Corpus direct to Your Classroom. *Independence: The Newsletter of the IATEFL Learner Independence Special Interest Group* (29): 10-14.
Krishnamurthy, R. (2008). Corpus-driven Lexicography. *Int J Lexicography* **21**(3): 231-242.
Krishnamurthy, R. & Kosem, I. (2007). Issues in creating a corpus for EAP pedagogy and research. *Journal of English for Academic Purposes* **6**(4): 356-373.
Landau, S. (2001). *Dictionaries: the Art and Craft of lexicography.* Cambridge: Cambridge University Press.
Landau, S.I. (1999). The New Oxford Dictionary of English (Review). *International Journal of Lexicography* **12**(3): 250-257.
LDOCE1: *Longman Dictionary of Contemporary English, 1st edition.* (1978). Harlow: Longman.

LDOCE2: *Longman Dictionary of Contemporary English, 2nd edition.* (1987). Harlow: Longman.
Lea, D. *Oxford Collocations Dictionary for Students of English.* (2002). Oxford: Oxford University Press.
LED and LED CD-ROM: *Longman Exams Dictionary.* (2006). Harlow: Longman.
Lee, D. (2001). Genres, Registers, Text Types, Domains, and Styles: Clarifying the Concepts and Navigating a Path through the BNC Jungle. *Language Learning & Technology* **5**(3): 37-72.
Leńko-Szymańska, A. (2008). Formulaic sequences in apprentice writing - does more mean better? In Frankenberg-Garcia, A. (ed.) *Proceedings of the 8th Teaching and Language Corpora Conference.* Lisbon: ISLA. 207-212.
Lew, R. (2002). Questionnaires in Dictionary Use Research: A Reexamination. In Braasch, A. & Poulsen, C. (eds.) *Proceedings of the Tenth EURALEX International Congress.* Copenhagen: Center for Sprogteknologi. 267-271.
Lew, R. (2009). Towards Variable Function-Dependent Sense Ordering in Future Dictionaries. In Bergenholtz, H., Nielsen, S. & Tarp, S. (eds.) *Lexicography at a Crossroads: Dictionaries and Encyclopedias Today, Lexicographical Tools Tomorrow.* Bern: Peter Lang. 237-264.
Lew, R. & Dziemianko, A. (2006). A New Type of Folk-inspired Definition in English Monolingual Learners' Dictionaries and its Usefulness for Conveying Syntactic Information. *International Journal of Lexicography* **19**(3): 225-242.
Lewin, B.A. (2005). Hedging: an exploratory study of authors' and readers' identification of 'toning down' in scientific texts. *Journal of English for Academic Purposes* **4**(2): 163-178.
Lewis-Beck, M., Bryman, A. & Liao, T.F. (eds.) (2004). *The SAGE Encyclopedia of Social Science Research Methods.* Thousand Oaks, California, USA: SAGE Publications.
Lexical Computing Ltd. (2007). Statistics used in the Sketch Engine. http://trac.sketchengine.co.uk/wiki/SkE/DocsIndex. Accessed: January 8th 2009.
Longman Language Activator. (2002). Harlow: Longman.
Louw, B. (1993). Irony in the Text or Insincerity in the Writer? The Diagnostic Potential of Semantic Prosodies. In Baker, M., Francis, G. & Tognini-Bonelli, E. (eds.) *Text and Technology.* Philadelphia/Amsterdam: John Benjamins.
Lynn, R.W. (1973). Preparing Word Lists: A Suggested Method. *RELC Journal* **4**(1): 25-32.
MacFarquhar, P. & Richards, J. (1983). On dictionaries and definitions. *RELC Journal* **14**(1): 111-124.
Marco, M.J.L. (2000). Collocational frameworks in medical research papers: a genre-based study. *English for Specific Purposes* **19**(1): 63-86.
Martinez, I.A. (2003). Aspects of theme in the method and discussion sections of biology journal articles in English. *Journal of English for Academic Purposes* **2**(2): 103-123.
Martínez, I.A., Beck, S.C. & Panza, C.B. (2009). Academic vocabulary in agriculture research articles: A corpus-based study. *English for Specific Purposes* **28**(3): 183-198.
Mauranen, A. (2003). The Corpus of English as Lingua Franca in Academic Settings. *TESOL Quarterly* **37**(3): 513-527.
McCreary, D. (2008). Looking Up "Hard Words" for a Production Test: A Comparative Study of the NOAD, MEDAL, AHD, and MW Collegiate Dictionaries. In Bernal, E. & DeCesaris, J. (eds.) *Proceedings of the Thirteenth EURALEX International Congress.* Barcelona: Universitat Pompeu Fabra. 1287-1293.
McCreary, D.R. (2002). American Freshmen and English Dictionaries: 'I Had Aspersions of Becoming an English Teacher'. *International Journal of Lexicography* **15**(3): 181-205.

McCreary, D.R. & Amacker, E. (2006). Experimental Research on College Students' Usage of Two Dictionaries: A Comparison of the Merriam-Webster Collegiate Dictionary and the Macmillan English Dictionary for Advanced Learners. In Corino, E., Marello, C. & Onesti, C. (eds.) *Euralex 2006: Proceedings.* Turin: Edizioni dell'Orso. 871-885.
McCreary, D.R. & Dolezal, F.T. (1999). A Study of Dictionary Use by ESL Students in an American University. *International Journal of Lexicography* **12**(2): 107-146.
McEnery, T., Xiao, R. & Tono, Y. (2006). *Corpus-Based Language Studies: an advanced resource book.* Abingdon: Routledge.
Meara, P. & English, F. (1987). Lexical errors and learners' dictionaries. *ERIC Document No. ED290322.*
MEDAL1: *Macmillan English Dictionary for Advanced Learners, 1st edition.* (2002). Oxford: Macmillan.
MEDAL2: *Macmillan English Dictionary for Advanced Learners, 2nd edition.* (2007). Oxford: Macmillan.
MICASE-based publications and presentations (2009). http://lw.lsa.umich.edu/eli/micase/publications.htm. Accessed: 20 September 2009.
MICASE Manual (2003). The Michigan Corpus of Academic Spoken English. http://www.lsa.umich.edu/eli/micase/MICASE_MANUAL.pdf. Accessed March 12-15th 2007.
MICASE: *The Michigan Corpus of Academic Spoken English.* http://quod.lib.umich.edu/m/micase/. Accessed June 2007-April 2010.
Microsoft Excel. (2002). Microsoft Corporation.
MICUSP: *The Michigan Corpus of Upper-level Student Papers.* http://micusp.elicorpora.info/.
Miller, G. & Gildea, P. (1987). How Children Learn Words. *Scientific American* **257**(3): 94-99.
Ming-Tzu, K.W. & Nation, P. (2004). Word Meaning in Academic English: Homography in the Academic Word List. *Applied Linguistics* **25**(3): 291-314.
Mitchell, E. (1983). *Search-Do Reading: Difficulties in Using a Dictionary.* Aberdeen: College of Education.
Moon, R. (1987). The Analysis of Meaning. In Sinclair, J. (ed.) *Looking up: An Account of the COBUILD Project in Lexical Computing.* London: Collins. 86-103.
Moon, R. (1992). 'There is reason in the roasting of eggs': a consideration of fixed expressions in native-speaker dictionaries. In Tommola, H., 1 & 2 (eds.) *Euralex '92 Proceedings I–II.* Tampere, Findland: University of Tampere. 493-502.
Moon, R. (1996). Data, Description, and Idioms in Corpus Lexicography. In Gellerstam, M., 1 & 2 (eds.) *Euralex '96 Proceedings I-II.* Gothenburg: Gothenburg University. 245-256.
Moon, R. (1998a). *Fixed Expressions and Idioms in English: A corpus-based approach* Oxford: Clarendon Press.
Moon, R. (1998b). Frequencies and Forms of Phrasal Lexemes in English. In Cowie, A.P. (ed.) *Phraseology: Theory, Analysis, and Applications.* Oxford: Oxford University Press. 79-100.
Moore, T. & Morton, J. (2005). Dimensions of difference: a comparison of university writing and IELTS writing. *Journal of English for Academic Purposes* **4**(1): 43-66.
Mudraya, O. (2006). Engineering English: A lexical frequency instructional model. *English for Specific Purposes* **25**(2): 235-256.
Murison-Bowie, S. (1993). *MicroConcord Manual.* Oxford: Oxford University Press.
MWCD and MWCD CD-ROM: *Merriam-Webster 11th Collegiate Dictionary.* (2003). Springfield: Merriam-Webster.
Myers, G. (2001). 'In my opinion': the place of personal views in undergraduate essays. In Hewings, M. (ed.) *Academic Writing in Context : Implications and Applications :*

Papers in Honour of Tony Dudley-Evans. Birmingham: University of Birmingham Press. 63-78.
Nation, I.S.P. (2001). *Learning Vocabulary in Another Language.* Cambridge: Cambridge University Press.
Nation, P. (1990). *Teaching and Learning Vocabulary.* Boston: Heinle & Heinle.
Nation, P. & Waring, R. (1997). Vocabulary size, text coverage and word lists. In Schmitt, N. & McCarthy, M. (eds.) *Vocabulary: Description, Acquisition and Pedagogy.* Cambridge: Cambridge University Press. 6-19.
Nattinger, J.R. & DeCarrico, J.S. (1992). *Lexical phrases and language teaching.* Oxford: Oxford University Press.
Nesi, H. (1999). A User's Guide to Electronic Dictionaries for Language Learners. *International Journal of Lexicography* **12**(1): 55-66.
Nesi, H. (2000). *The Use and Abuse of EFL Dictionaries. How learners of English as a foreign language read and interpret dictionary entries.* Tübingen: Max Niemeyer Verlag.
Nesi, H. (2009). A Multidimensional Analysis of Student Writing across Levels and Disciplines. In Edwardes, M. (ed.) *Proceedings of the British Association of Applied Linguistics (BAAL) Conference.* University of Swansea, 11-13 September 2008.
Nesi, H., Gardner, S., Forsyth, R., Hindle, D., Wickens, P., Ebeling, S., Leedham, M., Thompson, P. & Heuboeck, A. (2005). Towards the compilation of a corpus of assessed student writing: An account of work in progress. *Proceedings from the Corpus Linguistics Conference Series.* University of Birmingham, UK: http://www.corpus.bham.ac.uk/PCLC/NesiStudentWriting.doc.
Nesi, H. & Haill, R. (2002). A Study of Dictionary Use by International Students at a British University. *International Journal of Lexicography* **15**(4): 277-305.
Neubach, A. & Cohen, A. (1988). Processing Strategies and Problems Encountered in the Use of Dictionaries. *Dictionaries: Journal of the Dictionary Society of North America* **10**: 1-19.
New Oxford American Dictionary, 2nd edition. (2005). New York: Oxford University Press.
NODE and NODE CD-ROM: *The New Oxford Dictionary of English.* (1998). Oxford: Oxford University Press.
Norman, G. (2002). Description and Prescription in Dictionaries of Scientific Terms. *International Journal of Lexicography* **15**(4): 259-276.
Nth Grep Pro, version 2.3. http://www.nth-generation.com/nthgrep.php. Nth Generation. Downloaded: 17 October 2008.
Nurweni, A. & Read, J. (1999). The English Vocabulary Knowledge of Indonesian University Students. *English for Specific Purposes* **18**(2): 161-175.
Oakes, M. (1998). *Statistics for Corpus Linguistics.* Edinburgh: Edinburgh University Press.
Oakey, D.J. (2002). Formulaic language in English academic writing: A corpus-based study of the formal and functional variation of a lexical phrase in different academic disciplines. In Reppen, R., Fitzmaurice, S. & Biber, D. (eds.) *Using Corpora to Explore Linguistic Variation.* Philadelphia: John Benjamins. 111-129.
Oakey, D.J. (2005). Academic vocabulary in academic discourse: The phraseological behaviour of EVALUATION in Economics research articles. In Tognini-Bonelli, E. & Del Lungo Camiciotti, G. (eds.) *Strategies in Academic Discourse.* Amsterdam: John Benjamins. 169-183.
OALD3: *Oxford Advanced Learner's Dictionary of Current English, 3rd edition.* (1974). Oxford: Oxford University Press.
OALD4: *Oxford Advanced Learner's Dictionary of Current English, 4th edition.* (1989). Oxford: Oxford University Press.

ODE: *Oxford Dictionary of English.* (2005). Oxford: Oxford University Press.
OEC: *Oxford English Corpus.* http://www.askoxford.com/oec. Accessed: October 1st 2009.
OED: *Oxford English Dictionary.* Oxford: Oxford University Press. http://www.oed.com/.
Ogden, C.K. & Richards, I.A. (1923). *The meaning of meaning.*
Online Etymology Dictionary. http://www.etymonline.com/. Accessed: April 2009-April 2010.
Osselton, N.E. (2007). Innovation and Continuity in English Learners' Dictionaries: The Single-clause When-definition. *International Journal of Lexicography* **20**(4): 393-399.
Pajzs, J. (2009). On the Possibility of Creating Multifunctional Lexicographical Databases. In Bergenholtz, H., Nielsen, S. & Tarp, S. (eds.) *Lexicography at a Crossroads.* Bern: Peter Lang. 327-354.
Paquot, M. (2007). Towards a Productively-oriented Academic Wordlist. In Walinski, J., Kredens, K. & Gozdz-Roszkowski, S. (eds.) *Corpora and ICT in Language Studies: PALC 2005.* Frankfurt am Main: Peter Lang. 127-140.
PDF Ripper, version 2.01. http://www.pdfpdf.com/pdfconverter.html. PDF Bean Inc. Downloaded: 4 October 2008.
Pearson, J. (1996). The Expression of Definitions in Specialised Texts: A Corpus-based Analysis. In Gellerstam, M., 1 & 2 (eds.) *Euralex '96 Proceedings I-II.* Gothenburg: Gothenburg University. 817-824.
Praninskas, J. (1972). *American University Word List.* London: Longman.
Quirk, R. (1975). The social impact of dictionaries in the UK. In McDavid, R. & Ducket, A. (eds.) *Lexicography in English.* New York: New York Academy of Sciences. 76-88.
RAT: *The Reading Academic Text Corpus.* http://www.rdg.ac.uk/app_ling/corpus.htm. Accessed June 24th 2007.
Römer, U. (2009). The use of phraseological items in apprentice academic writing: Does nativeness matter? (conference presentation). *Aston Corpus Symposium.* UK, Birmingham: Aston University.
Rumshisky, A., Hanks, P., Havasi, C. & Pustejovsky, J. (2006). Constructing a Corpus-based Ontology using Model Bias. *FLAIRS 2006.* Melbourne Beach, Florida. 327-332.
Rundell, M. (1998). Recent Trends in English Pedagogical Lexicography. *International Journal of Lexicography* **11**(4): 315-342.
Rundell, M. (1999). Dictionary Use in Production. *International Journal of Lexicography* **12**(1): 35-53.
Rundell, M. (2002). How the MED was written? http://www.macmillandictionary.com/createhow.htm#. Downloaded in 2002 (no longer available online).
Rundell, M. (2006). More than One Way to Skin a Cat: Why Full-Sentence Definitions Have not Been Universally Adopted. In Corino, E., Marello, C. & Onesti, C. (eds.) *Euralex 2006: Proceedings.* Turin: Edizioni dell'Orso. 323-337.
Schmitt, N. (2000). *Vocabulary in Language Teaching.* Cambridge: Cambridge University Press.
Schmitt, N. & McCarthy, M. (eds.) (1997). *Vocabulary: Description, Acquisition and Pedagogy.* Cambridge: Cambridge University Press.
Scholfield, P. (1999). Dictionary Use in Reception. *International Journal of Lexicography* **12**(1): 13-34.
Scott, M. (1997). PC analysis of key words -- And key key words. *System* **25**(2): 233-245.
Scott, M. (1999). WordSmith Tools version 3. Oxford University Press.
Scott, M. (2007). WordSmith Tools version 4.0.0.387. Oxford University Press.

Scott, M. & Johns, T. (1993). MicroConcord. http://langbank.engl.polyu.edu.hk/corpus/microconcord.html. Oxford University Press. Accessed: April 2006 - April 2007.
Silver, M. (2003). The stance of stance: a critical look at ways stance is expressed and modeled in academic discourse. *Journal of English for Academic Purposes* **2**(4): 359-374.
Sinclair, J. (1985). Lexicographic evidence. In Ilson, R. (ed.) *Dictionaries, Lexicography and Language Learning.* Oxford: Pergamon Press. 81-94.
Sinclair, J. (1987). *Looking up: An Account of the COBUILD Project in Lexical Computing.* London: Collins.
Sinclair, J. (1991). *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.
Sinclair, J. (2004a). In Praise of the Dictionary. In Williams, G. & Vessier, S. (eds.) *Proceedings of the Eleventh EURALEX International Congress.* Lorient, France: Universite de Bretagne-Sud. 1-12.
Sinclair, J. (2004b). *Trust the text.* London: Routledge.
Sinclair, J. (2005). Corpus and Text - Basic Principles. In Wynne, M. (ed.) *Developing Linguistics Corpora: A Guide to Good Practice.* Oxford: Oxbow Books. 1-16.
Sketch Engine. http://www.sketchengine.co.uk/. Lexical Computing Ltd. Accessed: October 2008-February 2009.
SPSS Inc. Statistical Package for Social Sciences, version 12. Chicago, USA.
Statistics Canada (2009). Table 1. University enrolment by registration status, program level and gender. Available at http://www.statcan.gc.ca/daily-quotidien/090713/t090713a1-eng.htm. Accessed: 9. September 2009.
Stock, P. (1988). The structure and function of definitions. In Snell-Hornby, M. (ed.) *Zurilex '86 Proceedings.* Tübingen: Francke. 81-89.
Stotesbury, H. (2003). Evaluation in research article abstracts in the narrative and hard sciences. *Journal of English for Academic Purposes* **2**(4): 327-341.
Svensén, B. (1993). *Practical Lexicography.* Oxford: Oxford University Press.
Swales, J. (1990). *Genre Analysis: English in Academic and Research Settings.* Cambridge: Cambridge University Press.
Swales, J. (2004). *Research Genres: Exploration and Applications.* Cambridge: Cambridge University Press.
Swales, J., Ahmad, U.K., Chang, Y., Chavez, D., Dressen, D.F. & Seymour, R. (1998). Consider this: The role of imperatives in scholarly writing. *Applied Linguistics* **19:** 97-121.
T2K-SWAL: Biber, D., Conrad, S., Reppen, R., Byrd, P. & Helt, M. (2002). Speaking and Writing in the University: A Multidimensional Comparison. *TESOL Quarterly* **36**(1): 9-48.
Taylor, A. & Chan, A. (1994). 'Pocket electronic dictionaries and their use. In Martin, W., Meijs, W., Moerland, M., ten Pas, E., van Sterkenburg, P. & Vossen, P. (eds.) *Euralex '94 Proceedings.* Amsterdam: Euralex. 598-605.
Thompson, P. (2005). Points of focus and position: Intertextual reference in PhD theses. *Journal of English for Academic Purposes* **4**(4): 307-323.
Thompson, P. (2006). Assessing the contribution of corpora to EAP practice. In Kantaridou, Z., Papadopoulou, I. & Mahili, I. (eds.) *Motivation in Learning Language for Specific and Academic Purposes.* Macedonia: University of Macedonia.
Thompson, P. (2008). Disciplinary variation in assessed writing for undergraduate programmes. http://corpus.aston.ac.uk/Symposium08/sym_speakers/Thompson.pdf. Presented at the Aston Corpus Symposium, Aston University, Birmingham, May 2008. Accessed: 3 December 2008.

Thompson, S.E. (2003). Text-structuring metadiscourse, intonation and the signalling of organisation in academic lectures. *Journal of English for Academic Purposes* **2**(1): 5-20.
Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work.* Amsterdam: John Benjamins.
Tomaszczyk, J. (1979). Dictionaries: Users and Uses. *Glottodidactica* **12:** 103-119.
Tono, Y. (1984). On the dictionary user's reference skills (B.Ed. Dissertation). University of Tokyo.
TreeTagger. http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/. Institute for Computational Linguistics, University of Stuttgart.
Tribble, C. (2008). In this present paper... Some emerging norms in lingua franca English writing in the sciences? In Frankenberg-Garcia, A. (ed.) *Proceedings of the 8th Teaching and Language Corpora Conference.* Lisbon: ISLA. 307-309.
TshwaneDJe Human Language Technology http://tshwanedje.com/. Accessed: 8 October 2009.
TshwaneLex Professional, version 4.0.0. http://tshwanedje.com/. TshwaneDJe HLT. Downloaded: 15 August 2009.
Tucker, P. (2003). Evaluation in the art-historical research article. *Journal of English for Academic Purposes* **2**(4): 291-312.
U.S. Department of Education (2008). Table A-38-1. Percentage distribution of fall enrollment in degree-granting institutions, by percent combined enrollment of Black, Hispanic, Asian/Pacific Islander, and American Indian/Alaska Native students at institution, control and type of institution, and race/ethnicity: Academic year 2007. National Center for Education Statistics. Available at: http://nces.ed.gov/programs/coe/2009/section5/table-hec-1.asp. Accessed: 9. September 2009.
Unifier, version 4.0. http://www.melody-soft.com. Melody-Soft. Downloaded: 24 October 2008.
Universities UK (2009). Patterns of Higher Education Institutions in the UK: Ninth Report. Universities UK: http://www.universitiesuk.ac.uk/Publications/Documents/Patterns9.pdf.
Vongpumivitch, V., Huang, J.-y. & Chang, Y.-C. (2009). Frequency analysis of the words in the Academic Word List (AWL) and non-AWL content words in applied linguistics research papers. *English for Specific Purposes* **28**(1): 33-41.
W3: *Webster's Third New International Dictionary.* (2002). Springfield: Merriam-Webster.
Wang, J., Liang, S.-l. & Ge, G.-c. (2008). Establishment of a Medical Academic Word List. *English for Specific Purposes* **27**(4): 442-458.
Ward, J. (2005). The lexical aspect of reading English as a foreign language for engineering undergraduates. Unpublished PhD thesis.: UK: University of Birmingham.
Ward, J. (2007). Collocation and technicality in EAP engineering. *Journal of English for Academic Purposes* **6**(1): 18-35.
Ward, J. (2009). A basic engineering English word list for less proficient foundation engineering undergraduates. *English for Specific Purposes* **28**(3): 170-182.
West, M. (1953). *A General Service List of English Words.* London: Longman, Green and Co.
West, R. (1987). A Consumer's Guide to ELT Dictionaries. In Sheldon, L.E. (ed.) *ELT Textbooks and Materials: Problems in Evaluation and Development.* Oxford: Modern English Publications. 55-75.
Williams, G. (2008). Verbs of Science and the Learner's Dictionary. In Bernal, E. & DeCesaris, J. (eds.) *Proceedings of the Thirteenth EURALEX International Congress.* Barcelona: Universitat Pompeu Fabra. 797-806.
Williams, G.C. (1998). Collocational networks: interlocking patterns of lexis in a corpus of plant biology research articles. *International Journal of Corpus Linguistics* **3**(2): 151-171.

Williams, G.C. (2006). Advanced ESP and the Learner's Dictionary: Tools for the Non-Language Specialist. In Corino, E., Marello, C. & Onesti, C. (eds.) *Euralex 2006: Proceedings.* Turin: Edizioni dell'Orso. 795-801.

Williams, J. (1996). Enough Said: The Problems of Obscurity and Cultural Reference in Learner's Dictionary Examples. In Gellerstam, M., 1 & 2 (eds.) *Euralex '96 Proceedings I-II.* Gothenburg: Gothenburg University. 497-505.

Wray, A. (2000). Formulaic sequences in second language teaching: Principle and practice. *Applied Linguistics* **21**(4)**:** 463-489.

Xue, G.Y. & Nation, P. (1984). A University Word List. *Language Learning and Communication* **3:** 215-229.

Zgusta, L. (1971). *Manual of Lexicography.* The Hague: Mouton.

11. APPENDIX 1: CORPORA OF ACADEMIC ENGLISH

Table 98. A list of corpora of academic English.

Corpus and/or Author	Year	Size (words)	Spoken or Written	Contents	Availability
Jiaotong Daxue English for Science and Technology (JDEST) corpus	1985	4 M (now) (originally 1 M)	written	textbooks, academic works, science digests (physics, nuclear energy, metallurgy, computer science, aeronautics, mechanics, electrical engineering, chemical engineering, architectural engineering and shipbuilding)	status unknown
Hong Kong University of Science and Technology (HKUST) Learner Corpus	1992	25 M	written	essays from upper-secondary school and university students	available for use in research on a collaborative basis
MicroConcord -Academic (Scott & Johns, 1993)	1993	1 M	written	books and articles	free online access (http://langbank.engl.polyu.edu.hk/corpus/microconcord.html)
The Gledhill corpus (Gledhill, 1996; 2000)	1995	0.5 M	written	150 cancer research articles	not available
The Academic Corpus (Coxhead, 2000)	2000	3.5 M	written	414 articles, textbooks, monographs and manuals from 28 subject areas	not available
The Hyland corpus (Hyland, 2002)	2001	1.4 M	written	240 journal articles (3 articles from 10 leading journals in 8 disciplines): mechanical engineering, electrical engineering, marketing, philosophy, sociology, applied linguistics, physics, and microbiology	not available
The Uppsala Student English Corpus (USE) (Axelsson, 2003)	2001	1.2 M	written	1489 argumentative, reflective, and personal essays, literature and culture course assignments written by 440 Swedish university students of English	available at the Oxford Text Archive (http://www.ota.ahds.ac.uk)
Charles (2003) – Politics corpus	2003	0.2 M	written	8 MPhil theses in politics and international relations (written by native speakers)	not available

Corpus and/or Author	Year	Size (words)	Spoken or Written	Contents	Availability
Charles (2003) – Materials corpus	2003	0.3 M	written	8 PhD theses in materials science (written by native speakers)	not available
BAWE (British Academic Written English) (Nesi et al., 2005)	2007	6.5 M	written	upper-level student writing (mark 60+)	available at the Oxford Text Archive (for research only) (http://www.ota.ahds.ac.uk) OR free online access via the Sketch Engine (http://ca.sketchengine.co.uk/open/)
International Corpus of Learner English (ICLE), version 2	2009 (ongoing)	3.7 M	written	argumentative essays, literature examination papers (advanced learners of English from 16 different backgrounds)	available on CD-ROM
The Reading Academic Text (RAT) corpus	2007	over 1 M (native speakers)	written	38 PhD theses – 20 from the Faculty of Agriculture, 7 from the Department of Psychology, 6 from the Department of Food Science and Technology, 1 from the Department of Meteorology, and 3 from the Department of History	not available
Michigan Corpus of Upper-Level Student Papers (MICUSP)	2009	2.6 M	written	student writing (mark A or A-) from 4th year undergraduates to 3rd year graduates (around 829 papers from 16 disciplines)	free online access (http://search-micusp.elicorpora.info/simple/)
The PERC (the Professional English Research Consortium) corpus	2008	17 M	written	journal articles (1995-2002) from 22 science and technology domains	free online access until the end of June 2010 at http://scn02.corpora.jp/~perc04/ (registration is required)
TOEFL 2000 Spoken and Written Academic Language Corpus (T2K-SWAL)	2004	2.7 M (1.7 M written, 1 M spoken)	written and spoken	written: textbooks, course packs, course management, campus writing spoken: class sessions, classroom management office hours, study groups, service encounters	not available

Corpus and/or Author	Year	Size (words)	Spoken or Written	Contents	Availability
Michigan Corpus of Academic Spoken English (MICASE)	2002	1.7 M	spoken	lectures, meetings, interviews, etc.	free online access (http://micase.umdl.umich.edu/m/micase), or pay for CD-ROM
British Academic Spoken English (BASE) corpus	2005	1.2 M	spoken	160 lectures, 40 seminars	free online access via the Sketch Engine (http://ca.sketchengine.co.uk/open/)
The Corpus of English as Lingua Franca in Academic Settings (ELFA) (Mauranen, 2003)	2008	1 M	spoken	various speech events (650 non-native speakers from 51 different first language backgrounds)	available for research on request
Hong Kong Corpus of Spoken English (HKCSE) (Cheng et al., 2005)	2006 (ongoing)	0.9 M	spoken	50 hours of lectures, seminars, student presentations, tutorials and supervisions, workshops for staff	available at http://langbank.engl.polyu.edu.hk/HKCSE/

12. APPENDIX 2: PILOT SURVEY - QUESTIONNAIRE

DICTIONARY QUESTIONNAIRE

Name: _____

Age: _____

Native language: _____

I give permission that the information obtained with this questionnaire can be used for research and teaching purposes at Aston University.

Signature: _____

1) How many years have you studied English so far? _____

2) What course will you do after this pre-sessional course?

☐ undergraduate
☐ postgraduate

Which course? _____

3) Do you own a computer?

☐ YES
☐ NO

4) Have you ever received any training in how to use a dictionary?

☐ YES when (year): _____
how long did it last: _____
in which language was the training provided: _____
☐ TRAINING WAS AVAILABLE BUT I WAS NOT ABLE TO TAKE IT
☐ NO

5) Would you be willing to do a follow-up interview after dictionary-use sessions to discuss if your dictionary skills have changed in any way?

☐ YES (please provide email or other contact information): _____
☐ NO

6) How important is a dictionary for you during these activities? Please rate each element from 1-5 (with 5 being very important and 1 being not important)?

	5 very important	4	3	2	1 not important
writing essays	☐	☐	☐	☐	☐
writing emails, letters, CVs	☐	☐	☐	☐	☐
reading books	☐	☐	☐	☐	☐
listening to lectures	☐	☐	☐	☐	☐
speaking	☐	☐	☐	☐	☐

7) Below you will find a list of resources, marked with letters A, B, C, D, and E. Please indicate which version(s) (book, CD-ROM, etc.) you use and how often. You can write the name of MORE THAN ONE dictionary and publisher!

A) GENERAL MONOLINGUAL ENGLISH DICTIONARY

Publisher and name (if known):

_____

In this type of dictionary, all information on a word (definitions, examples, pronunciation, phrases, etc.) is given in English.
Some of the more known general monolingual English dictionaries include Oxford Dictionary of English, Chambers 21st-Century Dictionary, Collins English Dictionary, and Random House Webster's Unabridged Dictionary.

	more than once a day	once a day	more than once a week	once a week	never / do not own this type of dictionary
printed (book)	☐	☐	☐	☐	☐
CD-ROM	☐	☐	☐	☐	☐
electronic handheld	☐	☐	☐	☐	☐
online (internet)	☐	☐	☐	☐	☐

B) LEARNER'S MONOLINGUAL ENGLISH DICTIONARY

Publisher and name (if known):

_____

This type of dictionary is entirely in English like general monolingual English dictionary but is written for learners of the English language. It often includes information on grammar and usage.
Collins COBUILD Advanced Learner's Dictionary, Oxford Advanced Learner's Dictionary, Longman Dictionary of Contemporary English, and Macmillan English Dictionary for Advanced Learners are the most popular.

	more than once a day	once a day	more than once a week	once a week	never / do not own this type of dictionary
printed (book)	☐	☐	☐	☐	☐
CD-ROM	☐	☐	☐	☐	☐
electronic handheld	☐	☐	☐	☐	☐
online (internet)	☐	☐	☐	☐	☐

C) BILINGUAL DICTIONARY

Language combination: _____

uni-directional or bi-directional (see description below): _____

Publisher and name (if known):

_____

Bilingual dictionaries have entries in one language (usually the one being learned) and translations in another (usually the learner's mother tongue). Some bilingual dictionaries are ***uni-directional*** (they only offer translations from one language to another, for example English to French) and others *are* ***bi-directional*** (they consist of two halves, for example English to French and French to English).

	more than once a day	once a day	more than once a week	once a week	never / do not own this type of dictionary
printed (book)	☐	☐	☐	☐	☐
CD-ROM	☐	☐	☐	☐	☐
electronic handheld	☐	☐	☐	☐	☐
online (internet)	☐	☐	☐	☐	☐

D) THESAURUS

Publisher and name (if known):

_____

A thesaurus is basically a book of synonyms (words with similar meaning) and other related words, sometimes offering antonyms (words with opposite meaning). Examples are rarely provided.
All major publishers of general monolingual English dictionaries also offer thesauruses, such as Oxford Thesaurus of English, Collins Thesaurus A-Z, The Chambers Thesaurus, and Merriam-Webster's Thesaurus. A special example is Longman Language Activator which provides examples and other useful notes.

	more than once a day	once a day	more than once a week	once a week	never / do not own this type of dictionary
printed (book)	☐	☐	☐	☐	☐
CD-ROM	☐	☐	☐	☐	☐
electronic handheld	☐	☐	☐	☐	☐
online (internet)	☐	☐	☐	☐	☐

E) OTHER (Dictionary of Business English, Dictionary of Phrasal Verbs, etc.)

Name, language(s) and publisher: _____

8) What information do you usually look in a dictionary for?

	always	sometimes	never	not offered in my dictionary
I want to know the meaning of a word and therefore I read the **DEFINITION(S).**	☐	☐	☐	☐
I want to see more **EXAMPLES** of how the word is used in context.	☐	☐	☐	☐
I want to check how a word is **SPELLED.**	☐	☐	☐	☐
I want to hear (or see) how a word is **PRONOUNCED.**	☐	☐	☐	☐
I want to know some **FREQUENT PHRASES** in which a word is used.	☐	☐	☐	☐
I want to know some **SYNONYMS** (words with similar meaning) of the word	☐	☐	☐	☐
I want to know what a word means in my **NATIVE LANGUAGE.**	☐	☐	☐	☐
I want to read more about **THE USAGE AND GRAMMAR** of a word.	☐	☐	☐	☐
I want to see what other words frequently **COLLOCATE** with a word. (Collocates are words that are often used together, although they are not a fixed phrase; e.g. work + hard, hard + work; draw + picture; open/shut + door; bad/wicked + person; bad/poor + decision.)	☐	☐	☐	☐

9) Which version of a dictionary do you prefer?

☐ paper (book) version

☐ CD-ROM version

☐ electronic handheld version

☐ online (internet) version

Please explain why:

_____

_____

10) Is there any information which you think is often missing in dictionaries?

_____

_____

13. APPENDIX 3: MAIN SURVEY - QUESTIONNAIRE

PAGE 1

The use of English-English dictionaries by students and academics

My Surveys | Create Survey | My Deta

Welcome

Welcome to the Dictionary use survey. This UK-wide survey aims to establish which monolingual (English-English) dictionaries are used by university students and staff.
The survey is completed anonymously and takes 5-10 minutes to complete.
The purpose is to provide factual information for a PhD research which involves creating a model for a new dictionary for university students.

Iztok Kosem
PhD student in Corpus Linguistics
School of Languages and Social Sciences
Aston University
email: kosemi@aston.ac.uk

PAGE 2

Data Protection

All data collected in this survey will be held anonymously and securely. Any background information (e.g. age, native language) will be used for statistical purposes only.

PAGE 3

Dictionary use - part 1

Please note - once you click on CONTINUE you will **not** be able to return to this page.

Dictionary use at different activities

1. How important is a dictionary for you during these activities?

	very important	important	not very important	not important
a. Writing academic work	○	○	○	○
b. Writing emails, letters, CVs	○	○	○	○
c. Reading academic books, journals, etc.	○	○	○	○
d. Reading books, newspapers, etc.	○	○	○	○
e. Listening to academic lectures	○	○	○	○
f. Speaking with lecturers	○	○	○	○
g. Presenting your work	○	○	○	○

Dictionary format

2. Which dictionary format do you prefer?

- ○ paper dictionary
- ○ dictionary on CD-ROM
- ○ handheld (pocket) electronic dictionary
- ○ web dictionary (online)
- ○ no preference

Why do you prefer this dictionary format? (list some of the most relevant features) *(Optional)*

3. How often do you use these dictionary formats?

	all the time	often	rarely	never
a. paper dictionary	○	○	○	○
b. dictionary on CD-ROM	○	○	○	○
c. handheld (pocket) electronic dictionary	○	○	○	○
d. web dictionary (online)	○	○	○	○

English-English Dictionaries

4. Are you using any English-English dictionaries (where English words are defined in English) in any format?

- ○ Yes
- ○ No

If Yes, how long have you been using it?
Select an answer

5. Do you know the following dictionaries?

	I've never heard of it before.	I've heard of it, but haven't used it.	I use it occasionally.	I use it regularly.
a. Longman Exams Dictionary	○	○	○	○
b. Compact Oxford English Dictionary for University and College Students	○	○	○	○
c. Merriam-Webster Collegiate Dictionary	○	○	○	○

PAGE 4

Dictionary use - part 2

Dictionary features

6. What information do you usually look in a dictionary for?

	almost always	often	rarely	almost never
a. I want to know the meaning of a word and therefore I read the DEFINITION(S).	○	○	○	○
b. I want to see more EXAMPLES of how the word is used in context.	○	○	○	○
c. I want to check how a word is SPELLED.	○	○	○	○
d. I want to hear (or see) how a word is PRONOUNCED.	○	○	○	○
e. I want to know some FREQUENT PHRASES in which a word is used.	○	○	○	○
f. I want to know some SYNONYMS (words with similar meaning) of the word.	○	○	○	○
g. I want to read more about THE USAGE AND GRAMMAR of a word.	○	○	○	○
h. I want to see what other words frequently COLLOCATE with a word. (Collocates are words that are often used together, although they are not a fixed phrase; e.g. work + hard, hard + work; draw + picture; open/shut + door; bad/wicked + person; bad/poor + decision.)	○	○	○	○

7. To what extent to you agree or disagree with the following statements?

	strongly agree	partly agree	partly disagree	strongly disagree
a. I often look at only the first sense of a word.	○	○	○	○
b. If one dictionary does not provide me with the answer, I look into another one.	○	○	○	○
c. When buying a dictionary, the name of the publisher is a very important factor.	○	○	○	○

PAGE 5

Dictionary information

This question seems information on the **ENGLISH-ENGLISH** dictionaries only. If you are not using any English-English dictionaries, please skip this question.

Your English-English dictionaries

Please provide information about ONE OR TWO *English-English* dictionaries you are using. PLEASE TYPE AS MUCH INFORMATION AS POSSIBLE.

8. YOUR ENGLISH-ENGLISH DICTIONARIES

	DICTIONARY 1	DICTIONARY 2
a. Full title (e.g. the New Oxford Dictionary of English):		
b. Dictionary format (e.g. paper, CD-ROM, handheld electronic, or online):		
c. Type of dictionary (e.g. dictionary for learners, general dictionary, technical dictionary)		
d. Publisher (e.g. Cambridge):		
e. URL (if an online dictionary):		
f. Edition (if known):		
g. Year of publication:		

PAGE 6

Final Page

Please note - once you click on CONTINUE you will **not** be able to return to this page. Your answers will be submitted automatically.

About You

9. Are you a university student?

○ Yes ○ No

If yes:

a. Please select your status:
Select an answer

b. What is your programme of study?

10. Are you a member of university staff?

○ Yes ○ No

If yes, please select your status:
Select an answer

If you selected Other, please specify:

11. Please select the university where you study/work?
(The words "(the) University of ..." were moved to the end to make the alphabetical list easier to use. So, for example, "The University of Kent" can be found under "Kent, The University of".)

Select an answer

Department/School *(if known)* *(Optional)*

12. **Are you:**

○ a native-speaker of English
○ a non-native speaker of English

If non-native speaker of English:

a. What is your native language?

b. How many years have you been learning English?
Select an answer

c. How old were you when you started learning English?

13. Please select a country to describe your nationality. *(Optional)*

Select an answer

14. **How would you rate your English language proficiency for these activities?**

	Very poor	Poor	Fair	Good	Very good	Excellent
a. Listening	○	○	○	○	○	○
b. Reading	○	○	○	○	○	○
c. Speaking	○	○	○	○	○	○
d. Writing	○	○	○	○	○	○

15. What is your gender *(Optional)*

○ Female ○ Male

16. **How old are you?**

14. APPENDIX 4: CAJA TABLES

Table 99. 10 UK universities, HESA subject list, and in 5 dictionaries: Representation of initial 33 domain categories for CAJA (part 1).

	Worcester	Aston	Durham	Lancaster	Oxford	Cambridge	Birmingham	Manchester	London	Open
Anthropology	X	X	✓	X	≠	≠	X	X	≠	X
Archaeology	✓	X	✓	X	✓	≠	≠	X	✓	X
Architecture	X	X	X	X	X	✓	X	X	≠	X
Biochemistry	X	✓	≠	≠	✓	X	≠	✓	✓	X
Biology	✓	✓	≠	✓	≠	X	≠	✓	✓	≠
Business and Management	✓	✓	≠	≠	≠	≠	≠	≠	≠	✓
Chemistry	X	✓	✓	X	✓	X	✓	✓	✓	✓
Classics and Ancient History	X	X	✓	X	≠	≠	X	≠	✓	≠
Computer Science (Computing)	≠	✓	✓	✓	✓	✓	✓	✓	✓	≠
Creative Arts (Craft & Design)	≠	X	X	≠	X	X	X	X	≠	X
Cultural and Media studies	✓	≠	X	X	X	X	X	X	≠	✓
Drama, Theatre & Dance	≠	X	X	≠	X	X	≠	X	✓	X
Economics	✓	✓	✓	✓	≠	✓	✓	≠	✓	✓
Education	✓	≠	✓	✓	X	✓	✓	✓	✓	✓
Engineering	X	✓	✓	✓	≠	✓	✓	≠	✓	✓
Finance	✓	✓	≠	≠	X	X	✓	✓	✓	≠
Geography, Earth & Environ. studies	≠	X	≠	≠	≠	≠	✓	≠	≠	≠
History	✓	X	✓	✓	✓	✓	✓	≠	✓	✓
History of Art	X	X	X	X	✓	✓	X	X	✓	✓
Law	X	≠	✓	✓	✓	✓	✓	✓	✓	✓
Linguistics (Language & Literature)	≠	≠	≠	≠	≠	✓	≠	≠	✓	≠
Mathematics	X	✓	✓	≠	✓	✓	✓	✓	✓	✓
Medicine and Health sciences	≠	≠	≠	X	≠	≠	≠	≠	≠	≠
Music	X	X	✓	≠	✓	✓	✓	X	✓	✓
Philosophy	X	X	✓	≠	✓	✓	✓	X	✓	✓
Physics	X	X	✓	✓	✓	X	≠	≠	✓	✓
Politics, Government & Int Relations	X	≠	≠	✓	X	≠	≠	X	≠	≠
Psychology	X	✓	✓	✓	X	X	✓	≠	✓	✓
Social studies (Sociology)	≠	✓	≠	≠	X	≠	✓	≠	≠	≠
Sports	✓	X	✓	X	≠	X	≠	X	≠	X
Theology & Religion	X	X	✓	≠	≠	✓	≠	X	≠	≠
Veterinary Science	≠	X	X	X	X	≠	X	X	✓	X

Table 99. 10 UK universities, HESA subject list, and in 5 dictionaries: Representation of initial 33 domain categories for CAJA (part 2).

	HESA	NODE CD-ROM	CED CD-ROM	Chambers	MEDAL1	LED CD-ROM
Anthropology	X	✓	✓	✓	X	X
Archaeology	X	✓	✓	✓	X	X
Architecture	≠	✓	✓	✓	X	X
Biochemistry	≠	✓	✓	✓	X	X
Biology	≠	✓	✓	✓	X	X
Business and Management	≠	X	≠	≠	≠	X
Chemistry	X	✓	✓	✓	X	X
Classics and Ancient History	≠	X	≠	X	X	X
Computer Science (Computing)	≠	✓	✓	✓	✓	X
Creative Arts (Craft & Design)	✓	X	X	X	X	X
Cultural and Media studies	X	X	X	X	≠	X
Drama, Theatre & Dance	X	≠	≠	≠	X	X
Economics	X	✓	✓	✓	X	X
Education	✓	X	✓	✓	X	X
Engineering	✓	X	✓	✓	X	X
Finance	X	✓	✓	✓	X	X
Geography, Earth and Environ. studies	X	X	≠	≠	X	X
History	≠	≠	✓	≠	X	X
History of Art	≠	X	X	X	X	X
Law	✓	✓	✓	✓	✓	√
Linguistics (Language & Literature)	≠	✓	✓	✓	✓	≠
Mathematics	≠	✓	✓	✓	X	X
Medicine and Health sciences	≠	✓	✓	✓	✓	√
Music	X	✓	✓	✓	X	X
Philosophy	≠	✓	✓	✓	X	X
Physics	≠	✓	✓	✓	X	X
Politics, Government & Int. Relations	X	X	≠	≠	X	X
Psychology	X	✓	✓	✓	X	X
Social studies (Sociology)	✓	X	✓	✓	X	X
Sports	X	≠	✓	✓	X	X
Theology & Religion	X	✓	≠	✓	X	X
Veterinary Science	≠	X	✓	✓	X	X

Key:
✓ - subject with exactly the same name found; ≠ - subject with similar name found; X – subject not found

Table 100. CAJA: Number of texts by year.

SUBCORPUS	1993	1994	1995	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	total
Anthropology	1	1	2	9	0	0	3	0	7	18	37	224	46	0	348
Archaeology									4	3	30	139	136	0	312
Architecture							3	28	11	19	8	200	112	6	387
Arts and Art History								5	0	12	32	241	51	9	350
Biochemistry											0	758	37	13	808
Biology											14	603	68	63	748
Business and Management						26	3	25	16	51	20	327	24	0	492
Chemistry												607	41	16	664
Computer Science								2	0	0	6	439	71	16	534
Economics					3	11	11	16	16	31	10	273	53	7	431
Education								2	0	0	2	451	41	0	496
Engineering											8	547	18	8	581
Finance						4	7	0	4	17	11	325	92	0	460
Geography, Earth and Env. Studies										4	13	304	74	5	400
History						1	1	1	1	14	15	431	38	3	505
Law										3	17	295	50	40	405
Linguistics									6	10	30	363	64	0	473
Mathematics								13	18	0	5	303	0	47	386
Medicine and Health Sciences											10	368	77	5	460
Music						5	7	0	6	25	76	152	79	5	355
Philosophy										8	16	342	111	45	522
Physics										8	0	324	14	10	356
Politics, Gov. & Int. Relations											9	318	24	31	382
Psychology												326	2	15	343
Social Sciences										5	5	290	23	13	336
Sports							8	14	6	9	4	318	12	27	398
Theology and Religion							4	1	12	21	28	283	94	10	453
Veterinary Science								63	0	0	42	442	51	133	731
TOTAL	**1**	**1**	**2**	**9**	**3**	**47**	**47**	**170**	**107**	**258**	**448**	**9.993**	**1.503**	**527**	**13.116**
%	0.01	0.01	0.02	0.07	0.02	0.36	0.36	1.30	0.82	1.97	3.42	76.19	11.46	4.02	100.00

Table 101. CAJA: Number of authors per text by domain category.

SUBCORPUS	single author	two authors	three or more authors
Anthropology	305	27	16
Archaeology	160	61	91
Architecture	199	100	88
Arts and Art History	316	23	11
Biochemistry	37	103	668
Biology	50	155	543
Business and Management	137	176	179
Chemistry	39	105	520
Computer Science	77	188	269
Economics	141	175	115
Education	227	134	135
Engineering	19	144	418
Finance	122	191	147
Geography, Earth and Env. Studies	67	113	220
History	455	40	10
Law	301	79	25
Linguistics	257	116	100
Mathematics	91	158	137
Medicine and Health Sciences	16	67	377
Music	326	19	10
Philosophy	464	43	15
Physics	53	87	216
Politics, Government & Int. Relations	222	110	50
Psychology	46	93	204
Social Sciences	105	104	127
Sports	36	79	283
Theology and Religion	402	33	18
Veterinary Science	47	89	595
TOTAL	4.717	2.812	5.587
%	35.96	21.44	42.60

Table 102. Sketch Engine – POS-tagging: TreeTagger version of Penn Treebank Tagset.

POS Tag	Description	Example
CC	coordinating conjunction	and
CD	cardinal number	1, third
DT	determiner	the
EX	existential there	*there* is
FW	foreign word	d'hoevre
IN	preposition, subordinating conjunction	in, of, like
IN/that	*that* as subordinator	that
JJ	adjective	green
JJR	adjective, comparative	greener
JJS	adjective, superlative	greenest
LS	list marker	1)
MD	modal	could, will
NN	noun, singular or mass	table
NNS	noun plural	tables
NP	proper noun, singular	John
NPS	proper noun, plural	Vikings
PDT	predeterminer	*both* the boys
POS	possessive ending	friend*'s*
PP	personal pronoun	I, he, it
PP$	possessive pronoun	my, his
RB	adverb	however, usually, naturally, here, good
RBR	adverb, comparative	better
RBS	adverb, superlative	best
RP	particle	give *up*
SENT	Sentence-break punctuation	. ! ?
SYM	Symbol	/ [= *
TO	infinitive 'to'	*to* go
UH	interjection	uhhuhhuhh
VB	verb *be*, base form	be
VBD	verb *be*, past tense	was, were
VBG	verb *be*, gerund/present participle	being
VBN	verb *be*, past participle	been
VBP	verb *be*, sing. present, non-3d	am, are
VBZ	verb *be*, 3rd person sing. present	is
VH	verb *have*, base form	have
VHD	verb *have*, past tense	had

VHG	verb *have*, gerund/present participle	having
VHN	verb *have*, past participle	had
VHP	verb *have*, sing. present, non-3d	have
VHZ	verb *have*, 3rd person sing. present	has
VV	verb, base form	take
VVD	verb, past tense	took
VVG	verb, gerund/present participle	taking
VVN	verb, past participle	taken
VVP	verb, sing. present, non-3d	take
VVZ	verb, 3rd person sing. present	takes
WDT	wh-determiner	which
WP	wh-pronoun	who, what
WP$	possessive wh-pronoun	whose
WRB	wh-abverb	where, when
#	#	#
$	$	$
"	Quotation marks	' "
``	Opening quotation marks	' "
(	Opening brackets	({
)	Closing brackets	) }
,	Comma	,
:	Punctuation	- ; : -- ...

15. APPENDIX 5: MAIN SURVEY – TABLES AND FIGURES

Table 103. Main survey: Native languages reported by NNS students (n=171).*

German (19)	Finnish (2)	Indian
Chinese (16)	Italian (2)	Irish
Greek (12)	Kiswahili (2)	Kazakh
Polish (12)	Lithuanian (2)	Kokni
Urdu (11)	Malay (2)	Latvian
Russian (9)	Pashto (2)	Luganda
French (7)	Somali (2)	Malayalam
Gujarati (5)	Afrikaans	Mandarin Chinese
Japanese (5)	Bemba	Memon
Persian (4)	Bulgarian	Nbedele
Punjabi (4)	Cantonese	Ndebele-Zulu
Arabic (4)	Croatian	Pidgin English
Hindi (3)	Estonian	Romanian
Portuguese (3)	French and Creole	Saraiki
Shona (3)	French and German	Singhalese
Spanish (3)	German Czech	Somali
Tamil (3)	Hausa	Welsh
Vietnamese (3)	Hindco	Yoruba
Bengali (2)	Hungarian	

* - languages with more than one student are accompanied by the number of students in brackets

Figure 96. Main survey: Age distribution of the students.

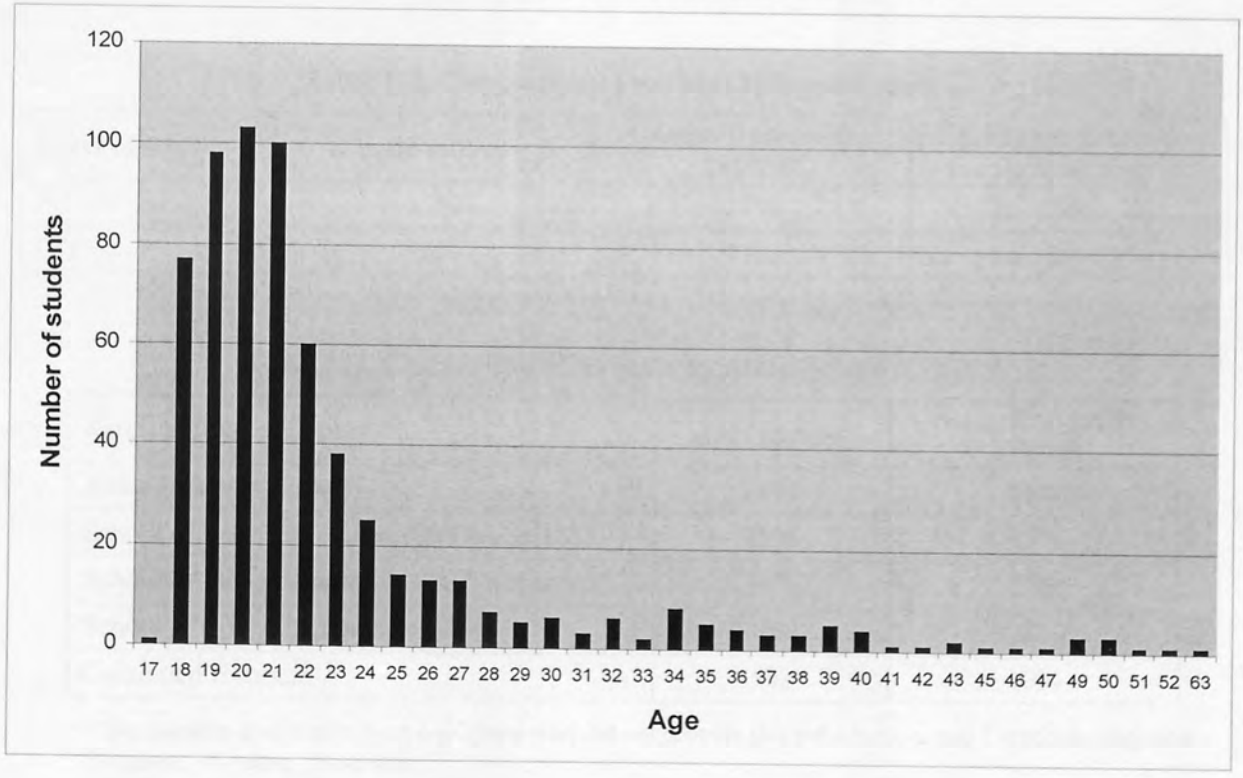


Table 104. Comparison of students by age group.*

Age group	main survey	Aston University 2007/08
20 and under	45%	47%
21-24	36%	33%
25 and over	19%	20%

* HESA does not provide data for student age groups.

Table 105. Comparison of students by gender.

gender	main survey*	Aston University 2007/08	UK Higher Education institutions 2007/08
Female	66%	49%	57%
Male	34%	51%	43%

* The data does not include 8 students who did not provide this information.

Table 106. Comparison of students by country of domicile.

Country of domicile	main survey*	Aston University 2007/08	UK Higher Education institutions 2007/08
UK	68%	74%	85%
Other EU	13%	5%	5%
Non-EU	19%	21%	10%

* The data does not include 32 students who did not provide this information.

Table 107. Comparison of students by level of study.

Level of study	main survey	Aston University 2007/08	UK Higher Education institutions 2007/08
PG	25%	27%	22%
UG	75%	73%	78%

Table 108. Comparison of students by Aston School of Study.

Aston School of Study	main survey*	Aston University 2007/08
Aston Business School	30%	36%
School of Engineering and Applied Science	17%	23%
School of Languages and Social Sciences	14%	8%
School of Life and Health Sciences	30%	23%
Combined Honours	9%	10%

* The number does not include 6 students who did not provide this information, and 7 students who were students of Aston University.

Figure 97. Main Survey: Frequency of use of four dictionary formats (by percentage of students).

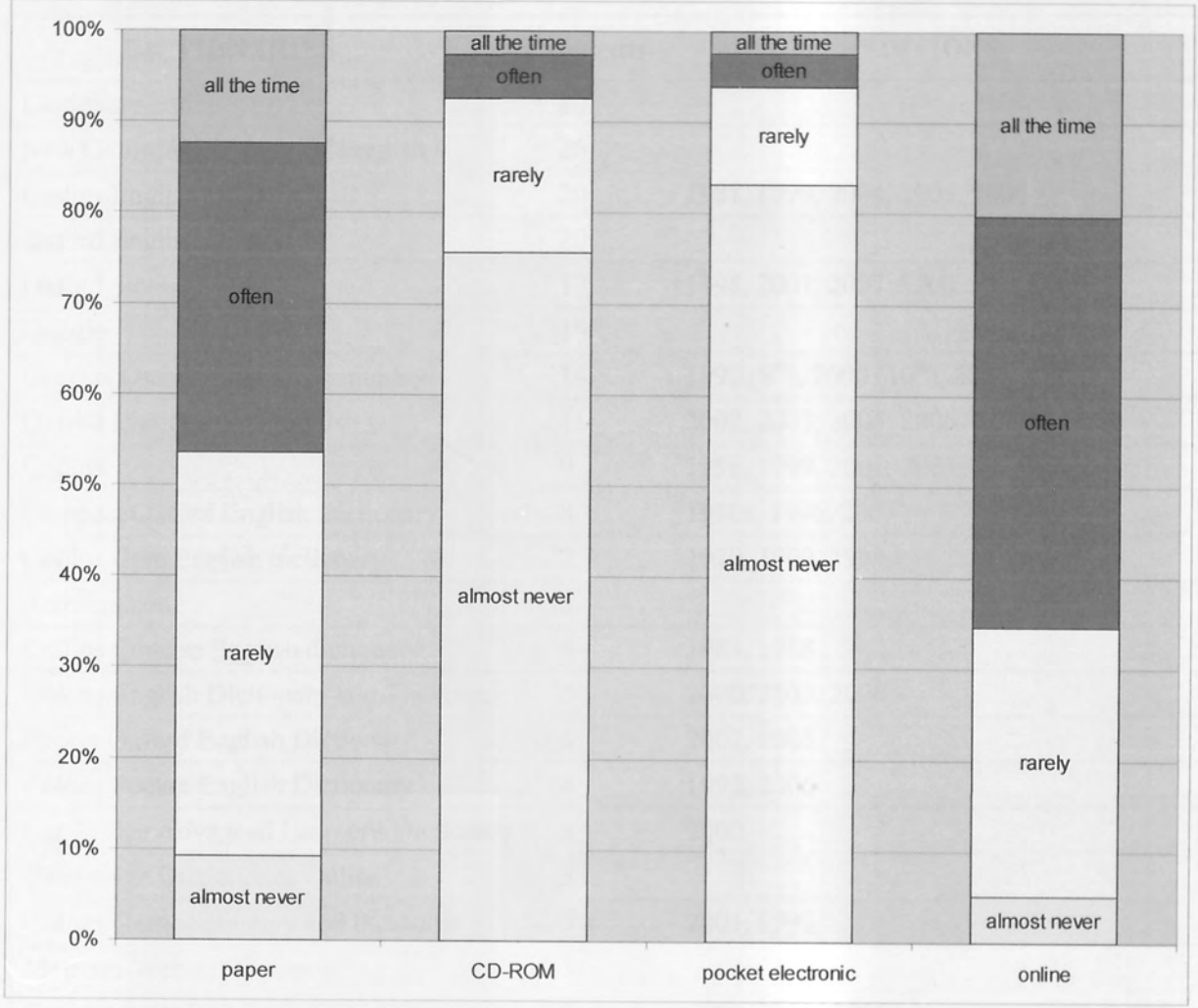


Table 109. Main survey: Monolingual English dictionaries reported by NS students (n=449).

DICTIONARIES	No. of students	EDITIONS
Dictionary.com	80	
New Oxford Dictionary Of English	26	
Collins English Dictionary	20	1981, 1999, 2004, 2005, 2006
Oxford English Dictionary	20	
Oxford dictionary	17	1998, 2001, 2007 + Askoxford.com
Google	15	
Concise Oxford English Dictionary	14	1990 (8^{th}), 2000 (10^{th}), 2004 (11^{th})
Oxford Dictionary of English	11	2002, 2003, 2005, 2006
Collins	9	1996, 1999, 2001, 2005, 2006
Compact Oxford English Dictionary	8	1980s, 1996, 2007
Collins Gem English dictionary	7	1970, 1990, 1994
Answers.com	5	
Collins Concise English dictionary	5	1984, 1988
Collins English Dictionary and Thesaurus	5	2000, 2003, 2006
Pocket Oxford English Dictionary	5	2002, 2005
Collins Pocket English Dictionary	4	1992, 2006
Cambridge Advanced Learner's Dictionary	3	2000
Cambridge Dictionaries Online	3	
Collins Gem Dictionary and thesaurus	3	2001, 1999
Merriam-Webster Online	3	
Paperback Oxford English Dictionary	3	1990, 2002, 2006
Yourdictionary.com	3	

* - Dictionaries reported by fewer than three students were not included in the list.

Table 110. Main survey: Monolingual English dictionaries reported by NNS students (n=171).

DICTIONARIES	No. of students	EDITIONS
Oxford Advanced Learner's Dictionary	17	2005, 2003, 2000, 1988
Dictionary.com	16	
Longman Dictionary of Contemporary English	10	2000, 2003, 2007
Oxford Dictionary of English	10	1989, 2000, 2006
New Oxford Dictionary of English	9	
Oxford dictionary	9	
Cambridge Advanced Learner's Dictionary	8	2003, 2005, 2007
Collins English Dictionary	8	1992, 2005, 2006
Concise Oxford English Dictionary	4	1996, 2001, 2004
Collins	3	2000
Wikipedia	3	
Collins Cobuild English Dictionary for Advanced Learners	2	
Collins Gem English Dictionary	2	1968, 1997
Google	2	
Leo.de	2	
Macmillan English Dictionary for Advanced Learners	2	
Merriam-Webster Online	2	
Oxford Concise Medical Dictionary	2	1998
Oxford English Dictionary	2	2005
Webster's English Dictionary	2	
Wordnet (Princeton university)	2	

* - Dictionaries in bold font were also named also by NS students.
** - Dictionaries reported by one student only were not included in the list.

Figure 98. Main survey: Mean ranks for attributed importance of dictionary use and reported English proficiency for related activities.

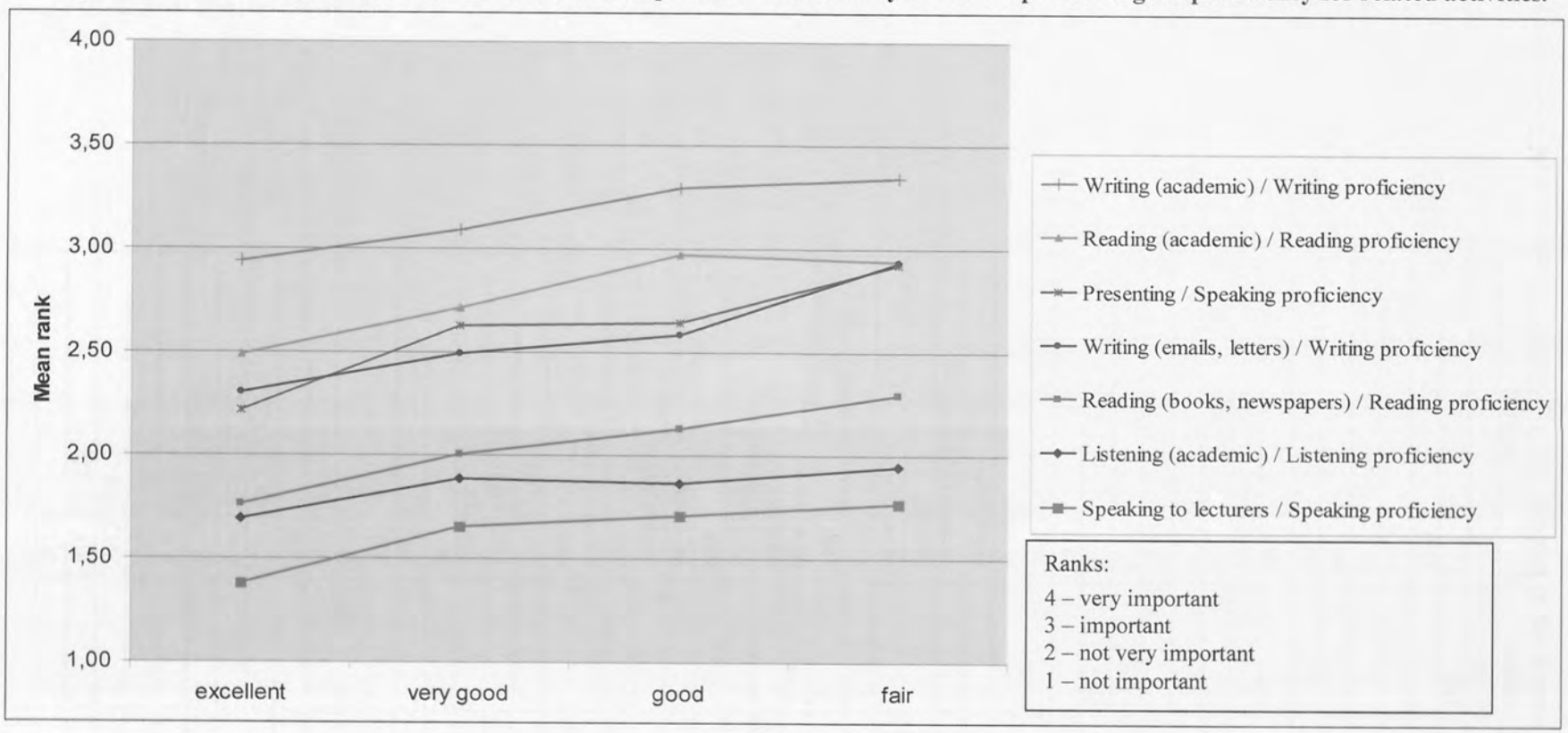


Table 111. Main survey: Mean rank data for Figure 98.

	excellent	very good	good	fair
Writing (academic) / Writing proficiency	2.94	3.09	3.30	3.34
Reading (academic) / Reading proficiency	2.49	2.71	2.97	2.94
Presenting / Speaking proficiency	2.22	2.63	2.64	2.92
Writing (emails, letters) / Writing proficiency	2.30	2.49	2.58	2.93
Reading (books, newspapers) / Reading proficiency	1.76	2.01	2.13	2.29
Listening (academic) / Listening proficiency	1.70	1.89	1.87	1.95
Speaking to lecturers / Speaking proficiency	1.38	1.66	1.71	1.77

16. APPENDIX 6: MACROSTRUCTURE OF DOAE - TABLES

Table 112. CAJA in Sketch Engine: First 50 lemmas in the lemma list (alphabetically ordered).

lemma	frequency
	23335
!	8482
!"31	1
!(2m	3
!8	1
!P	1
!U	1
!gg	1
!liquidity	1
!|t	1
!’	1
!”	1
"	306244
#	555
##	9
###	4
######	1
#########	1
###mdit###Bondage	1
##1078	1
##36	1
##38	1
##7	1
##75-76	1
##85-89	1
##87-88	1
##9	1
##95-96	1
##Difference	1
##P	4
##number##7	1
##p	4
##p < 0.001	1
##p < 0.01	3
#$shår§	1
#(	2
#((I	1
#((T	1
#(A	1
#(A0)b	1
#(C	1
#(D(x) ∩ D(y)) = 1	2
#(D(x) ∩ D(y)) = 2	1
#(D(x) ∩ D(y)) = 4	1
#(E1(¯j	1
#(Fi	1
#(G\GA/	1
#(Glk(Fp))/	1
#(Gu	2
#(I	2

Table 113. CAJA in Sketch Engine: Lemmas beginning with "extent" on the lemma list.

lemma	frequency
extent	15541
extent'.41	1
extent--be	1
extent--if	1
extent--work	1
extent-growth	1
extent-to	1
extent-with	1
extent.106	1
extent.16	1
extent.2	1
extent.35	1
extent.5	1
extent.53	1
extent.56	1
extent.89	1
extent/	1
extent13	1
extent3	1
extented	1
extention	2
extentions	1
extent - the	1
extent – by	1
extent–the	2
extent—and	1
extent—by	2
extent—conflicts	1
extent—did	1
extent—feed	1
extent—if	2
extent—it	1
extent—significant	1
extent—something	1
extent—that	1
extent'	9
extent"	2

Table 114. CAJA in Sketch Engine: Lemmas in the list containing the letters "attribut" (ordered alphabetically).

Lemma	Frequency
'ATTRIBUTE'	1
'attribute'	1
'attributes	1
'attributes'	1
'attributive'	1
'sub-attributes	1
'sub-attributes'	3
"attribute	1
"attribution-type	1
1-attribute	2
2-attribute	2
aAttribute	1
accomplished—attributes	1
across-attribute	1
attributability	3
attributability"	1
attributable	1329
ATTRIBUTE	6
Attribute	24
attribute	13005
'attribute	2
-attribute	2
attribute--	1
attribute(s	5
attribute".15	1
attribute/	2
attribute"	1
attribute19	1
attribute4	1
attribute—almost	1
attribute—amateurism—in	1
attribute--argument	1
attribute-based	32
attribute-benefit	1
attribute—classification	1
Attributed	8
attributed	1
attributed20	1
attributed—a	1
attributed—the	1
attributedto	5
attributee	4
Attribute—every	1
attribute-filler	1
attribute--for	1
Attribute-Grammar	1
attribute-grammar	1
attribute-host	1
attributeindependence	1
Attribute-independence	1

attribute-independence	4
attributeis	1
attributelevel	1
Attribute-Level	4
Attribute-level	4
attribute-level	59
attribute-managed	1
attribute-model	1
attribute-multiple-brands	2
attribute--nicotine	1
attribute-no-attribute	2
attribute-none	1
attribute-or	2
attribute--persistence--as	1
attribute-quality	8
attributer	2
attribute-rating	2
ATTributes	2
ATTRIBUTES	8
Attributes	30
'attributes	1
attributes > indicators	1
attributes.10	1
attributes.22	1
attributes.5	1
attributes/	3
Attributes'	1
attributes'	10
attributes—a	1
attributes—although	1
attributes-based	2
attributes—cultural	1
attributesg	1
attributes—greed	1
attributes--her	1
attributes—including	1
attributes-low	1
attributes—opening	1
attribute-specific	7
attributesplitting	1
attributes–properties	1
attributes—study	1
attributes--such	2
attributes—such	2
attributes--that	1
attributes-the	2
attributes--the	1
attributes—the	2
attributes--two	1
attribute--that	1
attribute-unique-brands	6
attribute-valuation	5
attributevalue	1
Attribute-Value	1

Attribute-value	1
attribute-value	27
attribute-value/	1
attribute-values	2
attribute-wise	3
attributi	1
Attributing	5
'attributing	1
Attribution	13
attribution	1602
attribution-	1
attribution.79	1
attribution/	4
attribution'	1
Attributional	14
attributional	122
attributionally	3
attribution-based	2
attribution-poor	1
Attributions	5
'attributions	1
attributions.1	1
attributions/	1
attribution-which	1
Attributive	2
attributive	102
attributively	5
'attributively	1
attributor	2
'attributor	1
attributor's	2
brand-attribute	11
brand-attribute-benefit	1
brand-plus-positioning-attribute	2
class-attribute	1
culture-attributes	1
earnings-attribute	1
earnings-attributes	2
factors-attributes	1
hierarchy-attributes	1
Influence-Attributed	3
initiative-attributes	1
irrelevant-attributes	2
item-by-attribute	1
j0-attribute	1
j-attribute	1
JoinAttribute	1
justice-attributes	1
k0-attribute	1
k-attribute	7
media-attribute	1
misattribute	11
misattributes	2
misattributing	2

misattribution	5
misattributions	4
mobilization-attributed	1
Multiattribute	3
multiattribute	13
multi-attribute	16
multiple-attribute	5
nameattribute	1
no-attribute	11
no-attribute-multiple	1
no-attribute-multiple-brands	1
no-attribute-unique-brands	1
nonattribute	1
nonattributive	1
no-positioning-attribute-no-preexposure	1
no-positioning-attribute-preexposure	1
one-attributes	1
overattribute	4
over-attribute	4
over-attributes	1
overattribution	3
person-attribute	1
positioning-attribute	1
positioning-attribute-no-preexposure	2
positioning-attribute-preexposure	1
procedures--attributes	1
product-attribute	2
reattributed	1
reattributes	1
reattribution	2
reattributions	1
self-attributed	1
self-attributes	2
self-attribution	5
singleattribute	1
single-attribute	4
sub-attributes	20
sub-attributes/	1
substance-attribute	1
system-attributed	1
type(t)=attribute	2
unattributed	15
Unnestattribute	1
valueattributes	1
value-attributes	2
value-attribution	1
well-attributed	1
within-attribute	2

Table 115. CAJA in Sketch Engine: Error-lemmas containing "attribut" (ordered alphabetically).

lemma	frequency
"attribute	1
'attribute	2
'attributes	1
'attributing	1
-attribute	2
1-attribute	2
2-attribute	2
ATTRIBUTE	6
ATTRIBUTES	8
ATTributes	2
Attribute	24
Attributed	8
Attributes	30
Attributes'	1
Attribute—every	1
Attributing	5
Influence-Attributed	3
JoinAttribute	1
Unnestattribute	1
aAttribute	1
accomplished—attributes	1
across-attribute	1
attribute".15	1
Attribute(s	5
Attribute--	1
Attribute--argument	1
Attribute--for	1
Attribute--nicotine	1
Attribute--persistence--as	1
Attribute--that	1
Attribute-benefit	1
Attribute-filler	1
Attribute-host	1
Attribute-independence	4
Attribute-managed	1
Attribute-model	1
Attribute-multiple-brands	2
Attribute-no-attribute	2
Attribute-none	1
Attribute-or	2
Attribute-rating	2
Attribute-unique-brands	6
attribute/	2
Attribute19	1
Attribute4	1
attributed20	1
attributedto	5
attributed	1
attributed—a	1
attributed—the	1
attributeis	1

attributes--her	1
attributes--such	2
attributes--that	1
attributes--the	1
attributes--two	1
attributes-based	2
attributes-low	1
attributes-the	2
attributes.10	1
attributes.22	1
attributes.5	1
attributes/	3
attributesg	1
attributesplitting	1
attributes > indicators	1
attributes–properties	1
attributes—a	1
attributes—although	1
attributes—cultural	1
attributes—greed	1
attributes—including	1
attributes—opening	1
attributes—study	1
attributes—such	2
attributes—the	2
attributes'	10
Attribute—almost	1
Attribute—amateurism—in	1
Attribute—classification	1
Attribute"	1
Attributi	1
brand-plus-positioning-attribute	2
class-attribute	1
culture-attributes	1
factors-attributes	1
hierarchy-attributes	1
Initiative-attributes	1
irrelevant-attributes	2
item-by-attribute	1
j-attribute	1
j0-attribute	1
justice-attributes	1
media-attribute	1
mobilization-attributed	1
nameattribute	1
one-attributes	1
person-attribute	1
procedures--attributes	1
product-attribute	2
substance-attribute	1
system-attributed	1
type(t)=attribute	2
value-attributes	2
valueattributes	1

well-attributed	1
within-attribute	2
‘ATTRIBUTE’	1
‘attributes	1
‘attributes’	1
‘attribute’	1

17. APPENDIX 7: RECORDING BASIC INFORMATION – TABLES

Table 116. DOAE: Recording basic information - a random sample of 20 concordance lines (out of 114) of FEATURE tagged with the NP tag.

1	Archaeology	STRATIGRAPHIC CONTEXT AND DISCUSSION OF FALE **FEATURE** AT BLOCK A Of the 15 sondages, only those
2	Archaeology	amounts of charcoal. One of structures (**Feature** 81 complex) measures at least 5 × 4 m
3	Archaeology	from a nearby adze quarry (Site -2510). At **Feature** 2 of Site -2509, 2 14C dates from a single
4	Architecture	interiors. HOK designed a space for Turner **Feature** Animation that used EnviroCoat paint in
5	Architecture	Map Service (WMS) -portrayal service Web **Feature** Service (WFS) -data service Web Coverage
6	Architecture	such as windows, doors from a Building **Feature** Service (BFS). A GetFeature request is
7	Art and Art History	predated the latest occupation represented by **Feature** 6. When visiting Guevavi, Kino requested
8	Biochemistry	BLAST report and generating GFF (General **Feature** Format) files (http: / /www. sanger. ac
9	Biochemistry	Gene Finding Format (GFF, a. k. a. General **Feature** Format), called GFF3. GFF3 (http: / /song
10	Computer Science	the bas-relief ambiguity is labeled B. **Feature** locations are indicated with small black
11	Computer Science	, which can provide quick browsing. (c) **Feature** computation complexity. The computation
12	Computer Science	3.4 Latent Semantic Indexing (LSI) for **Feature** Extraction We have not discussed document
13	Computer Science	4.3.1 Performance as a Function of LSI **Feature** Dimension. First, studied the effects of
14	Computer Science	fingerprint verification system and. (2) **Feature** fusion means to fuse the feature sets from
15	Engineering	Resource Bidding Mechanism With Utility **Feature** (RBMU): This mechanism addresses the efficiency
16	Linguistics	therapy work being undertaken. In table 5 **Feature** E: "Reference to aphasia", "Aphasic impairment
17	Linguistics	corpus. Table 1 Morphological features **Feature** Description Vietnamese word order Example
18	Linguistics	] in SPE-style feature systems. Unified **Feature** Theory has a set of binary features related
19	Linguistics	features of Type B adjectives (Total: 32) **Feature** Entries Examples Begin with / a/ alahaZiba
20	Linguistics	Section 5.1.1. 4.2.3 Translation Model and **Feature** Function Training. After pruning, a tuple

Table 117. DOAE: Recording basic information - concordance lines of ATTRIBUTE tagged with the NP tag.

1	Business and Management	THE MODERATING IMPACT OF PREPURCHASE **ATTRIBUTE** VERIFIABILITY We employed an attributional
2	Business and Management	, Hispanic/ Latino). Matched Valence of **Attribute** Pairs A similar problem of the use of two
3	Computer Science	differently in different trees. That may **Attribute** 1 Attribute 2 <=0.4 > 0.4 <=0.5 > 0.5 Class
4	Computer Science	rules can be translated as follows: Rule 1 **ATTRIBUTE** B (3.8,4.2] THEN Class=setosa; Rule 2 ATTRIBUTE
5	Computer Science	ATTRIBUTE B (3.8,4.2] THEN Class=setosa; Rule 2 **ATTRIBUTE** D (1.0,1.4] THEN Class=versicolor; Rule
6	Computer Science	(1.0,1.4] THEN Class=versicolor; Rule 3 **ATTRIBUTE** B (3.8,4.2] AND ATTRIBUTE D (1.0,1.4] THEN
7	Computer Science	Class=versicolor; Rule 3 ATTRIBUTE B (3.8,4.2] AND **ATTRIBUTE** D (1.0,1.4] THEN Class=virginica; These
8	Computer Science	attributes in a system called the Windsor **Attribute** Grammar Programming Environment (W/ AGE
9	Computer Science	following, where Es stands for entity set: data **Attribute** SENT_VAL Bool | NOUNCLA_VAL Es | ADJ_VAL
10	Computer Science	Entity r Event r Attribute and Quantity r **Attribute** Value and Quantity Value Most nominal concepts
11	Economics	of target information. Advertising New **Attribute** Information Consider the case of a marketer
12	Engineering	4.1 Efficient Gerrymandering on a Single **Attribute** 4.1.1 Efficient Partitioning Given a coming
13	Linguistics	e. g., kamennyj'stone') > Typing **Attribute** (e. g., severnyj'northern') Speakers'
14	Philosophy	numbers and pure sets). 4.2 Attributes and **Attribute** Kinds Attributes have been classified by
15	Social Sciences	Severity of Abuse in the Relationship; **Attribute** 1: Blame Attributed to Defendant; Attribute
16	Social Sciences	Attribute 1: Blame Attributed to Defendant; **Attribute** 2: Legal Responsibility of the Defendant
17	Social Sciences	of the Defendant Under the Law as It is; **Attribute** 3: Legal Responsibility of the Defendant
18	Social Sciences	Defendant Under the Law as It Should Be; **Attribute** 4: Guilt of the Defendant; Attribute 5:
19	Social Sciences	Be; Attribute 4: Guilt of the Defendant; **Attribute** 5: Reasonableness of the Force Used by

Table 118. DOAE: Recording basic information - concordance lines of ATTRIBUTE tagged with the JJ tag.

1	Business and Management	preferences of questionnaire designers. Fig. 2. **Attribute** > definition window. When IDS detects that
2	Business and Management	belief degrees. If needed, pressing the **Attribute** Definition button will popup the display
3	Business and Management	claims--search, experience. Category and **Attribute** Selection The three product categories
4	Computer Science	properties for validity as follows. (1) **Attribute** validity: the entity (ontology) has the
5	Computer Science	properties for validity as follows. (1) **Attribute** validity: the entity (ontology) has the
6	Computer Science	properties for validity as follows. (1) **Attribute** validity: the entity (ontology) has the
7	Computer Science	properties for validity as follows. (1) **Attribute** validity: the entity (ontology) has the
8	Computer Science	properties for validity as follows. (1) **Attribute** validity: the entity (ontology) has the
9	Computer Science	Event category, whereas adjectives are **Attribute** Values. The sememes in each category are
10	Psychology	student would be more questionable. Results **Attribute** rankings Participants showed a preference

Note: Concordance lines 4-8 are from the same Computer Science text, and while it may appear that this is the case of the same concordance line being displayed five times, each of the five concordance lines is in fact from a different part of the text.

18. APPENDIX 8: GRAMMATICAL RELATIONS IN WORD SKETCH

Table 119. Sketch Engine: Word Sketch - Codes and components for grammatical relations for noun headwords.

Code	Component(s)
A_subj	headword as subject + adjective used predicatively
AJ_premod	adjective as premodifier
AJP	adjective phrase
AJ_pert	pertainym (adjective meaning 'pertaining to X', never predicative)
and_or	phrase with and/or (e.g. *tea and coffee*)
AVP_post_mod	adverb phrase as post-modifier of headword
cl_(that)	indicative clause without *that*
cl_if	whether/if clause
cl_that	indicative clause with *that*
cl_that_cond	conditional clause with *that*
cl_that_subj	subjunctive clause with *that*
cl_wh	clause with *what, when, how, where, why*
it+	anticipatory 'it' construction
N_mod	headword modified by another noun
N_premod	headword as pre-modifier of another noun
object_of	verb with headword as object
Part_*specific*_post-mod	named particle as post-modifier of headword
possessor	modified by a noun in possessive case (e.g. *ruler* in *ruler's authority*)
PP_cl_wh	prepositional phrase + wh-clause
PP_NP_Ving	prepositional phrase + noun phrase with gerund
PP_obj_*specific*-i	object + prepositional phrase with named preposition
PP_PP_*specific*-i	prepositional phrase + prepositional phrase with named preposition
PP_Ving	prepositional phrase with gerund as object
PP_*for* Ving-to	the *for* + infinitive-*to* construction
PP_*specific*	prepositional phrase with named preposition, e.g. PP-*at,* PP-*by*, etc.
predicate	noun phrase with headword as predicate
predicate_*specific*	noun phrase with named preposition as predicate
premod_AJ	headword as pre-modifier of an adjective
premod_N	headword as pre-modifier of another noun
Supp_PP_*specific*	headword is object of named ('support') preposition
unary rels	unary relations (e.g. passive)
V_subj	headword as subject of a verb
Vinf-to	infinitive verb with *to*
Ving	gerundive verb phrase

Table 120. Sketch Engine: Word Sketch - Codes and components for grammatical relations for verb headwords.

Code	Component(s)
AJP	adjective phrase
and_or	phrase with and/or (e.g. *give and take*)
AVP	adverb phrase
AVP_mod	adverbial phrase as modifier
cl_(that)	indicative clause without *that*
cl_(that)_cond	conditional clause without *that*
cl_(that)_subj	subjunctive clause without *that*
cl_if	whether/if clause
cl_that	indicative clause with *that*
cl_that_cond	conditional clause with *that*
cl_that_subj	subjunctive clause with *that*
cl_wh	clause with *what, when, how, where, why*
it+	anticipatory 'it' construction
NP	all types of noun phrase
NP AVP	noun phrase + adverb phrase
NP cl_(that)	noun phrase + indicative clause without *that*
NP cl_that	noun phrase + indicative clause with *that*
NP NP	noun phrase + noun phrase
NP Vinf	noun phrase + infinitive verb without *to*
NP Vinf-to	noun phrase + infinitive verb with *to*
NP Ving	noun phrase + gerund
NP_AJP	noun phrase + adjective phrase
NP_part	noun phrase + particle
NP_PP	noun phrase + prepositional phrase
part_intrans	particle with verb used intransitively
part_NP	particle + noun phrase
Part_*specific*_x_obj	named particle + object
Part_*specific*	named particle
Part_*specific* NP	named particle + noun phrase
Part_*specific* PP-*specific*	named particle + preposition phrase with named preposition
PP_cl_wh	prepositional phrase + wh-clause
PP_NP_Vinf_to	prepositional phrase + noun phrase with infinitive with *to*
PP_NP_Ving	prepositional phrase + noun phrase with gerund
PP_PP_*specific*-i	prepositional phrase + prepositional phrase with named preposition
PP_Ving	prepositional phrase with gerund
PP_*specific*	prepositional phrase with named preposition (e.g. PP_*at*, etc.)

PP_*specific* cl_wh	prepositional phrase with named preposition + wh-clause
PP_*specific* Vinf-to	prepositional phrase with named preposition with infinitive with *to*
pro_object	pronoun as object
pro_subject	pronoun as subject
quote	quote
subj_NP	noun phrase as subject
unary rels	unary relations (e.g. passive)
Vinf	infinitive verb without *to*
Vinf-to	infinitive verb with *to*
Ving	gerund
Wh Vinf-to	wh-word with infinitive with *to*

Table 121. Sketch Engine: Word Sketch - Codes and components for grammatical relations for adjective headwords.

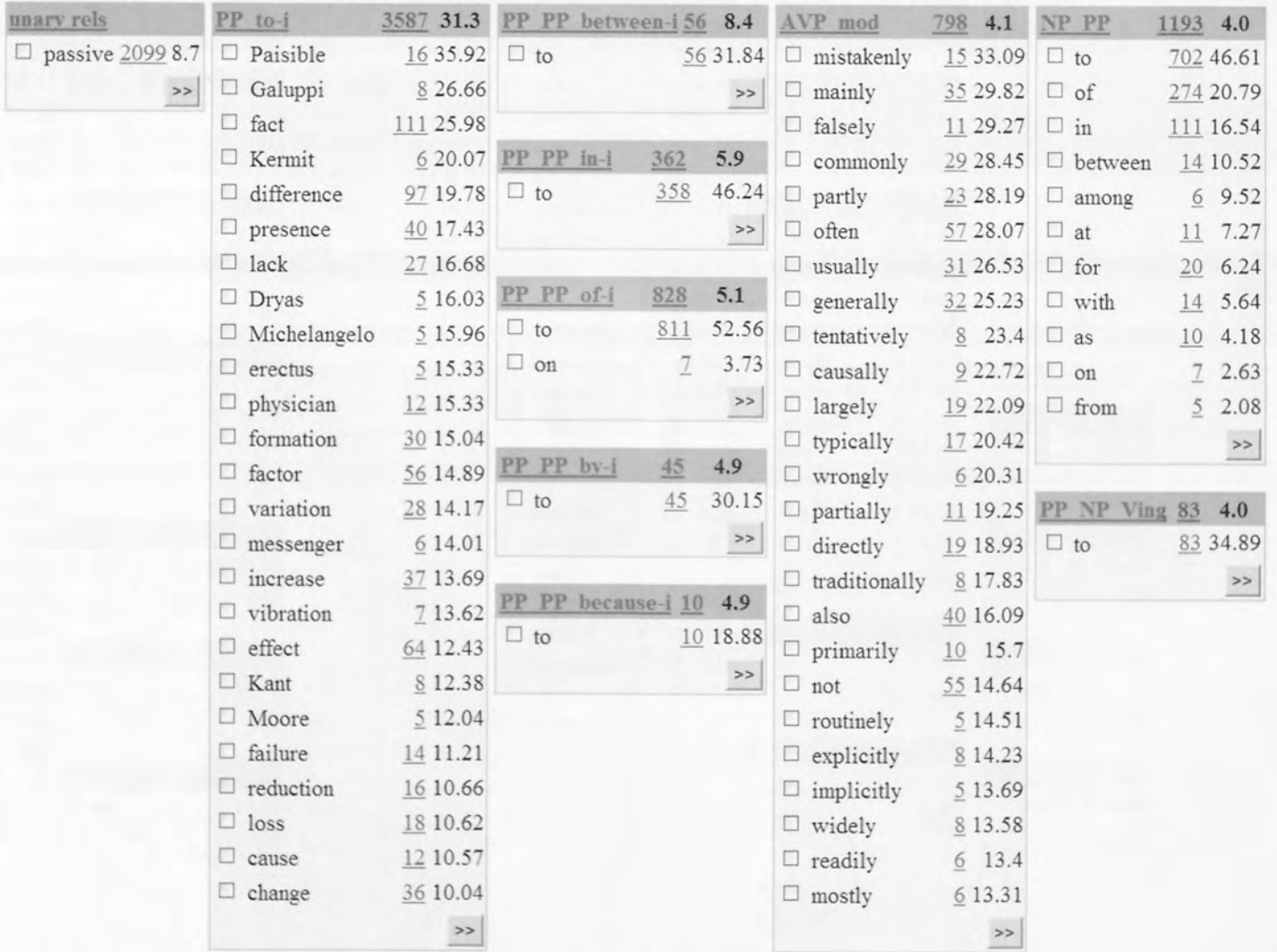
Code	Component(s)
AJP_premod	adjective phrase as pre-modifier
and_or	phrase with and/or (e.g. *big and small*)
AVP_premod	adverb phrase as pre-modifier
cl_(that)	indicative clause without *that*
cl_if	whether/if clause
cl_that	indicative clause with *that*
cl_that_subj	subjunctive clause with *that*
cl_wh	clause with *what, when, how, where, why*
comp_V	headword as complement after link verb
comp_V_NP	headword in noun phrase as complement after link verb
N_premod	noun as pre-modifier
PP_NP_Ving	prepositional phrase + noun phrase with gerund
PP_PP_*specific*-i	prepositional phrase + prepositional phrase with named preposition
PP_Ving	prepositional phrase with gerund
PP_*for* Ving-to	the *for* + infinitive-*to* construction
PP_*specific*	prepositional phrase with named preposition
PP_*specific* cl-wh	prepositional phrase with named preposition + wh-clause
premod_N	headword as pre-modifier of a noun
unary rels	unary relations (e.g. passive)
Vinf-to	verb phrase infinitive verb with *to*
Ving	gerundive verb phrase

Table 122. Sketch Engine: Word Sketch - Codes and components for grammatical relations for adverb headwords.

Code	Component(s)
AJP	modifying an adjective phrase
AVP	modifying an adverb phrase
AVP_premod	adverb phrase as pre-modifier
CL	modifying a clause (or sentence)
NP	modifying by noun phrase
PP	modifying a prepositional phrase
premod_VP	pre-modifying a verb
unary rels	unary relations (e.g. passive)
VP	modifying a verb

19. APPENDIX 9: MEANING ANALYSIS – TABLES AND FIGURES

Figure 99. DOAJ: Meaning analysis – Word sketch of *attribute* (verb) (page 1).



unary rels
☐ passive 2099 8.7
>>

PP_to-i	3587	31.3
☐ Paisible	16	35.92
☐ Galuppi	8	26.66
☐ fact	111	25.98
☐ Kermit	6	20.07
☐ difference	97	19.78
☐ presence	40	17.43
☐ lack	27	16.68
☐ Dryas	5	16.03
☐ Michelangelo	5	15.96
☐ erectus	5	15.33
☐ physician	12	15.33
☐ formation	30	15.04
☐ factor	56	14.89
☐ variation	28	14.17
☐ messenger	6	14.01
☐ increase	37	13.69
☐ vibration	7	13.62
☐ effect	64	12.43
☐ Kant	8	12.38
☐ Moore	5	12.04
☐ failure	14	11.21
☐ reduction	16	10.66
☐ loss	18	10.62
☐ cause	12	10.57
☐ change	36	10.04
>>		

PP_PP_between-i	56	8.4
☐ to	56	31.84
>>		

PP_PP_in-i	362	5.9
☐ to	358	46.24
>>		

PP_PP_of-i	828	5.1
☐ to	811	52.56
☐ on	7	3.73
>>		

PP_PP_by-i	45	4.9
☐ to	45	30.15
>>		

PP_PP_because-i	10	4.9
☐ to	10	18.88
>>		

AVP_mod	798	4.1
☐ mistakenly	15	33.09
☐ mainly	35	29.82
☐ falsely	11	29.27
☐ commonly	29	28.45
☐ partly	23	28.19
☐ often	57	28.07
☐ usually	31	26.53
☐ generally	32	25.23
☐ tentatively	8	23.4
☐ causally	9	22.72
☐ largely	19	22.09
☐ typically	17	20.42
☐ wrongly	6	20.31
☐ partially	11	19.25
☐ directly	19	18.93
☐ traditionally	8	17.83
☐ also	40	16.09
☐ primarily	10	15.7
☐ not	55	14.64
☐ routinely	5	14.51
☐ explicitly	8	14.23
☐ implicitly	5	13.69
☐ widely	8	13.58
☐ readily	6	13.4
☐ mostly	6	13.31
>>		

NP_PP	1193	4.0
☐ to	702	46.61
☐ of	274	20.79
☐ in	111	16.54
☐ between	14	10.52
☐ among	6	9.52
☐ at	11	7.27
☐ for	20	6.24
☐ with	14	5.64
☐ as	10	4.18
☐ on	7	2.63
☐ from	5	2.08
>>		

PP_NP_Ving	83	4.0
☐ to	83	34.89
>>		

Figure 99. DOAJ: Meaning analysis – Word sketch of *attribute* (verb) (page 2).

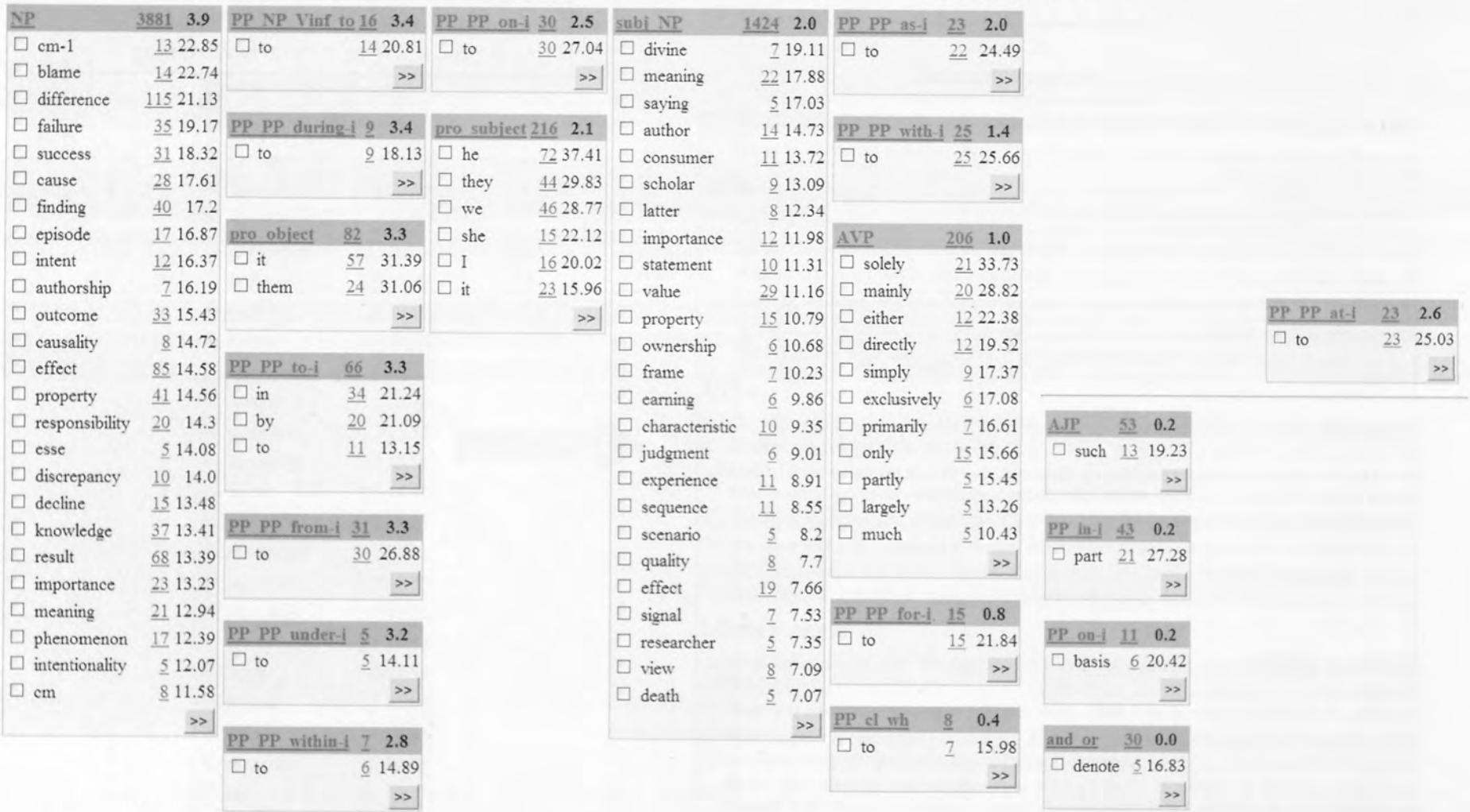


Figure 100. Model for DOAE: 5-step process of importing collocates and examples from CAJA (in Sketch Engine) to the DOAE database.

Step 1: Selecting collocates within grammatical relation in Word Sketch

PP to-i	3587	31.3
Paisible	16	35.92
Galuppi	8	26.66
fact	111	25.98
Kermit	6	20.07
difference	97	19.78
presence	40	17.43
lack	27	16.68
Dryas	5	16.03
Michelangelo	5	15.96
erectus	5	15.33
physician	12	15.33
formation	30	15.04
factor	56	14.89
variation	28	14.17
messenger	6	14.01
increase	37	13.69
vibration	7	13.62
effect	64	12.43
Kant	8	12.38
Moore	5	12.04
failure	14	11.21
reduction	16	10.66
loss	18	10.62
cause	12	10.57
change	36	10.04

>>

Step 2: Selecting examples

The difference in elution time of the sHSP peak relative to Luc is attributed to the presence of free
2) may be attributed to the presence of undifferentiated neuronal precursor cells as was demonstra

lack

Its poor agricultural productivity is attributed to the lack of proper irrigation facilities.
This may be partly attributed to the lack of high-resolution records in the region, as well as the prec
These have been attributed mainly to the lack of regulatory supervision and a consequent failure to
However, the dearth of primary research material is partly attributed to the lack of a suitable meth
This defect could not be attributed to the lack of sulfatide because fucoidan administration could nc
This failure was attributed to the lack of a P5 pocket in the groove of PapH.

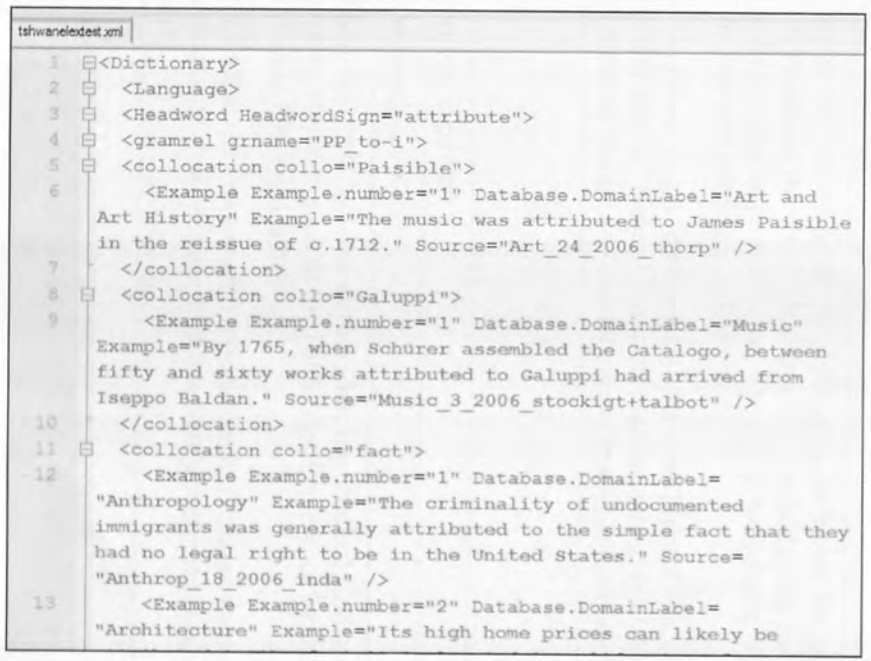
Dryas

The 14C dates from above and below this layer, the occurrence of the arctic-alpine molluscan speci Dryas cooling event (Dryas II) (GI-1d).
Most of these samples yielded rich insect assemblages, with the exception of samples P6 and P7 wh were probably sparse for environmental and climatic reasons, much larger samples would be needed
This finely laminated peat unit includes a loam-rich layer, locally overlain by a thin detrital layer that
The upper part of the peaty unit is strongly affected by cryoturbation (involutions and small local cra
Although no beetle fossils were found in the calcareous silts attributed to the Older Dryas, the pres ratio) suggests a significant cooling in the temperature.

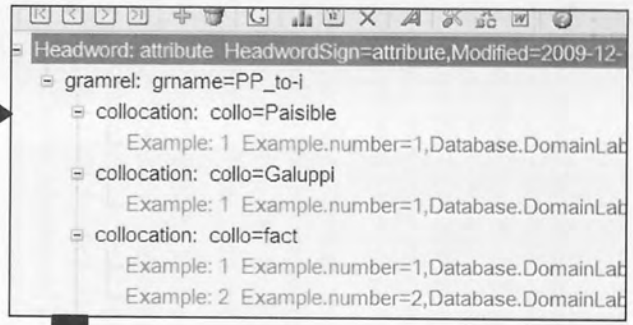
Michelangelo

Aside from the casts after Michelangelo's sculptures, only one other sculpture is given an attribution of 24 reales).
There was also a large painting of St Francis 'with a skull at his feet and a cross above it', attribute
Most noteworthy among these bronzes are a small Christ figure attributed to Michelangelo, a gilde two-figure sculptural group of a man and woman embracing.
In fact, there are some very important items missing from the 1613 inventory that were present in th Raphael, a St Anthony by Girolamo Muziano, Parmigianino's Holy Family Bergomasco's Perseus ca
This painting is very likely the large painting of St Francis listed in the 1613 inventory, attributed to

Step 3:
Importing collocates and examples
from Clipboard into XML document



Step 4:
Importing XML document
into TshwaneLex database



Step 5:
Copying collocates and examples
into database entry's meaning analysis
and adding notes

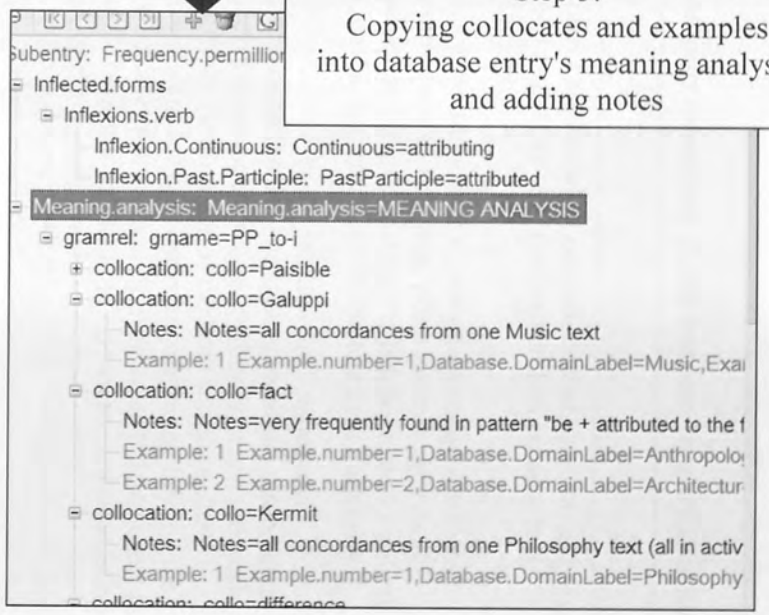


Table 123. DOAE: Initial seven meanings of *authority* (identified in 300 random concordance lines).

Meaning 1	power of control (of somebody)
Examples	The fallacy of appealing to authority happens when one claims that something is true or right simply because someone in **authority** says it is, rather than because it is supported by evidence or logical reasons. There is a dilemma here that needs to be addressed: In some lessons we observed, girls ' classes were different in atmosphere and outlook; often lessons were more self-sustained by the girls themselves, it was possible for the teacher to adopt a more relaxed style, and there was less tension and challenge to the teacher's **authority**. Even though the Methodist quadrilateral strives to maintain the balance and equilibrium amongst the four sources of **authority** - scripture, tradition, reason and experience - when there is conflict between the sources of authority, when the work of the Holy Spirit today goes against the instructions of the holy scripture, the Methodist Movement has leaned and leans towards experience, moving beyond the perceived normative position of the Bible: from accepting women as lay preachers in the eighteenth century - an unheard of novelty for the time - down to accommodating the ministry of homosexual people at the present time.

Meaning 2	a person with power or control
Examples	In the plot of a typical novel a wife/daughter achieves her goals by allowing a husband/father to think that an action that resulted in his loss of control in the home had been his idea in the first place. The use of subversive female wiles to outwit male **authorities** was not and is still not new. Thus Gombaud might represent and share the highest temporal **authority** in the duchy and also be 'bishop' - in itself nothing unique but with authority also understood as 'ducal' and overlapping with that of his brother.

Meaning 3	institutions in power
Notes	predominantly in plural
Examples	Children were sent by local **authorities**, the criminal justice system, poor law authorities, charities such as the Fairbridge Society, or by their parents (sometimes intending to join them later). It follows that the Oskarshamn model - even if applied meticulously - may not lead to the requisite trust where attitudes toward government and **authority** are quite different.

Meaning 4	a person or thing that is often referred to, and can be trusted (has a lot of knowledge on a certain subject)
Notes	mainly in singular
Examples	Significantly, he referred to the **authority** of Hume - the one British thinker who "can be numbered among the philosophes" of the French Enlightenment - without any strictures. There is no direct Scottish **authority** on the point, but, subject to one qualification, it is likely that Scots law would follow the authority of Roman-Dutch law in relation to a similar rule. That Dvořák had heard and studied Herbert's "spendid" concerto in the interim - in March 1894 to be exact - would suggest that the work played some part in his decision to write a cello concerto of his own; Jan Smaczny, an **authority** on the Dvořák concerto, even refers to the experience as "the road to Damascus."

Meaning 5	organization or department
Notes	Predominantly in upper case, and part of a proper noun (written with a capital).
Examples	Food consumption is deflated by a food price deflator, using regional prices collected by the Central Statistical **Authority**. This case study involved interviews with the officials who compiled the claim from North Yorkshire County Council (NYCC) and its seven District Councils, one Unitary Authority (York) and the police **authority** within the area of NYCC's jurisdiction.

Meaning 6	main or dominating feature
Notes	Found in modifier use. Perhaps a technical meaning (Computing)?
Examples	Its flow characteristic is peer-to-peer, i.e., there is no **authority** node in the network.

Meaning 7	an internet page that has many citations pointing to it
Notes	Definition obtained from the example. Computing label needed.
Examples	**Authorities** are pages that have many citations pointing to them, whereas hubs represent pages that have a lot of outgoing links.

Table 124. DOAE: Meaning analysis – Groupings of most salient collocates in the grammatical relation 'AJ_premod' of *authority* (noun).

RELATION	COLLOCATE(S)	MEANING(S)	NOTES
AJ_premod	*lawmaking, sovereign, decisional*	Meaning 1	singular uses of *authority* only
	monetary, central	Meaning 3	*authority* found in singular and plural
	fiscal		*authority* mostly in singular
	local, political, public, legitimate, colonial, national, religious, governmental, judicial, statutory, legal	Meanings 1 and 3	*authority* (singular) used in meanings 1 and 3, *authorities* (plural) used in meaning 3 only
	municipal, traditional		singular and plural uses of *authority* in both meanings
	civic, royal, ecclesiastical		*authority* (singular) in meaning 1, *authorities* (plural) in meaning 3
	parental	Meanings 1 and 2	*authority* (singular) in meaning 1, *authorities* (plural) in meaning 2
	supreme	Meanings 1, 2, 3	singular uses of *authority* only

Table 125. DOAE: Meaning analysis – Groupings of collocates in the grammatical relation 'object_of' of *AUTHORITY* (noun).

RELATION	PATTERN ELEMENT	COLLOCATE(S)	MEANING(S)	NOTES
object_of	Activity (of somebody with authority)	*delegate, exercise, devolve, grant, wield, claim, vest, assert, confer, lend, cede*	Meaning 1	*authority* always in singular
		give, provide	Meanings 1, 3	*authority* (singular) in meaning 1, *authorities* (plural) in meaning 3
	Activity (of somebody without authority)	*undermine, lack, usurp, justify, legitimize*	Meaning 1	*authority* always in singular
		challenge	Meanings 1, 3	*authority* more often found in singular than in plural
		obey, disobey	Meanings 1, 2, 3	singular and plural uses of *authority*
		recognize	Meanings 1, 4	*authority* mainly in singular
	Possession State	*have, possess*	Meaning 1	*authority* always in singular
	State	*be*	Meanings 1, 2, 3, 4	83% of uses in singular

Table 126. DOAE: Meaning analysis - Salient grammatical relations, pattern elements and collocates of Meaning 1 of *authority* (noun).

MEANING 1		
RELATION	**Pattern element**	**COLLOCATE(S)**
AJ_premod		*political, religious, public, legitimate, legal, traditional, royal, parental, sovereign, lawmaking, decisional, supreme, municipal, civic, ecclesiastical, national*
object_of	Activity (of somebody with authority)	*delegate, exercise, grant, claim, give, provide, devolve, wield, vest, assert, confer, lend, cede*
	Activity (of somebody without authority)	*challenge, undermine, recognize, lack, obey, disobey*
	Possession State	*have, possess*
V_subj		*do, rest, have, grant, establish*
PP_obj_*of*-i		*exercise, source, position, form, devolution, delegation*
N_mod		*state, enforcement*
premod_N		*figure, relation, relationship, structure*
PP_obj_*to*-i		*appeal, claim, submit, challenge*
PP_of-i	Institution	*church, council, state, court*
	Human | Sovereign	*king, teacher, pope, monarch*
	Entity | Object	*scripture, law*
possessor	Institution	*state, congress*
	Human | Sovereign	*ruler*
	Entity | Object	*law*

Table 127. DOAE: Meaning analysis - Salient grammatical relations, pattern elements and collocates of Meaning 2 of *AUTHORITY* (noun).

MEANING 2		
RELATION	**Pattern element**	**COLLOCATE(S)**
AJ_premod		*parental, supreme*
object_of	State	*be*

Table 128. DOAE: Meaning analysis - Salient grammatical relations, pattern elements and collocates of Meaning 3 of *AUTHORITY* (noun).

MEANING 3		
RELATION	**Pattern element**	**COLLOCATE(S)**
AJ_premod		*local, political, public, religious, traditional, national, central, monetary, legitimate, colonial, civic, municipal, royal, judicial, ecclesiastical, governmental, statutory*
object_of	State	*be*
	Activity (of somebody without authority)	*challenge, obey, disobey*
V_subj		*have, exercise, grant, establish, do, decide, issue, try, respond, ban, tolerate*
N_mod		*state, planning, government, health, education, tax, enforcement, police*
premod_N		*housing, planner*
PP_obj_*to*-i		*claim, submit, challenge, appeal*
PP_obj_*by*-i		*arrest, approve, back, issue, administer*

Table 129. DOAE: Meaning analysis - Salient grammatical relations, pattern elements and collocates of Meaning 4 of *AUTHORITY* (noun).

MEANING 4		
RELATION	**Pattern element**	**COLLOCATE(S)**
AJ_premod		*parental, supreme*
object_of	Activity (of somebody without authority)	*recognize*
	State	*be*
PP_obj_*of*-i		*position*

Table 130. DOAE: Analysis of compound candidates of *potential* (noun) in two Word Sketch relations.

grammatical relation	dictionaries + Wikipedia	explanation of meaning by corpus examples
AJ_premod		
endocochlear potential	Not found in any dictionary; definition found on the web http://oto2.wustl.edu/cochlea/ep.htm	corpus examples mention how it is generated; one example uses *called* but the definition is not totally clear
electrostatic potential	*electrostatic* found in some dictionaries (MWCD CD-ROM, NODE CD-ROM), but not the phrase; definition found on Wikipedia: http://en.wikipedia.org/wiki/Electrostatics	Nothing in the corpus (note: *be* is normally found as a collocate, but is only mentioned if something relevant is found)
event-related potential	nothing in any dictionary; definition found on Wikipedia: http://en.wikipedia.org/wiki/Event-related_potential	nothing in the corpus
mitochondiral membrane potential	part of *membrane potential* (see below) so not candidate entry (also nothing in corpus or dictionaries)	
resting (membrane) potential	definition in NODE CD-ROM and in Dictionary.com (Medical Dictionary + Encyclopaedia); Wikipedia mentions *resting* and *action potential* as *resting membrane potential* and *action membrane potential*	nothing in the corpus
electric potential	found in CED CD-ROM (used here) and Dictionary.com – appears to mean both *potential* and *action potential* – so need to have its own entry + symbol is V Wikipedia suggest it means the same as *electrostatic potential*	one example with *define*, but not explaining the compound
metastatic potential	nothing in dictionaries, nor directly on the web – *metastatic potential* simply means that something has potential to be metastatic (to spread to other organs), so it is a transparent compound and does not even belong to meaning 2 (it was moved)	nothing in the corpus
chemical potential	found as an entry in CED CD-ROM, NODE CD-ROM, and Dictionary.com (label tends to be Thermodynamics or Chemistry); also found on Wikipedia - http://en.wikipedia.org/wiki/Chemical_potential	μ is a symbol for chemical potential a few examples with *define, definition* found, but do not provide any definition of meaning
osteogenic differentiation potential	whole phrase not found in any dictionary, and not even in Wikipedia also, found mainly in one Engineering text, so not that relevant for the dictionary derivative of the *differentiation potential* (see below)	

adiabatic potential	nothing in dictionaries, not online (with the exception of articles) -seems to be transparent, as *adiabatic* means without loss or gain of heat (so *adiabatic potential* suggest something has potential not to lose or gain any heat)	nothing in the corpus
interatomic potential	nothing in dictionaries, not online (with the exception of articles) -partly transparent, but also highly technical so perhaps not an entry candidate	nothing in the corpus
electrochemical potential	nothing in dictionaries, found definition on Wikipedia: http://en.wikipedia.org/wiki/Electrochemical_potential -possible entry candidate, but perhaps too technical??	nothing in the corpus
postsynaptic potential(s)	nothing in dictionaries, only on Wikipedia: http://en.wikipedia.org/wiki/Postsynaptic_potential -possible entry candidate, but perhaps too technical??	nothing in the corpus
long-range potential(s)	nothing in dictionaries, not only (with the exception of articles) not an entry candidate	nothing in the corpus
intermolecular potential(s)	nothing in dictionaries, not only (with the exception of articles) not an entry candidate	nothing in the corpus
effective potential	nothing in dictionaries, found on Wikipedia: http://en.wikipedia.org/wiki/Effective_potential	nothing in the corpus (technical meanings mixed with general ones)

N_mod		
redox potential(s)	only in Dictionary.com (medical dictionary), where it is mentioned that this is another term for *oxidation-reduction potential* interestingly, Wikipedia (and some online articles) suggest that both *redox potential* and *oxidation-reduction potential* (not in CAJA) are forms of *reduction potential*	nothing in the corpus
membrane potential	only in Dictionary.com (medical dictionary); also on Wikipedia (http://en.wikipedia.org/wiki/Membrane_potential) -high occurrence so candidate for an entry!	*define, describe* in the corpus but not explaining the meaning
action potential(s)	found in CED CD-ROM, NODE CD-ROM, MWCD CD-ROM and Dictionary.com as an entry Also on Wikipedia: http://en.wikipedia.org/wiki/Action_potential -frequent, so entry candidate	*Applying a sufficiently strong stimulus to a cardiac cell leads to a prolonged elevation of transmembrane voltage v known as an* ***action potential****.*

zeta potential	entry in NODE CD-ROM, and found on Wikipedia: http://en.wikipedia.org/wiki/Zeta_potential -not a candidate due to low frequency?	*Another important characteristic of nanoparticles is the overall charge on the nanoparticle surface, namely* ***zeta potential****.*
Born-Oppenheimer potential(s)	nothing in dictionaries, nothing on Wikipedia (just in some online articles) -not a candidate due to all examples from a single text	mainly in plural, examples from a single text
quantum potential	nothing in dictionaries, nothing on Wikipedia (just in some online articles) -perhaps not a candidate due to most examples from a single text	*The* ***quantum potential*** *is the potential energy function of the wave field.* + one other example
electrode potential(s)	entry in CED CD-ROM, and on Wikipedia: http://en.wikipedia.org/wiki/Electrode_potential -78 hits – potentially an entry candidate	nothing in the corpus
Tersoff potential	nothing in dictionaries, not an entry on Wikipedia but mentioned -not a candidate as the examples come from a single text	nothing in the corpus
differentiation potential	not in dictionaries or Wikipedia -there are many types of differentiation potential; most examples from a single text, so not a candidate?	nothing in the corpus
reduction potential(s)	found in Dictionary.com and on Wikipedia: Wikipedia says it is the same as *redox potential* or *oxidation/reduction potential* http://en.wikipedia.org/wiki/Reduction_potential	nothing in the corpus
oxidation potential	found in Dictionary.com and on Wikipedia: Wikipedia (redirects to Reduction potential) says it is the same as *redox potential* or *oxidation/reduction potential* http://en.wikipedia.org/wiki/Reduction_potential -if *reduction potential* is made an entry, so should *oxidation potential*	nothing in the corpus

Figure 101. DOAE: Meaning analysis – Word sketch of *assortment* (noun).

assortment **Corpus of Academic Journal Articles (CAJA) freq = 346**

object_of	79	1.8
offer	7	15.86
skew	2	11.82
perceive	3	9.82
utilize	2	8.62
perform	3	7.86
present	3	7.59
consider	3	6.97
include	3	6.54
set	2	5.9
compare	2	5.44
associate	2	5.26
use	3	3.99
make	2	3.87
be	7	2.39
have	2	2.1

PP_on-i	28	9.8
choice	16	29.65
preference	5	17.12
strength	4	15.1
difficulty	2	9.32

PP_NP_Ving	4	4.0
of	4	9.37

PP_of-i	86	3.8
merchandise	3	17.46
product	5	11.34
task	3	8.25
good	2	7.53
group	2	3.84

AJ_premod	157	2.8
large	55	34.83
small	28	26.85
wide	14	25.02
positive	8	15.02
rich	3	11.23
odd	2	10.08
retail	2	9.8
usual	2	9.24
independent	3	8.43
diverse	2	8.03
great	3	7.15
unique	2	6.98
various	2	5.78
different	2	3.49

PP_PP_with-i	2	2.4
of	2	6.4

PP_PP_in-i	5	1.7
of	4	8.86

AVP_post_mod	6	1.7
carefully	2	14.07

V_subj	38	1.2
weaken	2	11.82
affect	4	11.69
lead	4	11.41
strengthen	2	10.9
arise	2	8.8
be	14	8.68
do	2	4.65

PP_obj_in-i	10	1.0
alternative	2	11.18

PP_obj_from-i	43	20.2
choice	21	31.65
choose	12	24.25
make	9	15.12

PP_PP_to-i	3	3.1
of	2	5.76

PP_obj_of-i	62	2.7
impact	27	33.01
diversity	5	16.59
quantity	2	8.7
context	3	8.43
consist	2	7.35
role	2	5.58
term	2	5.45
function	2	4.97
effect	2	4.21

and_or	77	1.3
availability	9	23.95
consumer	6	17.23
respondent	5	15.28
quality	6	14.21
articulation	2	10.72
G.	4	10.21
pricing	2	9.7
category	3	9.04
impact	3	8.9
service	3	8.41
choice	2	6.6
price	2	6.16
effect	3	5.69
line	2	5.63
point	2	4.92

N_mod	49	0.9
product	34	35.23
category	5	14.17
merchandise	2	14.09
chocolate	2	13.9

A_subj	10	2.9
likely	3	11.71
significant	2	8.66

PP_PP_of-i	7	0.9
on	4	12.67
of	2	4.41

premod_N	27	0.2
discrepancy	2	11.74
manipulation	2	10.79
set	3	9.46
variety	2	8.96
decision	2	7.74
size	2	7.04

Table 131. DOAE: Meaning analysis - Top 48 collocates of *albeit*, span 0+5 (ordered by MI3 value).

	Freq	T-score	MI	MI3
,	662	21.779	2.703	21.445
a	420	18.961	3.741	21.169
in	386	17.683	3.323	20.507
different	132	11.322	6.105	20.194
with	251	15.008	4.246	20.189
one	127	10.967	5.220	19.197
the	365	13.360	1.734	18.757
less	74	8.494	6.317	18.736
not	147	11.506	4.292	18.692
.	307	13.525	2.132	18.657
)	232	13.230	2.928	18.644
of	289	12.736	1.995	18.345
at	110	9.887	4.126	17.689
limited	39	6.193	6.901	17.472
lesser	19	4.350	8.861	17.357
than	81	8.637	4.632	17.311
very	55	7.278	5.748	17.311
lower	46	6.691	6.215	17.262
indirectly	18	4.233	8.823	17.163
reluctantly	7	2.645	11.408	17.023
somewhat	24	4.873	7.575	16.745
that	131	9.532	2.581	16.647
to	159	9.387	1.968	16.594
slightly	23	4.765	7.268	16.315
more	65	7.596	4.113	16.158
and	152	8.190	1.575	16.071
varying	18	4.222	7.709	16.049
degrees	17	4.104	7.732	15.907
weaker	14	3.729	8.167	15.782
for	95	8.065	2.535	15.675
ones	20	4.436	6.961	15.605
small	31	5.443	5.475	15.384
inefficiently	4	1.999	11.354	15.354
ways	25	4.925	6.054	15.342
unintentionally	5	2.235	10.676	15.320
weakly	11	3.306	8.295	15.214
redundantly	3	1.732	12.024	15.194
some	42	6.169	4.376	15.161
only	43	6.175	4.098	14.951
form	31	5.398	5.036	14.945
smaller	19	4.307	6.382	14.878
reduced	21	4.512	6.013	14.798
from	61	6.783	2.926	14.788
weak	16	3.964	6.781	14.781
tentatively	6	2.446	9.580	14.750
without	27	5.058	5.237	14.747
marginally	8	2.822	8.725	14.725
by	65	6.682	2.546	14.591

Table 132. DOAE: Types of change made to corpus examples.

Type of change	Example(s)
extra unnecessary context removed (e.g. preceding or following sentences or clauses, subordinate clauses, or titles)	• *There is no doubt that excessive lenience, derived at least in part from the defendants' condition as Indians, was behind the sentence of ten years banishment accorded to Captain Diego Andrés, who killed his wife with a club in Oaxaca in 1747; the pardon of Lorenzo Macapa, who, as discussed above, brutally kicked his wife and their unborn child to death near a California mission in 1776; the two-year exile of José Tomás Mendoza after he stoned his wife to death in central Mexico in 1805; or the pardon handed down by the ordinarily severe Acordada judges to Bernardino Antonio, a resident of Actopan, who beat his wife fatally in 1808 after they* ***argued*** *about the punishment of their child.* • *Traffic flow simulation The traffic flow data were* ***obtained*** *using the simulation and assignment of traffic in urban road networks (SATURN) traffic flow simulation software.* • *The leading Sydney newspapers, including the Sun and Daily Telegraph,* ***featured*** *full-page souvenir photographs of Sydney and its harbour from the air.*
text omitted because to avoid missing (irrelevant) context	*However, in August 2002 the process was stopped because the main retail chains did not want to incorporate these new products into their product* ***assortment****.*
Some academic writing conventions removed (mainly footnote numbers, but also some references)	• *Early opinions* ***attributed*** *the toxic effects of Abeta to its aggregates, whereas, more recently, oligomer participation in AD pathology has been generally accepted [22].* • *Ethical approval was* ***obtained*** *before the beginning of experiments (621-2531.31-20/ 01, Government of Lower Franconia, Würzburg, Germany).*
extra context added for clarity (normally, pronouns or other anaphora replaced with the original reference)	• *The system* replaced with *Green Building Rating System* (found in the preceding sentence)*:* *The Green Building Rating System awards points for design* ***features*** *that cover site development, water savings, energy efficiency, materials selection, and indoor environmental quality.* • *That year* replaced with *In 1938* (found in the preceding sentence)*:* *In 1938 the British government passed a Bill ruling that Australian films no longer counted as 'British' for their local quota,* ***thus*** *making them less attractive to British distributors.* • Added missing location *in Tunisia* (found in the preceding sentence)

	*One can **argue** that the management of economic change in Tunisia has been relatively good thus far.*
author's stress removed (e.g. single quotes)	*Those objects most highly valued by museums and collectors, and those which have received a great deal of scholarly interest, are often **attributed** to a particular 'artist' or 'school'.*
some (less known) abbreviations and acronyms replaced, omitted, or given in full	• Replacing the acronym *VHA* with *the agency* (phrase used in the preceding sentence). *In striving to make sense of the agency's closure, workers tended to **attribute** responsibility locally, blaming their employer for its betrayal and their union for its irrelevance or, worse, its damage.* • Omitting the acronym TPI (not standardised, specific to the article). *The textile processing industry (TPI) is regarded as a water intensive sector as it uses water as the principal medium for applying dyes and finishing agents and removing of impurities. The main environmental concern is **therefore** about the amount of water discharged and the chemical load it carries.*
some connective adverbial phrases removed	• *Before the Security Council Vietnam **justified** its intervention in Kampuchea principally on grounds of self-defence but also on humanitarian grounds.* • *First, it is not certain that we are **justified** in speaking of Middle Javanese as a single language.* • *Hence, we find that competition has a **significant** negative effect on firm performance.*
some evaluative comments removed	*But above all (so to speak), the most basic characteristic of religion is a belief in a higher being, a supreme otherworldly **authority** to whom ultimate allegiance is owed.*
combination of two or more types of change	*However, queries of this and a similar form did not lead to satisfactory results: As a consequence of Google's ranking mechanism, which prefers "**authorities**" (Brin and Page 1998), mainly portals of big organizations, companies, and others were retrieved.*
explanation added	***Thus** far, (=until this point in the article) we have reviewed two prototypical Western European stakeholder settings.*
obscure words and phrases removed	*Employees who are more dissatisfied with **various** aspects of their jobs are more likely to demand union representation, ceteris paribus.*

Key:

Items in bold (e.g. ***justified***) – headwords
Underlined items (*e.g. Hence,*) – omitted or added items

Table 133. Sketch Engine: Thesaurus: Synonym candidates for the noun *attribute* (sorted by similarity score)

Lemma	Score	Freq
trait	0.329	5535
characteristic	0.327	18244
feature	0.298	24979
dimension	0.294	13417
aspect	0.284	17518
quality	0.272	21825
indicator	0.267	7215
category	0.264	18925
criterion	0.261	12825
outcome	0.252	18029
phenomenon	0.25	9251
capability	0.249	4708
character	0.246	12696
meaning	0.244	14180
orientation	0.243	8262
entity	0.243	5624
objective	0.242	7969
motivation	0.241	6185
complexity	0.24	8454
concept	0.24	22271
location	0.239	13581
perception	0.238	11472
skill	0.238	12278
preference	0.238	10470
similarity	0.237	8148
choice	0.234	19145
goal	0.234	16536
circumstance	0.234	7195
name	0.234	11281
item	0.233	19688
representation	0.232	16473
dynamics	0.232	7823
benefit	0.232	13964
cue	0.232	3932
construct	0.231	4151
variable	0.23	40692
ability	0.229	20395
constraint	0.229	12733
strength	0.228	8937
kind	0.228	19003
configuration	0.227	5747
resource	0.227	19617
quantity	0.226	5724
property	0.226	30230
norm	0.226	8500
behaviour	0.225	10311
consideration	0.225	9870
element	0.225	25766
alternative	0.223	8131
topic	0.223	6745
option	0.222	10320
contribution	0.222	11226
diversity	0.221	5640
capacity	0.221	13681
intention	0.221	6672
combination	0.22	12052
requirement	0.22	12112
competence	0.22	3473
evaluation	0.22	11903
arrangement	0.22	5155

Table 134. Sketch Engine: Word Sketch: Synonyms of the noun *attribute* (circled) in grammatical relation 'and_or'.

and_or	945	1.0
☐ attribute	23	23.9
☐ governance	19	23.34
☐ element	33	20.47
☐ skill	22	19.87
☐ viz	5	18.35
☐ characteristic	33	18.1
☐ brand	12	17.44
☐ mode	14	15.28
☐ feature	17	14.08
☐ item	14	13.89
☐ tag	5	12.83
☐ quality	12	12.01
☐ name	7	10.4
☐ trait	5	9.57
☐ ability	8	9.44
☐ behavior	9	9.0
☐ experience	9	8.95
☐ criterion	6	8.5
☐ kind	6	7.68
☐ dimension	5	7.3
☐ level	13	6.97
☐ status	5	6.78
☐ value	12	6.67
☐ concept	5	6.21
☐ object	5	6.16

>>

Table 135. DOAE: Synonymy analysis of the first 10 synonym candidates of *attribute* (verb).

synonym candidate	estimated level of synonymy	Notes
RELATE	VERY LOW	-*relate* seems to be synonymous only to Meaning pattern 1 of the verb *attribute* – '*relate to* someone' has a completely different meaning than '*attribute to* someone' -shared syntactic patterns: *relate* sth *to* sth sth + *be* + *related to* sth -main differences: *relate* seems to imply than one thing is connected to another, which is not the sole cause of it, whereas *attribute* implies that one thing is a sole cause of another thing, or establishes a stronger connection between two things *attribute* contains a higher degree of author's involvement than *relate* (somebody can attribute something to another thing, but somebody cannot relate something to another thing) *relate* shows more absoluteness in its use – so, something is related (often *directly, closely, positively, negatively, significantly*), or is not related to another thing
REFER	VERY LOW	-mostly followed by *to*-prepositional phrase -has patterns with *to*, but the patterns are different from the ones found at *attribute*: sb *refer* sb *to* sth, or sb *is referred to* sth -*refer* is used more for describing things, or telling someone where something is (*referring* a reader to a part of the book)
ASCRIBE	HIGH	-mostly followed by *to*-prepositional phrase -shares many syntactic patterns with *attribute*, especially: a) sth + *be* + *ascribed* + *to* + sth/sb b) frequently occurs in the passive c) object 1 is a characteristic (e.g. *status, importance, value, property)*, or a finding/result (e.g. *difference, effect, result)* d) is often modified by adverb of frequency -it uses the pattern *ascribe* characteristic to something (similar to sense 3 at *attribute*) more often -while some its uses exhibit negative connotation, they are proportionally much less common than with *attribute*
LINK	LOW	-mostly followed by *to*-prepositional phrase -some patterns are similar to attribute, but the semantic properties of *link* are much closer to *relate* than *attribute* -less direct involvement from the author(s) – it is mainly used to report on previous findings
CONTRIBUTE	VERY LOW	-mostly followed by *to*-prepositional phrase -does not share many syntactic patterns with *attribute* -the main difference is that if something/sb contributes to something, it has played a role in creation/development of that thing – so, *to*-prepositional phrase contains the object that is ***the result of*** the action of the subject (as opposed to *attribute*, where *to*-prepositional phrase contains object 1 that is ***a cause of*** object 2)

ASSIGN	HIGH	-mostly followed by *to*-prepositional phrase -shares some syntactic patterns with *attribute*, especially the object 1 being a characteristic (e.g. *status, importance, value, property)*, or blame (e.g. *blame, responsibility)* -it has a much more neutral connotation -when a characteristic is being *assigned to* sth, it is being given to it, or you are saying that something has that characteristic
OWE	LOW	-mostly followed by *to*-prepositional phrase -semantically similar to *attribute*, but syntactic patterns are different; very frequent pattern is clause-beginning *owing* + *to* -shares with *attibute* some collocates that occur in *to*-prepositional phrase (e.g. *lack, fact, differences, changes, effect*) -the meaning displays less human involvement (author's voice)
LIMIT	VERY LOW	-often followed by *to*-prepositional phrase -completely different meaning – if something is limited to something, it occurs only in that case
RESTRICT	VERY LOW	-often followed by *to*-prepositional phrase -similar meaning to *limit*
EXPLAIN	LOW	-rarely followed by *to*-prepositional phrase -semantically similar to *attribute to* in the pattern *explained by*, when it also has similar collocates as object 1 (e.g. *differences, finding*) - this pattern represents slightly more than 10% of all occurrences of the verb *explain*

Table 136. DOAE: Identified missed meanings or patterns of sample entries, and action taken.

headword	missed meaning/pattern	corpus evidence	action
argue	to suggest, to demonstrate, to show it exists *The statement argues a change of attitude by the management.* (LED CD-ROM)	no examples found	NONE
	argue sb *into/out of (doing)* sth	no examples found	NONE
	BrE, informal *argue the toss*	no examples found	NONE
assortment	act of assorting	no examples found	NONE
	assortment of + [People]	no examples found	NONE
***attribute* (verb)**	to regard as produced by or originating in the time, period, place, etc., indicated *to attribute a work to a particular period* (Dictionary.com)	22 examples of *attribute to + period* found in CAJA	added a sense (sense 5)
***attribute* (noun)**	a material object recognized as symbolic of a person, especially a conventional object used in art to identify a saint or mythical figure (NODE CD-ROM)	6 examples found, but all from the same Arts and Art history text	not included in the entry, but examples saved and a note for an expert created
	Grammar an attributive adjective or noun (NODE CD-ROM)	7 examples from Linguistics found in CAJA	added a sense (sense 4)
	Logic the property, quality, or feature that is affirmed or denied concerning the subject of a proposition (CED CD-ROM)	found several related examples in Theology and Religion, and Philosophy concordance lines	added a sense (sense 5)
authority	permission to do sth *under the authority of somebody* *without sb's authority* *without the authority of somebody*	*under the authority of somebody* (67 examples in CAJA, 3 examples in BAWE, 5 examples in BASE) *without the authority of somebody* (2 examples in CAJA, 1 example in BAWE)	added a sense (sense 7) but without label *formal* as used in some learners dictionaries
	the power to influence other people	several examples found in all four corpora	added a sense (sense 8)
	confidence of having power to influence others *to speak with authority*	10 examples in CAJA 1 example in BAWE 1 example in MICASE	added as an example under sense 8, with pattern *speak with authority* in bold
	to have it on good authority	no examples found	NONE

headword	missed meaning/pattern	corpus evidence	action
et cetera	indicating that a list is too tedious or clichéd to give in full: *we've all got to do our duty, pull our weight, et cetera, et cetera.* (NODE CD-ROM)	1 example in CAJA 1 example in BAWE 33 examples in BASE 2 examples in MICASE	added as an example under *et cetera* with pattern *et cetera et cetera* in bold.
fact	*in actual fact*	37 examples in CAJA 18 examples in BAWE 13 examples in BASE 1 example in MICASE	replaced the pattern *in point of fact* due to higher frequency in spoken corpora
	a fact of life	21 examples in CAJA 3 examples in BAWE 1 example in MICASE 1 example in BASE	added a sense (sense 7) and a note about possible separate entry
	facts of life	1 example in BASE	NONE
	facts and figures	15 examples in CAJA 14 examples in BAWE 1 example in BASE	recorded in the database (not included in the entry due to transparent meaning)
	the fact remains	73 examples in CAJA, 19 examples in BAWE 1 example in BASE	added as a construction under sense 2 with an example
	and that's a fact (spoken)	3 examples in MICASE	recorded in the database
	is that a fact (spoken)	1 example in MICASE	NONE
	the facts speak for themselves	1 example in CAJA 1 example in MICASE 1 example in BASE	recorded in the database
***feature* (noun)**	a part of your face	several examples found, especially of phrase *facial feature:* 65 examples in CAJA 3 examples in BAWE 2 examples in BASE	added a sense (sense 2)
	newspaper or TV report that focuses on a particular subject	approximately 15 examples found in CAJA, all from Arts and Art History or Theology	added a sense (sense 3)
	a film of standard length	approximately 10 examples found in Arts and Art History texts alone	added a sense (sense 4)
	informal to imagine	no example found	NONE

headword	missed meaning/pattern	corpus evidence	action
justify	*Printing* to adjust space between words in a line	only few examples found	added a sense (sense 4) as it is believed that it will be useful to students
	Theology declare or make righteous in the sight of God (NODE CD-ROM)	found several examples in CAJA	added a sense (sense 5)
	Law to prove (a person) to have sufficient means to act as surety, etc., or (of a person) to qualify to provide bail or surety (CED CD-ROM)	no examples found	NONE
method	short for *method acting*	no examples found; however, examples of *method acting* were found	added *method acting* to entry candidates
	there's method in/to one's madness	no examples found	NONE
obtain	*formal* to exist	1 example in CAJA	recorded the meaning, made a note, but no sense was created in the entry
	to obtain something through something	331 examples in CAJA (nearly all in passive)	added *something is obtained through something* as construction under sense 1
***potential* (adjective)**	*Grammar* (of a verb or form of a verb) expressing possibility, as English *may* and *might*	no examples found in Linguistics and Education subcorpora	NONE
***potential* (noun)**	*Grammar* a potential verb or verb form	no examples found in Linguistics and Education subcorpora	NONE
significant	having special meaning *significant look/glance/smile/wink*	no examples of frequent collocate patterns found (-10+10 span used)	NONE
various	very different from each other often in pattern *be various* OR BrE *many and various* AmE *various and sundry*	*be various:* 10 examples in CAJA 6 examples in BAWE *many and various*: 4 examples in CAJA *various and sundry*: 5 examples in CAJA	added a sense (sense 2) and a note about the sense potentially requiring a label *written* or *mainly in writing*

20. APPENDIX 10: DOAE STYLE SETS – TABLES AND FIGURES

Figure 102. DOAE style sets: Default style and formatting settings.

justify /ˈʤʌstɪˌfaɪ/ *verb* [transitive]

WORD FORMS:
justify, justifies, justifying, justified

① If a thing or a person **justifies** something such a decision or an action, it provides evidence indicating why this decision or action is appropriate or correct.

- *Simulation examples were given to justify the theoretical conclusions.*
- *Managers may wonder what the advantages of investing in relationships with customers are and how they can justify this investment.*
- *The Security Council Vietnam justified its intervention in Kampuchea principally on grounds of self-defence but also on humanitarian grounds.*

something is justified by something

- *My decision to buy cigarettes is not justified by the fact that I cannot get rid of my intention to smoke.*
- *Retaining an HEI-based component is justified by a belief that this impacts in some beneficial way on workplace practice.*

② If someone **justifies** himself or herself, they prove their worthiness. If something **justifies** itself, it proves it exists for a good reason.

- *Over the course of the play, Hamlet will justify himself by straddling thought and action, interiority and exteriority, erecting a field of 'outness' around him that has stabilized his challenged masculinity for more than four centuries.*

③ **the end justifies the means** Used to say that the use of any methods, even bad ones, is acceptable if they lead to an important result.

④ **Printing** to adjust the text so that the lines begin or end, or both, at the same distance from the margin

⑤ **Religion** to declare or make righteous in the sight of God

WORD ORIGIN
Date: 1300-1400
Language: Old French
Origin: *justifier*, from Latin *justificare* (to make just) = *justus* (just) + *facere* (to make)

Table 137. DOAE style sets: Default setting: Fonts, font sizes, font styles, and colours.

Entry feature	Font	Font size	Font style	Colour
Headword sign	Verdana	13	bold	dark blue
Headword variant	Verdana	12	bold	dark blue
Pronunciation	Trebuchet MS	12		black
Word class	Arial	11	bold italics	light blue
Inflected forms	Verdana	10	italics	dark blue
Menu - text Domain label Subdomain label	Georgia Georgia Georgia	11 10 10	 bold bold	black red red
Subentry - lemma sign	Verdana	12	bold	dark blue
Sense				light blue background
Sense variant	Verdana	10	bold	dark blue
Patterns with sense status	Arial	11	bold	dark blue
Definition	Arial	11		black
Domain label, subdomain label	Arial	11	bold	red
Grammar label	Georgia	10	bold	light green
Regional label	Verdana	10	bold italics	purple
Example	Verdana	10	italics	black
Constructions	Arial	11	bold	light blue
Collocate box	Verdana	10		dark blue for title, black for text, light blue background
Frequency graph - text	Arial	11		
Etymology Date Language Origin	Arial Comic Sans MS Comic Sans MS Comic Sans MS	11 10 10 10	bold bold bold bold	black dark red green black
References	Georgia	11	bold small caps	dark blue
Usage note	Trebuchet MS	10		Black for text, pale yellow background

Figure 103. DOAE style sets: Black & white style and formatting settings.

justify /ˈdʒʌstɪˌfaɪ/ *verb* [transitive]

WORD FORMS:
justify, justifies, justifying, justified

❶ If a thing or a person **justifies** something such a decision or an action, it provides evidence indicating why this decision or action is appropriate or correct.

- *Simulation examples were given to justify the theoretical conclusions.*
- *Managers may wonder what the advantages of investing in relationships with customers are and how they can justify this investment.*
- *The Security Council Vietnam justified its intervention in Kampuchea principally on grounds of self-defence but also on humanitarian grounds.*

something is justified by something

- *My decision to buy cigarettes is not justified by the fact that I cannot get rid of my intention to smoke.*
- *Retaining an HEI-based component is justified by a belief that this impacts in some beneficial way on workplace practice.*

❷ If someone **justifies** himself or herself, they prove their worthiness. If something **justifies** itself, it proves it exists for a good reason.

- *Over the course of the play, Hamlet will justify himself by straddling thought and action, interiority and exteriority, erecting a field of 'outness' around him that has stabilized his challenged masculinity for more than four centuries.*

❸ **the end justifies the means** Used to say that the use of any methods, even bad ones, is acceptable if they lead to an important result.

❹ *Printing* to adjust the text so that the lines begin or end, or both, at the same distance from the margin

❺ *Religion* to declare or make righteous in the sight of God

WORD ORIGIN
Date: 1300-1400
Language: Old French
Origin: *justifier*, from Latin *justificare* (to make just) = *justus* (just) + *facere* (to make)

Table 138. DOAE style sets: Black & white setting: Fonts, font sizes, font styles, and colours.*

Entry feature	Font	Font size	Font style	Colour
Headword sign	Verdana	13	bold	**black**
Headword variant	Verdana	12	bold	**black**
Pronunciation	Trebuchet MS	12		black
Word class	Arial	**12**	bold italics	**dark grey**
Inflected forms	Verdana	10	italics	**black**
Menu - text Domain label Subdomain label	Georgia Georgia Georgia	11 10 10	 **italics** **italics**	black **black** **black**
Subentry - lemma sign	Verdana	12	bold	**black**
Sense				**light grey background**
Sense variant	Verdana	10	bold	**black**
Patterns with sense status	**Arial Unicode MS**	**12**	bold	**black**
Definition	Arial	11		black
Domain label, subdomain label	**Georgia**	**11**	**italics**	**black**
Grammar label	Georgia	10	bold	**light grey**
Regional label	**Arial**	**11**	**italics**	**black**
Example	Verdana	10	italics	black
Constructions	Arial	11	bold	**black**
Collocate box	Verdana	10		**black for title/text, light grey for background**
Frequency graph - text	Arial	11		
Etymology Date Language Origin	Arial Comic Sans MS Comic Sans MS Comic Sans MS	11 10 10 10	bold bold bold bold	**dark grey** **black** **black** black
References	Georgia	11	bold small caps	**black**
Usage note	Trebuchet MS	10		Black for text, **white background**

* Items in bold are the ones that have been modified from the default setting.

Figure 104. DOAE style sets: Medium-size style and formatting settings.

justify /ˈdʒʌstɪˌfaɪ/ *verb* [transitive]

WORD FORMS:
justify, justifies, justifying, justified

① If a thing or a person **justifies** something such a decision or an action, it provides evidence indicating why this decision or action is appropriate or correct.

- *Simulation examples were given to justify the theoretical conclusions.*
- *Managers may wonder what the advantages of investing in relationships with customers are and how they can justify this investment.*
- *The Security Council Vietnam justified its intervention in Kampuchea principally on grounds of self-defence but also on humanitarian grounds.*

something is justified by something

- *My decision to buy cigarettes is not justified by the fact that I cannot get rid of my intention to smoke.*
- *Retaining an HEI-based component is justified by a belief that this impacts in some beneficial way on workplace practice.*

② If someone **justifies** himself or herself, they prove their worthiness. If something **justifies** itself, it proves it exists for a good reason.

- *Over the course of the play, Hamlet will justify himself by straddling thought and action, interiority and exteriority, erecting a field of 'outness' around him that has stabilized his challenged masculinity for more than four centuries.*

③ **the end justifies the means** Used to say that the use of any methods, even bad ones, is acceptable if they lead to an important result.

④ **Printing** to adjust the text so that the lines begin or end, or both, at the same distance from the margin

⑤ **Religion** to declare or make righteous in the sight of God

WORD ORIGIN
Date: 1300-1400
Language: Old French
Origin: *justifier*, from Latin *justificare* (to make just) = *justus* (just) + *facere* (to make)

Figure 105. DOAE style sets: Large-size style and formatting settings.

justify /ˈdʒʌstɪˌfaɪ/ *verb* [transitive]

WORD FORMS:
justify, justifies, justifying, justified

① If a thing or a person **justifies** something such a decision or an action, it provides evidence indicating why this decision or action is appropriate or correct.

- *Simulation examples were given to justify the theoretical conclusions.*
- *Managers may wonder what the advantages of investing in relationships with customers are and how they can justify this investment.*
- *The Security Council Vietnam justified its intervention in Kampuchea principally on grounds of self-defence but also on humanitarian grounds.*

something is justified by something

- *My decision to buy cigarettes is not justified by the fact that I cannot get rid of my intention to smoke.*
- *Retaining an HEI-based component is justified by a belief that this impacts in some beneficial way on workplace practice.*

② If someone **justifies** himself or herself, they prove their worthiness. If something **justifies** itself, it proves it exists for a good reason.

- *Over the course of the play, Hamlet will justify himself by straddling thought and action, interiority and exteriority, erecting a field of 'outness' around him that has stabilized his challenged masculinity for more than four centuries.*

③ **the end justifies the means** Used to say that the use of any methods, even bad ones, is acceptable if they lead to an important result.

④ **Printing** to adjust the text so that the lines begin or end, or both, at the same distance from the margin

⑤ **Religion** to declare or make righteous in the sight of God

WORD ORIGIN
Date: 1300-1400
Language: Old French
Origin: *justifier*, from Latin *justificare* (to make just) = *justus* (just) + *facere* (to make)

Figure 106. DOAE style sets: Senses and frequent patterns in the entry *fact* (Chemistry settings).

fact /fækt/ *noun* [countable]

① **in fact; *also* as a matter of fact, in actual fact** Used to indicate that you are about to provide more detailed information about what has just been said, or that you will provide the information that is in contrast with what has just been said.

② **the fact that** Used when you want to emphasize that some situation or information is true.

FREQUENT PATTERNS
to **ignore/consider** the fact (that)
to **contradict/explore** the fact (that)

③ a piece of information that is known to be true

④ an actual occurrence; an actual event

after the fact after the event that has just been mentioned

⑤ **the fact is / the fact of the matter is** Used to introduce a statement in which you make an important point about what has just been said.

⑥ **FACT** Biochemistry abbreviation for **facilitates chromatin transcription** *SEE* **FACT**

⑦ Computing The kind of clause used in logic programming which has no subgoals and so is always true (always succeeds).

⑧ **a fact of life** something, often unpleasant, that is true and cannot be avoided

⑨ **a stylized fact** a simplified presentation of an empirical finding

⑩ Philosophy a situation, a proposition

⑪ **after/before the fact** Law after/before a criminal act

Figure 107. DOAE style sets: Senses and frequent patterns in the entry *fact* (Combined Honours: Linguistics and Psychology settings).

fact /fækt/ *noun* [countable]

① **in fact;** *also* **as a matter of fact, in actual fact** Used to indicate that you are about to provide more detailed information about what has just been said, or that you will provide the information that is in contrast with what has just been said.

② **the fact that** Used when you want to emphasize that some situation or information is true.

FREQUENT PATTERNS
to **ignore/consider** the fact (that)
to **highlight/overlook/obscure** the fact (that)

③ a piece of information that is known to be true

④ an actual occurrence; an actual event

after the fact after the event that has just been mentioned

⑤ **the fact is / the fact of the matter is** Used to introduce a statement in which you make an important point about what has just been said.

⑥ **a fact of life** something, often unpleasant, that is true and cannot be avoided

⑦ **a stylized fact** a simplified presentation of an empirical finding

⑧ **Philosophy** a situation, a proposition

⑨ **after/before the fact** **Law** after/before a criminal act

⑩ **FACT** **Biochemistry** abbreviation for **facilitates chromatin transcription** *SEE* **FACT**

⑪ **Computing** The kind of clause used in logic programming which has no subgoals and so is always true (always succeeds).

Figure 108. DOAE: The entry *attribute* (style set: non-native speaker student of Business and Management).

attribute

MENU

noun

1. a characteristic or feature
2. a positive characteristic
3. **Logic** property of a subject in proposition
4. **Grammar** attributive adjective or noun
5. **Computing** a data item with a certain value

verb

1. assign a finding to something
2. assign a characteristic to something
3. **attribute blame/responsibility**
4. assign authorship to someone
5. assign object to a particular period

attribute /ˈætrəˌbjuːt/ *noun* **[countable]**

① a characteristic or feature of a thing or person

- *Note that of the 48 available attributes, quality is the most important attribute selected by consumers.*
- *An example of **a positive attribute** is "The candidate can concentrate very well over long periods."*
- *Fader and Hardie exploit the notion that categories with a large number of alternatives can usually be described according to a substantially smaller, stable set of attributes (e.g., brand, size, flavor, form).*
- *What is evident here is a strong focus on foreign languages, humanities and the intellectual and cultural attributes of the Italian people.*
- *Subjects were verbally instructed to pay particular attention to the attributes of various products whose advertisements were included in the booklet.*
- *Silica materials have poor mechanical attributes, which limit their applications.*

FREQUENT PATTERNS
key/essential attribute(s)
job/earnings/search attribute(s)

② a quality or positive characteristic of a person or institution

- *Persistence was recognized by seventeen of the scientists as the single most important **personal attribute** necessary for success.*
- *The quality of local schools is another attribute that may be critical to a strategy to attract creative workers.*
- *In Mühlenberg's view no Greek philosopher or theologian before Gregory mentions infinity as an attribute of God, because infinity was connected with imperfection and the material world.*

③ **Logic** the property, quality, or feature that is affirmed or denied concerning the subject of a proposition

④ **Grammar** an adjective or noun modifying a noun; an attributive adjective or noun

⑤ **Computing** a data item with a certain value that describes a property of an object, entity, or file

- *The network file contains a set of attributes such as street length, the name of the street, and alternative name of the street.*
- *Each update has unit cost. The cost of each query is uniformly distributed at random between 1 to 20. In this set of experiments, the range **attribute values** were between 1,000 and 30,000.*

attribute /əˈtrɪbjuːt/ *verb* **[transitive]**

① If a finding or result **is attributed to** something, it is believed that the finding or result was caused by that thing.

- *There is a discrepancy between the actual experimental values and the theoretical predictions, which has been attributed to the presence of absorption.*
- *However, in the Prussian experience the initiative to participate in overseas commerce must be largely attributed to the government bureaucracy in Berlin.*
- *The high cost of housing helps developers meet their requirements and continue to profit, but most importantly the program's success should be attributed to strong support from the City's leadership.*

attribute something to something

- *Microsoft attributed this growth in Internet Explorer's usage share to the replacement of AOL's Booklink browser with Internet Explorer rather than coming at the expense of Netscape Navigator.*

• *The authors attribute the differences between Vietnam regions and provinces to differences in climatic and environmental factors.*

• *On the contrary, performance-oriented students tend to attribute failure to a lack of abilities.*

be attributed to the fact that

• *The criminality of undocumented immigrants was generally attributed to the simple fact that they had no legal right to be in the United States.*

② If you **attribute** a characteristic or property **to** a thing or person, you say that it has that characteristic or property.

• *There was, however, an important exception: proficiency in mathematics was attributed more to boys than to girls (Räty, 2003).*

• *Khan et al. (2000) found that those in this sector tend to attribute much greater importance to costing techniques and inventory control and less to implementing strategy.*

• *Unlike horror movies and other situations in which we seem to enjoy negative emotions, documentaries offer the potential for the same kind of knowledge attributed to tragedy.*

③ If you **attribute blame** or **responsibility** for something bad **to** a person or organization, you say or think they caused that thing to happen.

• *Chavez's supporters and opponents have both attributed to him considerable responsibility for the resurgence of Latin America's left - most recently with the election of Evo Morales in Bolivia.*

• *Customers reporting two failures will attribute blame for the failures to the firm more strongly after the second failure than after the first failure.*

④ If you **attribute** something, such as a statement or a work of art, **to** a person, you believe that person is its author.

• *This painting is very likely the large painting of St Francis listed in the 1613 inventory, attributed to Michelangelo.*

• *Emilio Bigi, Giuseppe Corsi and most modern literary scholars now attribute the text to Castellani.*

• *The paper distinguishes between two different senses of 'genius' found in Kant's Critique of Judgement, and criticizes an argument commonly attributed to Kant.*

⑤ If you **attribute** an object **to** a particular period, you say it comes from that period.

• *The oldest, the Lower Fortress was attributed to the period of King Solomon.*

Figure 109. DOAE: The entry *attribute* (style set: native-speaker student of Engineering).

attribute

MENU

verb

1. assign a finding to something
2. assign object to a particular period
3. assign a characteristic to something
4. **attribute blame/responsibility**
5. assign authorship to someone

noun

1. a characteristic or feature
2. a positive characteristic
3. **Computing** a data item with a certain value
4. **Logic** property of a subject in proposition
5. **Grammar** attributive adjective or noun

attribute *verb*

① If a finding or result **is attributed to** something, it is believed that the finding or result was caused by that thing.

- *There is a discrepancy between the actual experimental values and the theoretical predictions, which has been attributed to the presence of absorption.*
- *However, in the Prussian experience the initiative to participate in overseas commerce must be largely attributed to the government bureaucracy in Berlin.*
- *The high cost of housing helps developers meet their requirements and continue to profit, but most importantly the program's success should be attributed to strong support from the City's leadership.*

attribute something to something

- *Microsoft attributed this growth in Internet Explorer's usage share to the replacement of AOL's Booklink browser with Internet Explorer rather than coming at the expense of Netscape Navigator.*
- *The authors attribute the differences between Vietnam regions and provinces to differences in climatic and environmental factors.*
- *On the contrary, performance-oriented students tend to attribute failure to a lack of abilities.*

be attributed to the fact that

• *The criminality of undocumented immigrants was generally attributed to the simple fact that they had no legal right to be in the United States.*

② If you **attribute** an object **to** a particular period, you say it comes from that period.

• *The oldest, the Lower Fortress was attributed to the period of King Solomon.*

③ If you **attribute** a characteristic or property **to** a thing or person, you say that it has that characteristic or property.

• *There was, however, an important exception: proficiency in mathematics was attributed more to boys than to girls (Räty, 2003).*
• *Khan et al. (2000) found that those in this sector tend to attribute much greater importance to costing techniques and inventory control and less to implementing strategy.*
• *Unlike horror movies and other situations in which we seem to enjoy negative emotions, documentaries offer the potential for the same kind of knowledge attributed to tragedy.*

④ If you **attribute blame** or **responsibility** for something bad **to** a person or organization, you say or think they caused that thing to happen.

• *Chavez's supporters and opponents have both attributed to him considerable responsibility for the resurgence of Latin America's left - most recently with the election of Evo Morales in Bolivia.*
• *Customers reporting two failures will attribute blame for the failures to the firm more strongly after the second failure than after the first failure.*

⑤ If you **attribute** something, such as a statement or a work of art, **to** a person, you believe that person is its author.

• *This painting is very likely the large painting of St Francis listed in the 1613 inventory, attributed to Michelangelo.*
• *Emilio Bigi, Giuseppe Corsi and most modern literary scholars now attribute the text to Castellani.*
• *The paper distinguishes between two different senses of 'genius' found in Kant's Critique of Judgement, and criticizes an argument commonly attributed to Kant.*

attribute *noun*

① a characteristic or feature of a thing or person

• *Note that of the 48 available attributes, quality is the most important attribute selected by*

consumers.

- *An example of **a positive attribute** is "The candidate can concentrate very well over long periods."*
- *Fader and Hardie exploit the notion that categories with a large number of alternatives can usually be described according to a substantially smaller, stable set of attributes (e.g., brand, size, flavor, form).*
- *What is evident here is a strong focus on foreign languages, humanities and the intellectual and cultural attributes of the Italian people.*
- *Subjects were verbally instructed to pay particular attention to the attributes of various products whose advertisements were included in the booklet.*
- *Silica materials have poor mechanical attributes, which limit their applications.*

FREQUENT PATTERNS
key/essential attribute(s)

② a quality or positive characteristic of a person or institution

- *Persistence was recognized by seventeen of the scientists as the single most important **personal attribute** necessary for success.*
- *The quality of local schools is another attribute that may be critical to a strategy to attract creative workers.*
- *In Mühlenberg's view no Greek philosopher or theologian before Gregory mentions infinity as an attribute of God, because infinity was connected with imperfection and the material world.*

③ **Computing** a data item with a certain value that describes a property of an object, entity, or file

- *The network file contains a set of attributes such as street length, the name of the street, and alternative name of the street.*
- *Each update has unit cost. The cost of each query is uniformly distributed at random between 1 to 20. In this set of experiments, the range **attribute values** were between 1,000 and 30,000.*

④ **Logic** the property, quality, or feature that is affirmed or denied concerning the subject of a proposition

⑤ **Grammar** an adjective or noun modifying a noun; an attributive adjective or noun

21. APPENDIX 11: DISCUSSIONS: ENTRY COMPARISONS – TABLES

Table 139. Entry for *significant* in DOAE and 10 existing dictionaries.

DOAE	1. A **significant** difference, correlation, etc. between the observed value and the hypothesis is too big to be attributed to chance. • *Wright and Cropanzano (2000, p. 92) identify a significant relationship between staff well-being and their performance in the workplace.* • *Evenness values, although larger than those reported in, still display a significant increase through time (rs=+0.893, p=0.007).* **statistically significant** • *The results of this final search corroborated the previous findings: men cite themselves slightly more than women (12.1 % as opposed to 11.1 % in articles; 14.4 % as opposed to 13.1 % in reports), but the differences are not statistically significant.* • *Differences were analyzed by one-way ANOVA test, by using SPSS software and considered statistically significant at P < 0.05 and P < 0.01.* 2. A **significant** amount or change is very large or considerable. • *First, the developer will pay $200,000 in fees, a significant increase in past fee levels.* • *A significant amount of research has been conducted on this barrier island system as well as the other barriers in Louisiana.* • *But hostilities were brutal in Croatia, especially in its eastern border region where Serbs comprised a significant portion of the population (see Grandits and Promitzer 2000).* • *Hookworm infection causes significant loss of blood, resulting in severe anemia.* • *Price levels and e-mail coupons do not have a significant effect on order size.* 3. A fact or an event that is **significant** is considered important or noticeable. • *The fact that the British Government was prepared to pay £150,000 for an expedition that identified anthropological activity as a primary objective is significant.* • *In the 1997 Eurobarometer survey, immigration turns out to be one of the three most significant political or social issues.* • *Examining the relation between official and individual narratives of the past should be highly significant for the understanding of learning.* • *The Instituto Cervantes/Spanish Cultural Center in Damascus has played a significant role in promoting the Mediterranean dimensions of Syrian culture through its programming.*
CED CD-ROM	1. having or expressing a meaning; indicative 2. having a covert or implied meaning; suggestive 3. important, notable, or momentous 4. (Statistics) of or relating to a difference between a result derived from a hypothesis and its observed value that is too large to be attributed to chance and that therefore tends to refute the hypothesis

NODE CD-ROM	1. sufficiently great or important to be worthy of attention; noteworthy: a significant increase in sales. 2. having a particular meaning; indicative of something: in times of stress her dreams seemed to her especially significant. • suggesting a meaning or message that is not explicitly stated: *she gave him a significant look.* 3 ***Statistics*** of, relating to, or having significance.
MWCD CD-ROM	1. having meaning; especially: SUGGESTIVE *a significant glance* 2. a: having or likely to have influence or effect :IMPORTANT *a significant piece of legislation*; also: of a noticeably or measurably large amount *a significant number of layoffs , producing significant profits* b: probably caused by something other than mere chance *statistically significant correlation between vitamin deficiency and disease*
LED e-LDOCE	1. having an important effect or influence, especially on what will happen in the future ▪ *His most significant political achievement was the abolition of the death penalty.* ▪ *Please inform us if there are any significant changes in your plans.* ▪ *In the eighteenth century, the written word was rarely a significant factor in the life of the general public.* ▪ *The slight difference in the way men and women are affected by the drug is not really significant* ***significant for*** ▪ *The result is highly significant for the future of the province.* ▪ *The problems on the U.S. stock markets were not significant for Europe.* ***it is significant that*** ▪ *It is significant that the writers of the report were all men.* 2. large enough to be noticeable or have noticeable effects ▪ *A significant number of drivers fail to keep to speed limits.* ▪ *A significant part of Japan's wealth is invested in the West.* ▪ *There is a significant difference between the number of home births now and ten years ago.* ▪ *The rise in temperature is not statistically significant.* ▪ *The proportion of the population that is overweight is now significant.* 3. a significant look, smile etc has a special meaning that is not known to everyone: ▪ *He gave me a significant look.*
COBUILD CD-ROM	1. A **significant** amount or effect is large enough to be important or affect a situation to a noticeable degree. *Most 11-year-olds are not encouraged to develop reading skills; a small but significant number are illiterate.* *...foods that offer a significant amount of protein...* *It is the first drug that seems to have a very significant effect on this disease.* 2. A **significant** fact, event, or thing is one that is important or shows something. *Time would appear to be the significant factor in this whole drama.* *...a very significant piece of legislation...*

	I think it was significant that he never knew his own father. 3. A **significant** action or gesture is intended to have a special meaning. *Mrs Bycraft gave Rose a significant glance.*
e-MED	1. very large or noticeable *I think we can save a significant amount of time.* *The increase in enrolment this year is significant.* *a significant portion of the population* 2. very important *Davis was one of the most significant musicians of the last century.* *There's been some significant progress.* 3. having a special meaning that only some people understand *The look he gave her seemed to be significant.*
e-CALD	**significant** (important) important or noticeable *There has been a significant increase in the number of women students in recent years.* *The talks between the USA and the USSR were very significant for the relationship between the two countries.* **significant** (special meaning) having a special meaning *She looked at him across the table and gave him a significant smile.* *Do you think it's significant that he hasn't replied to my letter yet?*
e-OALD	1. large or important enough to have an effect or to be noticed: a highly significant discovery ** The results of the experiment are not statistically significant. * There are no significant differences between the two groups of students. * Your work has shown a significant improvement. * These views are held by a significant proportion of the population. * It is significant that girls generally do better in examinations than boys. * The drug has had no significant effect on stopping the spread of the disease.* 2. having a particular meaning *It is significant that he changed his will only days before his death.* 3. [usually before noun] having a special or secret meaning that is not understood by everyone *a significant look / smile*
Dictionary.com	1. important; of consequence. 2. having or expressing a meaning; indicative; suggestive: a significant wink. 3. *Statistics* of or pertaining to observations that are unlikely to occur by chance and that therefore indicate a systematic cause.

Table 140. Entry for *argue* in DOAE and 10 existing dictionaries.

DOAE	1. [transitive] If someone **argues** a view or an idea in an article or book, they present the idea and support it with evidence. Note: **argue** is very often followed by a **that**-clause. • *Mazrui (1999: 1) also argued that Africa developed the West.* • *The question, I will argue in this essay, cannot be answered without a clearer sense of how Greece relates to Rome in Shelley's work.* • *Some researchers have argued that a decentralized economy will have difficulty in fully exploiting the growth returns of general purpose technologies.* • *"Some frames," argues Gamson (1992: 135), "have a natural advantage because their ideas and language resonate with a broader political culture."* **article/paper/essay argues that...** • *This article argues that literacy is an important sociological phenomenon, but one largely under-researched in British sociology.* **it could/can/might/may be argued that...** • *It could be argued that, by suggesting complete openness the liberals initially tried to handle the issue differently.* **one could/can/might/may argue that...** • *In almost simplistic terms, one could argue that, without sound, there would be no music.* **as someone argues somewhere / as argued somewhere** • *As argued elsewhere (Steadman and Palmer 1995; Palmer and Steadman 2004), the key to this distinction appears to be the behaviour of the listeners.* • *No doubt state leaders in new states often follow the ethnic model of nation-building, but, as I argue below (=later in the article), this is not the only possible solution, and normally not the best one in order to survive and flourish as nations.* **as argued by someone / as someone argues** • *Stravinsky's music was never intended to be complete in itself, but to be made complete by the choreography, as Irene Alm has persuasively argued.* • *The results are consistent with the hypothesis that capital markets are integrated, as argued by Campbell and Hamao (1992) and Harvey (1991), among others.* 2. [intransitive] If you **argue for** or you **argue in favour** of an idea or theory, you agree with it and provide evidence that supports it. • *Vygotsky (1978) argued for the importance of language as both a psychological and cultural tool.* • *This article argues in favour of putting into place a legal framework for feedback intermediaries.* 3. [intransitive] If you **argue against** an idea or theory, you provide evidence that opposes it. • *Ian Sefton is physics educator who has strongly argued against the mistaken view that electrons possess potential energy.* • *Nevertheless, four pieces of circumstantial evidence argue against this narrower view.* 4. [intransitive] If you **argue** with someone about/over something, you discuss it because you have different opinions. • *Living cells had of course been seen through the microscope before, and a range of their activities had been described and*

	argued over. • *The committee also argued about and negotiated the unorthodox multimedia exhibition style.* 5. [intransitive] If you **argue with** someone or someone's view, you disagree with it. • *Few would argue with the idea that we should maximise positive value and minimise evil.* • *According to Derrida, the machine is dangerous because it is the opposite of life; it is 'pure representation' and 'never runs by itself'. So far it is difficult to argue with Derrida.* 6. [intransitive] If people **argue**, they talk angrily to each other because they disagree. • *Again, the father went to argue with the doctors. The chief physician went almost wild, according to Hallvard.* **argue about/over something** • *Bernardino Antonio, a resident of Actopan, beat his wife fatally in 1808 after they argued about the punishment of their child.*
CED CD-ROM	1. ***intr*** to quarrel; wrangle *they were always arguing until I arrived* 2. ***intr; often foll by: for*** *or* ***against*** to present supporting or opposing reasons or cases in a dispute; reason 3. ***tr; may take a clause as object*** to try to prove by presenting reasons; maintain 4. ***tr; often passive*** to debate or discuss *the case was fully argued before agreement was reached* 5. ***tr*** to persuade *he argued me into going* 6. ***tr*** to give evidence of; suggest *her looks argue despair*
NODE CD-ROM	1. [REPORTING VERB] give reasons or cite evidence in support of an idea, action, or theory, typically with the aim of persuading others to share one's view: [with CLAUSE] *sociologists argue that inequalities in industrial societies are being reduced* | [with DIRECT SPEECH] *'It stands to reason,' she argued.* [with OBJ.] (argue someone into/out of) persuade someone to do or not to do (something) by giving reasons: *I tried to argue him out of it.* 2. [no OBJ.] exchange or express diverging or opposite views, typically in a heated or angry way: *don't argue with me* | figurative *I wasn't going to argue with a gun* | [with OBJ.] *she was too tired to argue the point.* PHRASES **argue the toss** *informal, chiefly Brit.* dispute a decision or choice already made.
MWCD CD-ROM	intransitive verb 1. to give reasons for or against something: REASON *argue for a new policy*

	2 : to contend or disagree in words: DISPUTE *argue about money* transitive verb 1 : to give evidence of: INDICATE *the facts argue his innocence* 2. to consider the pros and cons of: DISCUSS *argue an issue* 3. to prove or try to prove by giving reasons: MAINTAIN *asking for a chance to argue his case* 4. to persuade by giving reasons: INDUCE *couldn't argue her out of going*
LED, e-LDOCE	1. [intransitive] to disagree with someone in words, often in an angry way: ▪ *We could hear the neighbours arguing.* **argue with** ▪ *Gallacher continued to argue with the referee throughout the game.* **argue about** ▪ *They were arguing about how to spend the money.* **argue over** ▪ *The children were arguing over which TV programme to watch.* 2. [intransitive and transitive] to state, giving clear reasons, that something is true, should be done etc **argue that** ▪ *Croft argued that a date should be set for the withdrawal of troops.* ▪ *It could be argued that a dam might actually increase the risk of flooding.* **argue for/against (doing) something** ▪ *Baker argued against cutting the military budget.* ▪ *All the available evidence argues against this theory.* ▪ *She argued the case for changing the law.* ▪ *The researchers put forward a well-argued case for banning the drug.* ▪ *They argued the point (=discussed it) for hours without reaching a conclusion.* 3. argue somebody into/out of doing something British English to persuade someone to do or not do something: ▪ *Joyce argued me into buying a new jacket.* 4. [transitive] formal to show that something clearly exists or is true synonym demonstrate: ▪ *The statement argues a change of attitude by the management.* 5. argue the toss British English informal to continue to argue about a decision that has been made and cannot be changed: ▪ *There was no point arguing the toss after the goal had been disallowed.*
COBUILD CD-ROM	1. If one person **argues** with another, they speak angrily to each other about something that they disagree about. You can also say that two people **argue**. *The committee is concerned about players' behaviour, especially arguing with referees.*

	They were still arguing; I could hear them down the road. 2. If you tell someone not to **argue** with you, you want them to do or believe what you say without protest or disagreement. *Don't argue with me.* *The children go to bed at 10.30. No one dares argue.* 3. If you **argue** with someone about something, you discuss it with them, with each of you giving your different opinions. *He was arguing with the King about the need to maintain the cavalry at full strength.* *They are arguing over foreign policy.* *The two of them sitting in their office were arguing this point.* 4. If you **argue** that something is true, you state it and give the reasons why you think it is true. *His lawyers are arguing that he is unfit to stand trial.* *It could be argued that the British are not aggressive enough.* 5. If you **argue for** something, you say why you agree with it, in order to persuade people that it is right. If you **argue against** something, you say why you disagree with it, in order to persuade people that it is wrong. *The report argues against tax increases.* *I argued the case for an independent central bank.* 6. If you **argue**, you support your opinions with evidence in an ordered or logical way. *I've argued deductively from the text.* *I'd like to argue in a framework that is less exaggerated.* 7. If you say that no-one can **argue** with a particular fact or opinion, you are emphasizing that it is obviously true and so everyone must accept it. (SPOKEN) *We produced the best soccer of the tournament. Nobody would argue with that.*
e-MED	1. [intransitive] if people argue, they speak to each other in an angry way because they disagree *Those girls are always arguing!* **argue with:** *Don't argue with me – you know I'm right.* **argue about/over:** *We used to argue over who should drive.* 1.a. [intransitive/transitive] to discuss something with someone who has a different opinion from you The programme gives people a chance to argue their ideas. **argue about/over:** *They are still arguing over the details of the contract.* 2. [intransitive/transitive] to give reasons why you believe that something is right or true *Successful economies, she argues, are those with the lowest taxes.* **argue for/against:**

	Woolf's report argued for (=supported) an improvement in prison conditions. **argue that:** *Reuben opposed the new road, arguing that it wasn't worth spending $25 million to cut seven minutes off drivers' journey times.* *Several people stood up to argue against (=say they do not support) moving the students to the new school.* **phrase** **argue someone into/out of (doing) something** *British* to persuade someone to do/not to do something *I've managed to argue him out of going to the match.*
e-CALD	**argue** (disagree) [I] to speak angrily to someone, telling them that you disagree with them *The children are always arguing.* *Kids, will you stop arguing with each other?* *They were arguing over/about which film to go and see.* **argue** (give reasons) [I or T] to give the reasons for your opinion, idea, belief, etc. *The minister argued for/in favour of/against making cuts in military spending.* [+ that] *The minister argued that cuts in military spending were needed.* *You can argue the case either way.* **argue** (show) [T] to show that something is true or exists *The evidence argues a change in policy.* **argue the toss** *UK informal disapproving* to disagree with a decision or statement *It doesn't matter what you say, he'll always argue the toss!*
e-OALD	1. [v] ~ (with sb) (about / over sth) to speak angrily to sb because you disagree with them: *My brothers are always arguing. *We're always arguing with each other about money. * I don't want to argue with you—just do it! * He's offering to pay so who am I to argue?* 2. ~ (for / against sth)| ~ (for / against doing sth) to give reasons why you think that sth is right/wrong, true/not true, etc., especially to persuade people that you are right: *[v] They argued for the right to strike. * [vn] She argued the case for bringing back the death penalty. * He was too tired to argue the point (= discuss the matter). * a well-argued article * [v that] He argued that they needed more time to finish the project. * [vn that] It could be argued that laws are made by and for men. [also v wh-]* 3. [vn] (formal) to show clearly that sth exists or is true: *These latest developments argue a change in government policy.* **idiom *argue the toss*** (*BrE, informal*) to continue to disagree about a decision, especially when it is too late to change it or it is not

	very important **phrasal verb *argue sb into/out of doing sth*** to persuade sb to do/not do sth by giving them reasons: *They argued him into withdrawing his complaint.* **phrasal verb *argue with sth*** (usually used in negative sentences) (*informal*) to disagree with a statement: *He's a really successful man—you can't argue with that.*
Dictionary.com	–verb (used without object) 1. to present reasons for or against a thing: *He argued in favor of capital punishment.* 2. to contend in oral disagreement; dispute: *The Senator argued with the President about the new tax bill.* –verb (used with object) 3. to state the reasons for or against: *The lawyers argued the case.* 4. to maintain in reasoning: *to argue that the news report must be wrong.* 5. to persuade, drive, etc., by reasoning: *to argue someone out of a plan.* 6. to show; prove; imply; indicate: *His clothes argue poverty.*

22. APPENDIX 12: DOAE SAMPLE ENTRIES (DEFAULT SETTINGS)

albeit /ɔːlˈbiːɪt/ *conjunction*

You use **albeit** to introduce a comment or fact which reduces the importance of the thing being discussed.

- *Spain is one European country where stories are still told, albeit less frequently now than in the past.*
- *Swiss cities provide another interesting, albeit relatively unknown, case of public debt.*
- *The laws embraced two distinct (albeit partially overlapping) models of state intervention.*

USAGE NOTE:
The comment with **albeit** is normally presented in the form of a subordinate clause, or is provided in brackets.

WORD ORIGIN
Date: 1300-1400
Origin: *al be it* (shortened form of "although it be")

analysis of variance *noun*

SEE ANOVA

ANOVA *noun*

(abbreviation for **analysis of variance**)

Statistics Statistical technique for determining the degree of difference or similarity between two or more groups of data.

- *Multiple comparisons between more than three groups were carried out using an ANOVA test.*

USAGE NOTE:
In academic writing, the abbreviation ***ANOVA*** is far more common than the full form ***analysis of variance***.

argue /ˈɑːgjuː/ *verb*

WORD FORMS:
argue, argues, arguing, argued

① [transitive] If someone **argues** a view or an idea in an article or book, they present the idea and support it with evidence. Note: **argue** is very often followed by a ***that***-clause.

- *Mazrui (1999: 1) also argued that Africa developed the West.*
- *The question, I will argue in this essay, cannot be answered without a clearer sense of how Greece relates to Rome in Shelley's work.*
- *Some researchers have argued that a decentralized economy will have difficulty in fully exploiting the growth returns of general purpose technologies.*
- *"Some frames," argues Gamson (1992: 135), "have a natural advantage because their ideas and language resonate with a broader political culture."*

article/paper/essay argues that...

- *This article argues that literacy is an important sociological phenomenon, but one largely under-researched in British sociology.*

it could/can/might/may be argued that...

- *It could be argued that, by suggesting complete openness the liberals initially tried to handle the issue differently.*

one could/can/might/may argue that...

- *In almost simplistic terms, one could argue that, without sound, there would be no music.*

as someone argues somewhere / as argued somewhere

- *As argued elsewhere (Steadman and Palmer 1995; Palmer and Steadman 2004), the key to this distinction appears to be the behaviour of the listeners.*
- *No doubt state leaders in new states often follow the ethnic model of nation-building, but, as I argue below* ***(=later in the article)****, this is not the only possible solution, and normally not the best one in order to survive and flourish as nations.*

as argued by someone / as someone argues

- *Stravinsky's music was never intended to be complete in itself, but to be made complete by the choreography, as Irene Alm has* ***persuasively argued****.*
- *The results are consistent with the hypothesis that capital markets are integrated, as argued by Campbell and Hamao (1992) and Harvey (1991), among others.*

② [intransitive] If you **argue for** or you **argue in favour of** an idea or theory, you agree with it and provide evidence that supports it.

- *Vygotsky (1978) argued for the importance of language as both a psychological and cultural tool.*
- *This article argues in favour of putting into place a legal framework for feedback intermediaries.*

③ [intransitive] If you **argue against** an idea or theory, you provide evidence that opposes it.

- *Ian Sefton is physics educator who has strongly argued against the mistaken view that electrons possess potential energy.*
- *Nevertheless, four pieces of circumstantial evidence argue against this narrower view.*

④ [intransitive] If you **argue** with someone about/over something, you discuss it because you have different opinions.

- *Living cells had of course been seen through the microscope before, and a range of their activities had been described and argued over.*
- *The committee also argued about and negotiated the unorthodox multimedia exhibition style.*

⑤ [intransitive] If you **argue with** someone or someone's view, you disagree with it.

- *Few would argue with the idea that we should maximise positive value and minimise evil.*
- *According to Derrida, the machine is dangerous because it is the opposite of life; it is 'pure representation' and 'never runs by itself'. So far it is difficult to argue with Derrida.*

⑥ [intransitive] If people **argue**, they talk angrily to each other because they disagree.

- *Again, the father went to argue with the doctors. The chief physician went almost wild, according to Hallvard.*

argue about/over something

- *Bernardino Antonio, a resident of Actopan, beat his wife fatally in 1808 after they argued about the punishment of their child.*

WORD ORIGIN

Date: 1300-1400

Language: Old French

Origin: *arguer* (to assert, charge with), from Latin *arguere* (to make clear)

assortment /əˈsɔːtmənt/ *noun* [countable]

WORD FORMS:
assortment, assortments

An **assortment** is a selection or range of similar things, such as products.

- *The main retail chains did not want to incorporate new salads into their **product assortment**.*

assortment of something

- *The items included an assortment of knives, bells, bracelets, cups, and cheap metal Hindu figurines.*
- *The United States took in the widest and most evenly balanced assortment of ethnic groups.*

attribute

MENU

verb

1. assign a finding to something
2. assign a characteristic to something
3. assign authorship to someone
4. **attribute blame/responsibility**
5. assign object to a particular period

noun

1. a characteristic or feature
2. a positive characteristic
3. **Computing** a data item with a certain value
4. **Grammar** attributive adjective or noun
5. **Logic** property of a subject in proposition

attribute /əˈtrɪbjuːt/ *verb* [transitive]

> WORD FORMS:
> ***attribute, attributes, attributing, attributed***

① If a finding or result **is attributed to** something, it is believed that the finding or result was caused by that thing.

- *There is a discrepancy between the actual experimental values and the theoretical predictions, which has been attributed to the presence of absorption.*
- *However, in the Prussian experience the initiative to participate in overseas commerce must be largely attributed to the government bureaucracy in Berlin.*
- *The high cost of housing helps developers meet their requirements and continue to profit, but most importantly the program's success should be attributed to strong support from the City's leadership.*

attribute something to something

- *Microsoft attributed this growth in Internet Explorer's usage share to the replacement of AOL's Booklink browser with Internet Explorer rather than coming at the expense of Netscape Navigator.*
- *The authors attribute the differences between Vietnam regions and provinces to differences in climatic and environmental factors.*
- *On the contrary, performance-oriented students tend to attribute failure to a lack of abilities.*

be attributed to the fact that

- *The criminality of undocumented immigrants was generally attributed to the simple fact that they had no legal right to be in the United States.*

② If you **attribute** a characteristic or property **to** a thing or person, you say that it has that characteristic or property.

- *There was, however, an important exception: proficiency in mathematics was attributed more to boys than to girls (Räty, 2003).*
- *Khan et al. (2000) found that those in this sector tend to attribute much greater importance to costing techniques and inventory control and less to implementing strategy.*
- *Unlike horror movies and other situations in which we seem to enjoy negative emotions, documentaries offer the potential for the same kind of knowledge attributed to tragedy.*

③ If you **attribute** something, such as a statement or a work of art, **to** a person, you believe that person is its author.

- *This painting is very likely the large painting of St Francis listed in the 1613 inventory, attributed to Michelangelo.*
- *Emilio Bigi, Giuseppe Corsi and most modern literary scholars now attribute the text to Castellani.*
- *The paper distinguishes between two different senses of 'genius' found in Kant's Critique of Judgement, and criticizes an argument commonly attributed to Kant.*

④ If you **attribute blame** or **responsibility** for something bad **to** a person or organization, you say or think they caused that thing to happen.

- *Chavez's supporters and opponents have both attributed to him considerable responsibility for the resurgence of Latin America's left - most recently with the election of Evo Morales in Bolivia.*
- *Customers reporting two failures will attribute blame for the failures to the firm more strongly after the second failure than after the first failure.*

⑤ If you **attribute** an object **to** a particular period, you say it comes from that period.

- *The oldest, the Lower Fortress was attributed to the period of King Solomon.*

WORD ORIGIN
Date: 1400-1500
Language: Latin
Origin: *attribuere* (assign to)

attribute /ˈætrəˌbjuːt/ *noun* [countable]

WORD FORMS:
attribute, attributes

① a characteristic or feature of a thing or person

- *Note that of the 48 available attributes, quality is the most important attribute selected by consumers.*

• *An example of **a positive attribute** is "The candidate can concentrate very well over long periods."*
• *Fader and Hardie exploit the notion that categories with a large number of alternatives can usually be described according to a substantially smaller, stable set of attributes (e.g., brand, size, flavor, form).*
• *What is evident here is a strong focus on foreign languages, humanities and the intellectual and cultural attributes of the Italian people.*
• *Subjects were verbally instructed to pay particular attention to the attributes of various products whose advertisements were included in the booklet.*
• *Silica materials have poor mechanical attributes, which limit their applications.*

FREQUENT PATTERNS
key/essential attribute(s)
job/earnings/search attribute(s)

② a quality or positive characteristic of a person or institution

• *Persistence was recognized by seventeen of the scientists as the single most important **personal attribute** necessary for success.*
• *The quality of local schools is another attribute that may be critical to a strategy to attract creative workers.*
• *In Mühlenberg's view no Greek philosopher or theologian before Gregory mentions infinity as an attribute of God, because infinity was connected with imperfection and the material world.*

③ Computing a data item with a certain value that describes a property of an object, entity, or file

• *The network file contains a set of attributes such as street length, the name of the street, and alternative name of the street.*
• *Each update has unit cost. The cost of each query is uniformly distributed at random between 1 to 20. In this set of experiments, the range **attribute values** were between 1,000 and 30,000.*

④ Grammar an adjective or noun modifying a noun; an attributive adjective or noun

⑤ Logic the property, quality, or feature that is affirmed or denied concerning the subject of a proposition

WORD ORIGIN
Date: 1400
Language: Latin
Origin: *attributum* (anything attributed)

authority /ɔːˈbrɪtɪ/ *noun*

WORD FORMS:
authority, authorities

MENU

1. power to control people or activities
2. government department
3. **(the authorities)** organizations in charge of a country
4. expert
5. important written work
6. person with power
7. official permission
8. personal quality
9. **Computing** type of internet page

① **[uncountable]** If an institution or person has **authority**, they have the right or power to control people or activities.

- *The city planners made it clear that they had no authority to alter the plan.*
- *In the household older females can have authority over younger males, especially mothers over sons.*
- *If conflicts occurred, they were about the extent to which work teams could* ***exercise authority over*** *the activities of militant students.*
- *Informational advantages brought about by specialization were perhaps the only means by which it could challenge the legislative authority of the European Council.*
- *Immigrant Asian clients may see the therapist as an expert* ***authority figure*** *who can help them solve their problems.*

FREQUENT PATTERNS
to **challenge/undermine** authority
to **delegate/exercise/grant/claim** authority
political/religious/public/legal/sovereign authority

② (*also* **Authority**) [countable, usually singular] An **authority** is a government department or official organization that is responsible for a certain area of activities, and has the power to make decisions.

- *The **planning authority** (Staffordshire County Council) received some 80 letters of objection, a few letters of support and a petition against the scheme with 3000 signatures on it.*
- *The analysed sample of firms is taken from the registers compiled by the Finnish Tax Authority.*
- *At the same time, however, it urged **health authorities** to put out to tender contracts for hospital cleaning, catering and laundry.*
- *They described the case of how the Port Authority of New York and New Jersey responded to the homelessness issue that was affecting the organization in the 1980s.*

RELATED ENTRIES: LOCAL AUTHORITY

③ [uncountable] **The authorities** are organizations or people that are in charge of a certain country or area.

- *The company itself undertakes the testing of the new products and then the new products mostly also have to be tested and approved by the authorities.*
- *Since 1982, the Chinese authorities have undertaken a nationwide campaign to check and update household registration before the census.*
- *American military authorities decided permission was needed, however, and began to warn troops in the U.K. against marrying local women.*

④ [countable] A person who is **an authority on** something is considered to be an expert on a particular subject.

- *On the basis of this relatively slim experience Simpson began to be recognized as an authority on Russia.*
- *The editors of the journal Current Anthropology solicit comments from recognized authorities.*

⑤ [countable] An **authority** is a written work that is often cited in support of a particular argument.

- *By the time Lacan gave his address to the Psychoanalytical Congress, there were more authorities that could be cited regarding the age of mirror recognition.*

⑥ [countable] An **authority** is a person in a position of power.

- *Students undergo the process of learning through traditional educational methods in which the teacher is the authority rather than a facilitator.*
- *The most basic characteristic of religion is a belief in a higher being, a supreme otherworldly authority to whom ultimate allegiance is owed.*

⑦ [uncountable] **Authority** is official permission to do something.

- *A manufacturer of VCRs was liable for authorisation of copyright infringement because of the ability of VCRs to copy without authority, television shows and movies.*
- *Section 6 of the Human Rights Act makes it 'unlawful for a court to act in a way which is incompatible with a Convention right,' without express authority from Parliament.*

⑧ [uncountable] if someone has **authority**, they are knowledgeable or behave in a way that other people listen to them.

- *He **spoke with the authority** of a wise elder.*

⑨ [countable] Computing An **authority** is an internet page that has many citations pointing to it.

- *As a consequence of Google's ranking mechanism, which prefers authorities, mainly portals of big organizations, companies, and others were retrieved.*

WORD ORIGIN
Date: 1200-1300
Language: Old French
Origin: *auctorité*, from Latin *auctoritas* (opinion, decision), from *auctor* (author)

CEO /ˌsiː iː əʊ/ *noun* [countable]

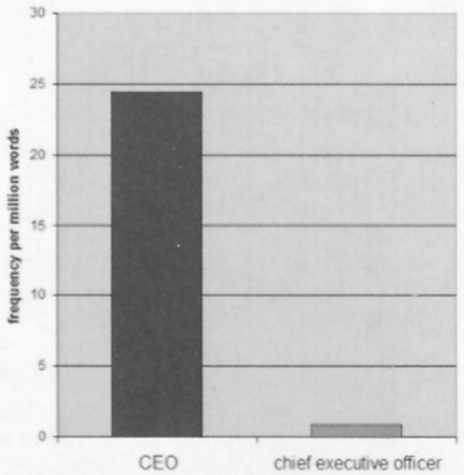
WORD FORMS:
CEO, CEOs

(abbreviation for chief executive officer)

The **CEO** of a firm or company is the person who is responsible for managing the company.

• *On July 19, 1996, Al Dunlap was hired as CEO of Sunbeam, a struggling small appliance maker.*

In academic writing, the abbreviation **CEO** is far more common than the full form **chief executive officer**.



chief executive *noun* [countable]

WORD FORMS:
chief executive, chief executives

① The **chief executive** of a firm or company is the person who is responsible for managing the company.

RELATED ENTRIES: *ALSO:* CHIEF EXECUTIVE OFFICER, CEO

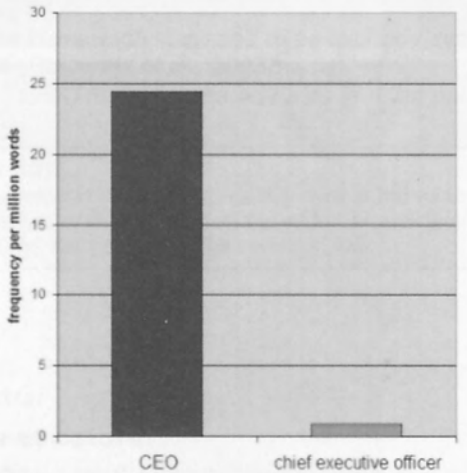
② **the Chief Executive** the president of the US

chief executive officer *noun* [countable]

WORD FORMS:
chief executive officer, chief executive officers

The **chief executive officer** of a firm or company is the person who is responsible for managing the company. *ALSO:* **CEO**

In academic writing, the abbreviation **CEO** is far more common than the full form **chief executive officer**.



electric potential *noun*

Electricity the quantity determining the energy of charge in an electric field or of mass in a gravitational field *ALSO:* POTENTIAL :②

et al. /ɛt'ɑl/ *abbreviation*

et al. means *and others* and is used to save space in academic writing when you are providing a reference for a book or article with three or more authors, but you do not want to name all the authors

- *Schools are increasingly expected to be explicitly accountable for teacher performance (Fitzgerald et al., 2003).*
- *Jain et al. (1985) looked at the attitudes of Asian patients in Birmingham to GP services.*

> USAGE NOTE:
> The abbreviation **et al.** is not usually used in the section 'References' or 'Bibliography' at the end of an article or a book as all authors need to be named in full.

etc., also **etc** /ɪt'sɛtrə/

(abbreviation for et cetera)

etc. means *'and others'* and is used at the end of a list of items, and indicates that only some items out of many have been listed

- *Each athlete indicated biographical information on the last questionnaire sheet (name, age, etc.).*
- *Public services, such as the post office, policemen, fire brigades, architects, veterinary surgeons etc. would have to co-operate with the schools in showing pupils how these services work.*

et cetera, also **etcetera** /ɪtˈsɛtrə/

et cetera is used at the end of a list of items, and indicates that only some items out of many have been listed. In writing, the abbreviation **etc.** is much more common.

- *Recently, the nonprofit Creative Commons has promoted similar licenses for other types of creative works - photos, film, music, et cetera.*
- *He has a fight with one of these guys and wins, and another guy and loses **et cetera et cetera**.*

FACT *noun* [uncountable]

(abbreviation for **facilitates chromatin transcription**)

Biochemistry a heterodimeric protein complex

- *Fewer studies have explored the role of FACT in recombination and DNA-damage response.*

fact /fækt/ *noun* [countable]

WORD FORMS:
fact, facts

MENU

1. **in fact / as a matter of fact**
2. **the fact that**
3. a piece of true information
4. an actual event
5. **the fact of the matter is / the fact is**
6. **a fact of life**
7. **stylized fact**
8. Law a criminal act
9. Philosophy a situation
10. Computing a clause in logic programming
11. Biochemistry **FACT**

① **in fact;** *also* **as a matter of fact, in actual fact** Used to indicate that you are about to provide more detailed information about what has just been said, or that you will provide the information that is in contrast with what has just been said.

- *54 out of 56 informants stated that they sang daily or often with young children. In fact, 30 % sang two or more hours per day.*
- *Construction of the scaffolding was scheduled to take three months but in fact was completed in just nine weeks.*
- *Prices of stocks in Egypt do not change much on a daily basis. As a matter of fact, prices of many stocks do not even change on a weekly basis.*

② **the fact that** Used when you want to emphasize that some situation or information is true.

- *Like most national survey samples, women are overrepresented in ours (56%),* ***reflecting the fact*** *that women between the ages of 19 and 85 comprise 52% of the population.*
- *The fact that a high majority of programs derive funds from their city government is evidence of the prevalence of a strong private/public partnership in Main Street communities.*
- *The criminality of undocumented immigrants was generally attributed to the simple fact that they had no legal right to be in the United States.*

explained/supported/complicated/etc. by the fact that

- *Roten and Mullineaux (2002) find that commercial banks charge lower fees than investment banks. This may be explained by the fact that in their sample commercial banks have only just entered the market.*

given the fact that

- *Given the fact that the period under discussion is far removed from what most of my informants could easily remember, some gaps exist in the data.*

due to the fact that

- *They do not want to become retired, insignificant old people with no meaningful tasks left to accomplish. This view was mainly expressed by males, and may be due to the fact that the sample did not include many women in high academic or administrative positions.*

despite the fact that / in spite of the fact that

- *Participants were informed that there was only one correct, best answer for each question despite the fact that several answers might appear reasonable.*

the fact remains that

- *Although academic motivation has received much conceptual and empirical focus, the fact remains that an abundance of high school students lack academic motivation (Snyder & Hoffman, 2002; Statistics Canada, 2002).*

in light/view of the fact that

- *Our sample is interesting in the light of the fact that most studies have concentrated on collecting data from lawyers working in law firms.*

FREQUENT PATTERNS
to **ignore/consider** the fact (that)
to **highlight/overlook/obscure** the fact (that)
to **contradict/explore** the fact (that)

③ a piece of information that is known to be true

- *It is **a well-known fact** that the Schilderswijk is home to a large number of irregular immigrants.*
- *Indeed, speaking is itself one of those general facts of human nature that make us the creatures we are.*
- *Observations are represented as sets of low-level numeric facts, such as distance.*
- *It must be stressed that that other person only may have 'rights of custody' depending on the facts of each particular case.*

④ an actual occurrence; an actual event

• *For instance, the Holocaust is a historical fact and as such, the subject of historical research.*
• *In his most recent novel Next Michael Crichton blurs fact and fiction to engage with the question of chimeras and the fear, fascination, and discomfort they produce.*

after the fact after the event that has just been mentioned
• *The fall of Negroponte was thus one of the first events in Renaissance history to be recorded in print more-or-less immediately after the fact.*

⑤ **the fact is / the fact of the matter is** Used to introduce a statement in which you make an important point about what has just been said.
• *I don't know what caused this fear of being alone - maybe not having any brothers or sisters. The fact is that yes, friendship is important, but you can count on friendship only up to a certain point.*
• *While successive cabinets have been aware of the tyranny of the militarists, and have suffered from it, the fact of the matter is they have never actively resisted it.*

⑥ **a fact of life** something, often unpleasant, that is true and cannot be avoided
• *Death is an inevitable fact of human life.*

⑦ **a stylized fact** a simplified presentation of an empirical finding

⑧ **after/before the fact** Law after/before a criminal act
• *In keeping with the Western model, this legislation focuses on punishing wrongdoing after the fact, rather than preventative measures.*

⑨ Philosophy a situation, a proposition
• *If the singular fact is, say, a is to the left of b, we must distinguish between the general facts that all things are to the left of b and that a is to the left of all things.*

⑩ Computing The kind of clause used in logic programming which has no subgoals and so is always true (always succeeds).

⑪ **FACT** Biochemistry abbreviation for **facilitates chromatin transcription** *SEE* **FACT**

WORD ORIGIN
Date: 1500-1600
Language: Latin
Origin: *factum* (something done), from *facere* (to make)

feature /ˈfiːtʃə/

MENU

noun

1. characteristic
2. part of your face
3. newspaper article or TV report
4. film

verb

1. have as an important part
2. **to feature in**
3. **be featured in**

feature *noun* [countable]

WORD FORMS: ***feature, features***

① A **feature** is an important characteristic or part of something.

- *We suggest that this development of micro-politics is a **distinctive feature** of the current economy.*
- *The South African industry had a number of features that favoured its survival.*
- *The model presented here retains several **important features** of the original algorithm.*
- *The Green Building Rating System awards points for design features that cover site development, water savings, energy efficiency, materials selection, and indoor environmental quality.*
- *Manual feature extraction was used herein with two examples presented to illustrate the approach.*

FREQUENT PATTERNS
distinguishing/striking/unique/prominent feature(s)
salient/key/characteristic/common/main/essential feature(s)
linguistic/grammatical/semantic/syntactic/stylistic feature(s)
morphological/histopathological/spectral feature(s)
to **share/have/possess/exhibit** a feature or features
to **capture/identify/reveal/highlight** a feature or features
to **extract/select** a feature or feature(s)

② **Features** are parts of your face such as your eyes, mouth, or nose.

- *Evolution of our species has continued since that time with a marked trend toward smaller teeth and less robust **facial features** .*
- *He accentuated her features by highlighting her paunchy cheeks and the circles beneath her eyes.*

③ A **feature** is a newspaper article or TV report that focuses on a particular subject

- *In 2003, the Herald published **a feature on** the appeal of Buddhism entitled: "Is This the Answer to Your Prayers? "*
- *In 1930, popular fan magazine feature writer Dorothy Calhoun was moved to point out that youth could even be a guarantor of stardom after one's death.*

④ A **feature** is a film of standard length. *ALSO:* **FEATURE FILM**

- *Three years after the German reception of Pulp Fiction, there was a profusion of knock-offs, including Fatih Akin's debut feature 'Kurz und schmerzlos' (Short Sharp Shock, 1998).*

WORD ORIGIN
Date: 1300-1400
Language: Anglo-French
Origin: *feture* (shape, form), from Latin *facere* (to do, to make)

feature *verb*

WORD FORMS:
feature, features, featuring, featured

① [transitive] If something **features** a thing or person, that thing or person is an important part of it.

- *Live performances also feature songs by other local or international bands, which may be special requests from members of the audience.*
- *The leading Sydney newspapers featured full-page souvenir photographs of Sydney and its harbour from the air.*
- *The first issue of the 1958 Architectural Journal featured four articles related to the microdistrict.*
- *Gold tree and Silver tree and Snow White are similar because they feature an older woman who is jealous of a younger woman for her beauty.*

② [intransitive] If a thing or person **features in** something, they are an important part of it.

- *Stories about struggles and ordeals **feature prominently** in the reflections of Australian archaeologists on their fieldwork.*
- *A "Magazin" and two other buildings also feature in this area on the 1586 map.*

③ [transitive] If a character or topic **is featured in** an advertisement or piece of research, it has a central role in it.

- *The main character featured in this commercial is an old lady in a rural area.*
- *Four countries are featured in this article: Scotland, Greece, Slovenia, and Israel.*

feature film *noun* [countable]

WORD FORMS:
feature film, feature films

a film of standard length

justified /'ʤʌstɪfaɪd/ *adjective*

① If someone **is justified in** doing something, they are right in doing it.

- *It is not certain that we are justified in speaking of Middle Javanese as a single language.*
- *Were the missionaries justified in requiring that the Indians give up their long hair in order to receive baptism?*

② If something **is justified**, it is done legitimately or proves to be correct.

- *The European Court of Justice considered whether Danish legislation restricting the free movement of workers was justified.*
- *This suspicion of instability proved to be justified when in 1990 a civil war broke out and people had to flee.*

③ Printing (of a text) adjusted in spacing so that the lines begin or end, or both, at the same distance from the margin

- *The subheadings for those texts written by student 1 are left justified, those by student 2 are right justified and those by student 3 are centered.*

justify /ˈdʒʌstɪˌfaɪ/ *verb* [transitive]

WORD FORMS:
justify, justifies, justifying, justified

① If a thing or a person **justifies** something such a decision or an action, it provides evidence indicating why this decision or action is appropriate or correct.

- *Simulation examples were given to justify the theoretical conclusions.*
- *Managers may wonder what the advantages of investing in relationships with customers are and how they can justify this investment.*
- *The Security Council Vietnam justified its intervention in Kampuchea principally on grounds of self-defence but also on humanitarian grounds.*

something is justified by something

- *My decision to buy cigarettes is not justified by the fact that I cannot get rid of my intention to smoke.*
- *Retaining an HEI-based component is justified by a belief that this impacts in some beneficial way on workplace practice.*

② If someone **justifies** himself or herself, they prove their worthiness. If something **justifies** itself, it proves it exists for a good reason.

- *Over the course of the play, Hamlet will justify himself by straddling thought and action, interiority and exteriority, erecting a field of 'outness' around him that has stabilized his challenged masculinity for more than four centuries.*

③ **the end justifies the means** Used to say that the use of any methods, even bad ones, is acceptable if they lead to an important result.

④ **Printing** to adjust the text so that the lines begin or end, or both, at the same distance from the margin

⑤ **Religion** to declare or make righteous in the sight of God

WORD ORIGIN
Date: 1300-1400
Language: Old French
Origin: *justifier*, from Latin *justificare* (to make just) = *justus* (just) + *facere* (to make)

local authority *noun* [countable]

WORD FORMS:
local authority, local authorities

British English the organization in a particular area or city that is responsible for providing public services

- *Every demolition project needs a successful permit application from the local authority before actual work commences.*
- *The children were scattered in nine mainstream schools across four local education authorities.*

method /ˈmɛθəd/ *noun* [countable]

WORD FORMS:
method, methods

a (systematic) way of doing research or some other activity

- *They **used** both qualitative and quantitative **methods** to collect and analyse their data.*
- *We next describe our data source and **method of analysis**, after which we present our statistical findings.*
- *Over much of the last four decades, fertility control methods have been generally available for wealthier women through private health clinics (Barroso 1984).*

method for doing something

- *Categorization tools normally have a method for ranking the documents in order of which documents have the most content on a particular topic.*

method of doing something

- *This facilitated an opportunity for other pupils to argue an alternative method of solving the equation.*

FREQUENT PATTERNS
numerical/statistical/quantitative/qualitative method(s)
analytical/scientific method(s)
alternative/traditional/new/standard method(s)
multigrid/iterative method(s)
method(s) of **inquiry/estimation/assessment**
method(s) of **data collection**
to **develop/devise/improve/discuss** a method
to **employ/utilize/adopt/apply/implement** a method
to **propose/present/introduce** a method
to **validate** a method

WORD ORIGIN
Date: 1500-1600
Language: Latin
Origin: *methodus*, from Greek *methodos* (pursuit of knowledge) = meta- (after) + hodos (way)

obtain /əb'teɪn/ *verb* [transitive]

WORD FORMS:
obtain, obtains, obtaining, obtained

① If you **obtain** a result or data, you get it by doing research or conducting an experiment.

- *These **results** are comparable to those **obtained by** other authors, even though samples had different origin.*
- *A survey approach was adopted for this study and **the data were obtained** by means of a questionnaire prepared in Turkish.*
- *Endurance time was monitored, and the **values obtained** in 3 trials per time point were averaged for further analysis.*
- *There is considerable pedagogical value in allowing students to first obtain an overview, then go into details, and finally, through a process of synthesis, come up with a design.*

obtain information (from someone)

- *The demand information may be obtained directly from customers or estimated from their past ordering history.*
- *He obtained information about beliefs and rites, but does not seem to have been able to witness many rituals or dances at first hand.*

something is obtained through something

- *Indeed, many insights were gained through action research that would not have been obtained through other methods.*

② If you **obtain** something, you get it.

- *As of the spring of 2005, Palestinian journalists first needed to obtain a work permit to enter Israel.*
- *Auto thieves have adapted to these changes by illegally obtaining keys to accomplish their misdeeds.*

obtain consent/permission (to do something)

- *Written informed consent was obtained from all participating patients before enrollment.*
- *After obtaining permission to conduct research from the restaurant's highest-ranking manager, I began collecting examples of Anglo Spanish.*

obtain something from someone

- *Laborers also obtained a variety of basic materials from companies, such as clothing, tobacco, food and containers for food storage.*
- *Secondary antibodies were all obtained from Jackson ImmunoResearch Laboratories.*
- *Ruth B. Grossman obtained her Ph.D. from Boston University in neurolinguistics in 2001.*

③ To **obtain** a substance means extracting it from something by using a particular method or technique.

- *Biopsy was then obtained by scissors from the stroma around the dominant follicle.*
- *In total, 45 samples were obtained from the Lake Grusha core.*
- *A crude membrane preparation was then obtained by high-speed centrifugation (Type 55 Ti rotor, 45,000rpm for 2h).*

④ If you **obtain** a certain skill or ability, you learn it.

- *In the early stages of the research project, the researcher cooperated and worked with various Tamil interpreters, Tamil refugees who had obtained reasonable skills in Norwegian.*

WORD ORIGIN
Date: 1400-1500
Language: Old French
Origin: *obtenir*, from Latin *obtinere* (to take hold of, acquire)

potential /pəˈtɛnʃəl/

adjective [only before noun]

possible or likely in the future

- *In the case of government procurement, the **potential benefits** include cheaper goods of higher quality and greater cost saving.*
- *The erect pose of the deer is a response to a **potential threat**.*
- *The authors give examples of the **potential impact** of climate change on the insurance, automotive and petroleum industries, among others.*
- *They also chopped down fruit trees, which had the effect of depriving people of **potential sources** of food.*
- *Search engines (on the Internet, for example) as such are of course of great value to potential buyers once they know what to search for.*

FREQUENT PATTERNS
potential **problem/conflict/pitfall/bias**
potential **danger/harm**
potential **target/application**
potential **partner/candidate/competitor**
potential **customer/entrant/adopter**

noun

WORD FORMS:
potential, potentials

① [uncountable] possibility or capability to achieve something in the future

have the potential to do something

- *Renewable sources have the potential to provide energy supply for an indefinite period of time.*

potential for something

- *Multimedia instructional environments are widely recognized to have great potential for improving the way people learn.*
- *The expert report on tourism concludes that there is a large potential for development in the area, and it outlines a rather extensive development plan based on a zoning system.*

(the) potential of something

- *The increasing awareness of the potential of computer-assisted language learning (CALL) within ESL and EFL programs has necessitated a broadening of research into its use and effectiveness.*
- *This and other chapters point out that changes are taking place, but that work must be ongoing in order for museums to* ***realize the full potential of*** *their social agency.*

- *Gould explores the potentials and dangers of the Internet.*
- *It is also an ideal testing ground for us to understand how investors respond to innovative events with highly uncertain commercial potentials.*
- *There are a variety of processes that might give rise to influenza strains with pandemic potential.*

② **Electricity** the quantity determining the energy of charge in an electric field or of mass in a gravitational field
ALSO: **ELECTRIC POTENTIAL**

- *Synthesis was carried out in a SDS-water medium (1 mg/ ml) in the presence of pyrrole (0.01 M) at constant potential of 1.0 V for 20 min.*
- *These currents declined slowly at a holding potential of −80 mV until they fully deactivated.*

WORD ORIGIN
Date: 1300-1400
Language: Old French
Origin: *potencial*, from Late Latin *potentialis*, from Latin *potentia* (power)

ribonucleic acid *noun*

SEE **RNA**

RNA /ɑːr en ˈeɪ/ *noun* [countable]

WORD FORMS:
RNA, RNAs

(abbreviation for **ribonucleic acid**)

Biochemistry any of a group of nucleic acids, present in all living cells, that play an essential role in the synthesis of proteins

• *RNA was extracted using Trizol (Invitrogen, Inc.), following the recommendations of the manufacturer.*

significant /sɪg'nɪfɪkənt/ *adjective*

WORD FORMS:
more significant, most significant

① Statistics A **significant** difference, correlation, etc. between the observed value and the hypothesis is too big to be attributed to chance.

- *Wright and Cropanzano (2000, p. 92) identify a significant relationship between staff well-being and their performance in the workplace.*
- *Evenness values, although larger than those reported in, still display* ***a significant increase*** *through time (rs=+0.893, p=0.007).*

statistically significant

- *The results of this final search corroborated the previous findings: men cite themselves slightly more than women (12.1 % as opposed to 11.1 % in articles; 14.4 % as opposed to 13.1 % in reports), but the differences are not statistically significant.*
- *Differences were analyzed by one-way ANOVA test, by using SPSS software and considered statistically significant at P < 0.05 and P < 0.01.*

② A **significant** amount or change is very large or considerable.

- *First, the developer will pay $200,000 in fees, a significant increase in past fee levels.*
- *A significant amount of research has been conducted on this barrier island system as well as the other barriers in Louisiana.*
- *But hostilities were brutal in Croatia, especially in its eastern border region where Serbs comprised a significant portion of the population (see Grandits and Promitzer 2000).*
- *Hookworm infection causes significant loss of blood, resulting in severe anemia.*
- *Price levels and e-mail coupons do not have a significant effect on order size.*

③ A fact or an event that is **significant** is considered important or noticeable.

- *The fact that the British Government was prepared to pay £150,000 for an expedition that identified anthropological activity as a primary objective is significant.*

• *In the 1997 Eurobarometer survey, immigration turns out to be one of the three* ***most significant*** *political or social issues.*

• *Examining the relation between official and individual narratives of the past should be* ***highly significant*** *for the understanding of learning.*

• *The Instituto Cervantes/Spanish Cultural Center in Damascus has played a significant role in promoting the Mediterranean dimensions of Syrian culture through its programming.*

RELATED ENTRIES: SIGNIFICANT OTHER

WORD ORIGIN
Date: 1500-1600
Language: Latin
Origin: *significare* (to signify)

significant other *noun* [countable]

WORD FORMS:
significant other, significant others

① an important person with whom you have a close relationship, for example a parent, friend, or co-worker.

• *Matthew's connection to significant others in his life are primarily mediated by email and telephone.*

② a spouse or lover

• *Finally, for those relatively few individuals who did not have a working spouse, significant other or were not supported, in part, by a divorced or separated partner's income, spousal contribution to family income was scored a zero.*

state-of-the-art, also **state of the art** /steɪtəv-ðiː-ɑːt/

adjective [usually before noun]

State-of-the-art thing or method is the best available because it uses the most modern and recent knowledge or materials.

- *State-of-the-art technologies have often led consumers to skip generations of incremental improvements.*
- *Innovation strategies create state-of-the-art products that are beyond what rivals can offer.*
- *The space center will be equipped with state-of-the-art facilities such as an assembly complex and a ground test facility.*

noun [uncountable]

The **state of the art** is the latest knowledge or methods achieved in a particular field.

- *The current state of the art does not provide us with many applicable proposals for solving these architectural and methodological problems.*
- *Most researchers agree that the current state-of-the-art in clustering is spectral clustering.*
- *A more recent book representing the state of the art in social archaeology, A Companion to Social Archaeology (Meskell and Preucel 2004), is considerably better at including feminist and gender perspectives.*

subsequent /ˈsʌbsɪkwənt subsequent1.wav/ *adjective*

① **Subsequent** activity or event follows something mentioned earlier. *SEE* SUBSEQUENTLY

- *The first analysis includes both tests and **subsequent analyses** examine each test separately.*
- *The next section develops a theoretical model of political representation in committee reports and derives four hypotheses to be tested in **subsequent sections**.*
- *Peak heart rate for the males was relatively stable during the first 3 years in the program before decreasing in subsequent years.*

② **Subsequent to** an event means after that event.

- *Subsequent to the questionnaire the experimenter showed the pupils how to use the program.*

WORD ORIGIN
Date: 1400-1500
Language: Latin
Origin: *subsequens* (following on), from *subsequi* (to follow closely)

subsequently /ˈsʌbsɪkwəntlɪ/ *adverb*

later or after an event that has already been mentioned *SEE* SUBSEQUENT :①

- *By pressing the 'Enter' key on the keyboard they confirmed their choice. Subsequently, the computer notified them whether their judgment was correct or not.*
- *The interviews were recorded on video and the responses were subsequently analysed for common themes.*

take /teɪk/ *noun*

WORD FORMS:
take, takes

① [uncountable] Somebody's **take on something** is their view or perspective on it.

- *Glanvill had his own take on the question, but ultimately saw that this speculative approach had its limits.*
- *And perhaps the journalistic term 'spin' is little more than a popular take on policy discourse.*

② [countable] A **take** is one of a series of recordings of a song, or a scene in a film, out of which the best one is selected for release.

- *McCartney then proceeded during the evening, in a three and a half hour session beginning at 7.00 p.m. (Lewisohn 1988, p. 59), to record his guitar and vocal part onto a four-track tape with the second take marked 'best' (Coleman 1995, p. 42).*

③ [countable] A **take** is a scene that is filmed without interruption.

- *This darkness falls across his features twice during the short take in which he asks for the eyes of the Oracle.*

WORD ORIGIN
Date: 1000-1100
Language: Old Norse
Origin: *taka*, related to Gothic *taken* (to touch)

taken /'teɪkən/

verb

THE PAST PARTICIPLE OF TAKE

adjective

If you **are taken with** something or someone, you find them very interesting or attractive.

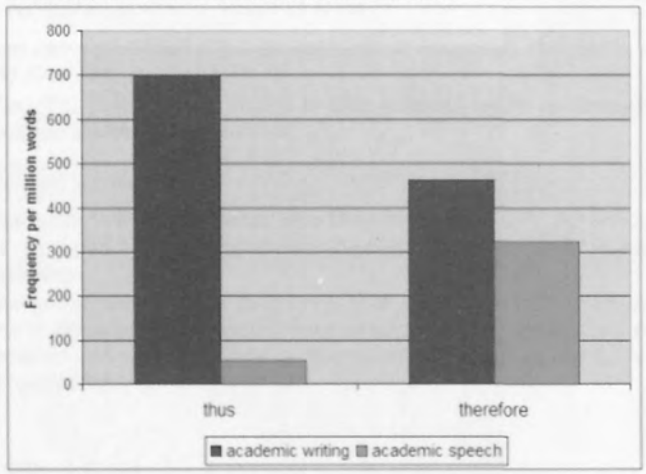
- *During the viewing, Meru and her younger brother were especially taken with their older cousin's new red Audi convertible, whose arrival they replayed several times.*
- *Haydn himself was very taken with contemporary war heroes.*

therefore /ˈðɛəˌfɔː/ *adverb*

used to introduce the result or conclusion of something that has just been mentioned SYNONYM **THUS** :①

- *Modern readers do not approach the Bible with the same information as the implied reader, whose cultural background vastly differed from ours. Therefore, we may be surprised by information that would be obvious to ancient readers.*
- *Two of these students did not sit the final examination and therefore had to be eliminated from the sample.*
- *Many of the countries' leaders understand that they cannot secure their borders on their own, and that they need external funds and expertise . Most have therefore been eager to cooperate with international organizations, although often not with each other.*

Sense 1 of **thus** and **therefore** have the same meaning, but **thus** is used more frequently in academic writing. In academic speech, however, **therefore** is much more common.



WORD ORIGIN
Date: 1100-1200
Language: Old English
Origin: there + for

thus /ðʌs/ *adverb*

① used to introduce a result or conclusion of something that has just been mentioned SYNONYM **THEREFORE :1**

- *The categorization literature is enormous, and no single article could survey it all. Thus, we focus on a restricted subset of the entire literature.*
- *In 1938 the British government passed a Bill ruling that Australian films no longer counted as 'British' for their local quota, thus making them less attractive to British distributors.*
- *Courts seem to be willing to support the view that eBay only functions as a facilitator and is thus not liable for any fraudulent operations.*

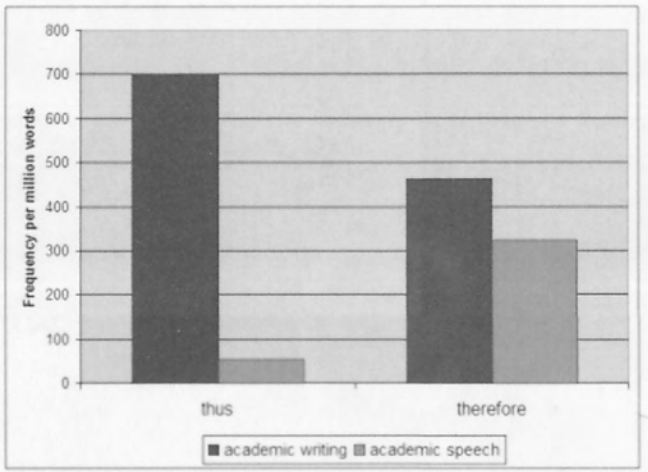
② **thus far** so far, until now, until this point

- *One can argue that the management of economic change in Tunisia has been relatively good thus far.*
- *Thus far,* **(=until this point in the article)** *we have reviewed two prototypical Western European stakeholder settings.*

③ in this manner, in this way

- *The system was evacuated and then heated at 70°C for 4 hours under a continuous flow of hydrogen gas. This gray material thus obtained was used in catalytic experiment with suitable substrates.*
- *The government is the institution that the Prince is in charge of, and Rousseau defines it thus: 'I therefore call Government or supreme administration the legitimate exercise of the executive power, and Prince or Magistrate the man or the body charged with that administration'.*

Sense 1 of **thus** and **therefore** have the same meaning, but **thus** is used more frequently in academic writing. In academic speech, however, **therefore** is much more common.



WORD ORIGIN

Date:

Language: Old English

Origin: thus

thusly /ˈðʌslɪ/ *adverb*

in this manner, in this way SYNONYM **THUS** :③

• *In a work published in 1875 he ranked the following three religions thusly, Hinduism as the closest to Christianity, then Buddhism, and, finally, Islam.*

took /tʊk/ *verb*

THE PAST TENSE OF **TAKE**

various /'vɛərɪəs/ *adjective* [usually before noun]

① If you talk about **various** things, you mean several different things of the same type.

- *Employees who are more dissatisfied with **various aspects** of their jobs are more likely to demand union representation.*
- *With the exception of one, all these films are films that deal with various kinds of natural disasters.*
- *Documents can be categorized in **various ways**, for example, by subject, genre, or the sentiment expressed in the document.*
- *The design projects are presented by the entire team at various stages of the project.*

FREQUENT PATTERNS
various **combinations/factors/reasons/sources**
various **disciplines/techniques**
various **species/tissues/solvents**

② If you describe things as **various**, they are very different from each other.

- *The processes involved could be highly various, and include both horizontal and vertical transmission.*

many and various ***British English*** / **various and sundry** ***American English***

- *The explanations given for this conflict were many and various.*
- *Various and sundry port taxes on spices were collected.*

WORD ORIGIN
Date: 1500-1600
Language: Latin
Origin: *varius* (changing)

23. APPENDIX 13: CD-ROM

CD-ROM Contents:

- File with instructions on how to access and use the material on the CD-ROM
- DOAE database (sample DOAE entries)
- tlReader: free viewer program for viewing the DOAE database (and other TshwaneLex databases)
- word sketches for all the sample DOAE entries
- CAJA lemma lists:
    - complete lemma list (ordered by frequency)
    - lemma list of lemmas with frequency ≥ 5 (ordered alphabetically)