|  |
| --- |
|   |
| Interlingual Identifiers of an L1 German speaker writing in English. |
|   |
|  |
|  |
|   |

Ria Perkins

Candidate Number: 210757

Masters Dissertation,

Forensic Linguistics.

Aston University

Table of Contents

[Abstract 4](#_Toc241864818)

[Chapter 1. – Introduction. 4](#_Toc241864819)

[1.1 Potential and Implications of this Research 4](#_Toc241864820)

[1.2 Orientation to Previous Research 5](#_Toc241864821)

[1.3 Organisation of this Research 6](#_Toc241864822)

[1.4 Hypothesis 6](#_Toc241864823)

[Chapter 2 – Literature Review. 8](#_Toc241864824)

[2.1 Interlanguage Literature 8](#_Toc241864825)

[2.2 Methodology Literature 11](#_Toc241864826)

[Chapter 3. – Methodology 16](#_Toc241864827)

[3.1 Academic definition of ‘Interlanguage’ 16](#_Toc241864828)

[3.2 Structure of this Project 16](#_Toc241864829)

[3.3 Methodology of Analysis of Corpora of Student Texts 17](#_Toc241864830)

[3.3.1. Gathering of Data 17](#_Toc241864831)

[3.3.2 Initial analysis 18](#_Toc241864832)

[3.3.3 Model for full analysis 19](#_Toc241864833)

[3.3.4 Statistical analysis 23](#_Toc241864834)

[Chapter 4 – Discussion and Analysis 25](#_Toc241864835)

[4.1 Details of Analysis 25](#_Toc241864836)

[4.2 Results and discussion of initial analyses 27](#_Toc241864837)

[4.3 SPSS Analysis 33](#_Toc241864838)

[4.4 Chapter Conclusion 36](#_Toc241864839)

[Chapter 5 – Application 38](#_Toc241864840)

[5.1 Introduction 38](#_Toc241864841)

[5.2 Methodology 40](#_Toc241864842)

[5.3 Analysis and Discussion 43](#_Toc241864843)

[5.4 Conclusions 45](#_Toc241864844)

[Chapter 6 – Evaluation of Methods. 47](#_Toc241864845)

[6.1 Introduction and Method. 47](#_Toc241864846)

[6.2 Analysis and Discussion 48](#_Toc241864847)

[6.3 Conclusions 49](#_Toc241864848)

[7.1 Summary of Findings 51](#_Toc241864849)

[7.2 Limitations and Indications 52](#_Toc241864850)

[7.3 Implications 53](#_Toc241864851)

[Bibliography 55](#_Toc241864852)

[Appendices. 59](#_Toc241864853)

# Abstract

This research analyses the potential of interlanguage as an investigative tool. It looks at a combination of academic and ‘real-life’ data sources, and demonstrates that it is possible to ascertain if the author of an anonymous text is a native German speaker or not. The findings have an undeniable significance for authorship analysis as well as implications for the wider investigative and intelligence community. Although this research focuses on identifying German native speakers, the principles will be generalisable, with the intent of being extended to other languages later.

# Chapter 1. – Introduction.

## 1.1 Potential and Implications of this Research

This research is firmly based within the area of authorship analysis – an area of forensic linguistics that is attracting growing attention. In particular this research relates to interlanguage, which according to Grant (2008) and Koppel, Schler, and Zigdon (2005), has the potential to be an invaluable investigative tool. The potential importance of such an investigative tool can be seen in the well publicised case of the Lindberg Kidnapping. Experts unanimously agreed that the ransom note was most likely written by a German national who had spent time in America. Later the German national Bruno Hauptmann was found guilty and sentenced to death. The goal of this research is to develop a new kind of investigative tool for authorship analysis to enable an analyst to ascertain whether an anonymous text in English was written by a native speaker of German. Such information had obvious valuable implications for intelligence gathering and for police investigations. Although this project is developing the tool, concentrating on German, the principles will be generalisable.

## 1.2 Orientation to Previous Research

The term *interlanguage* was first coined by Selinker (1974). It is very closely related to the field of contact linguistics although as Selinker, notes Weinreich’s works have seldom been referred to in the literature on contrastive analysis, interlanguage and second language acquisition (SLA), despite the fact that he “may be the scholar whose insights have proven most important to the continuing discovery of interlanguage.” (Selinker, 1992, p.26). My research will be based on early work by Weinreich and Selinker and more recent research on ‘interlanguage’ in particular Faerch and Kasper *Strategies in Interlanguage Communication* (1983), Kasper and Blum-kulka’s *Interlanguage Pragmatics* (1993), Winford’s *Introduction to Contact Linguistics* (2003) in which he discusses aspects of interlanguage in the context of both language acquisition and contact linguistics. I will also be referring to Thomas McArthur’s almost encyclopaedic book, the *Oxford Guide to World English* (2003), which documents the variation of English around the world including in countries where English is not an official language. Background literature will be discussed more extensively in the next chapter entitled “Literature Review”.

This research is innovative because, unlike the majority of literature on interlanguage, it does not focus exclusively on source texts produced by students. Rather the student texts are be the initial area of investigation. Any style markers identified will then be examined (and possibly added to) in internet language. David Crystal has discussed the much publicised informality of internet language in many works, he illustrates the situation by writing “The electronic medium [...] presents us with a channel which facilitates and constrains our ability to communicate in ways that are fundamentally different from those found in other semiotic situations.” (Crystal, 2006, p. 5) This builds on Labov’s work on field methods for sociolinguistic research, in which he ascertained that style is related to the amount of attention a speaker pays to their speech and in turn that the less attention is paid to speech the more systematic data it provides for linguistic analysis (Labov, Field Methods of the Projecct in Lingusitic Change and Variation, 1984, p. 29). As this research is intended to create a model for analysis in an investigative or authorship analysis scenario, the use of internet language means the research is built on the type of language that is more likely to require analysis in a real-life situation.

While it would be theoretically possible to research this area without an understanding of German, an understanding of German enables a more thorough investigation due to a better knowledge of the possible motivations and influences on the interlanguage in question. It is therefore beneficial that I hold a Bachelor of Arts degree in German. I have also lived and worked in Germany enabling me to gather cultural knowledge that will most likely provide invaluable insight during the linguistic analysis.

## 1.3 Organisation of this Research

This research will be split into two main section of analysis, one that focuses on the student texts, and one that focuses on the internet text. There will then be a third section that evaluates the predictive abilities of the findings.

## 1.4 Hypothesis

 This research will investigate one main hypothesis: that it is possible to ascertain from the language used (including errors and stylistic choices) whether the author of a text written in English, is a native English (L1 English) speaker or a native German (L1 German) speaker.

# Chapter 2 – Literature Review.

## 2.1 Interlanguage Literature

There has been much research relating to Interlanguage. The majority of this research is related to Second Language Acquisition (SLA) takes a pedagogical perspective; with the aim of identifying interlanguage errors in order to prevent them. In contrast this research is interested in interlanguage features, not to eradicate them, but to identify them as investigative tools. Despite the difference in perspectives, a lot of the existing literature is still very valuable to this research. In Hopkin’s article entitled *Contrstive Analysis, Interlanguage, and the Learner* (1982) he wrote that “ CA would be able to predict the errors of the FL learner (cf. Wardhaugh's "strong hypothesis" [1970]) and provide an integrated and scientifically motivated basis for error therapy, textbook construction, etc.” (Hopkins, 1982, p. 32) Although he was referring more specifically to contrastive analysis when he wrote that, it is a recurring theme that can be seen linked to the majority of research in Interlanguage.

There term interlanguage was first coined by Selinker (1974). However, there had previously been much discussion around the topic. The field of Second Language Acquisition has long been interested in the language produced by learners of second languages. Lado (1957) indicated that learner errors could be predicted solely from studying the native language and seeing where the linguistic systems differed from the target language. This form of native language (NL) transfer was later questioned as researchers (in particular Richards, 1971) demonstrated a systematic recurrence of errors that could not be explained by either native or target language influence.

Selinker introduced the term *interlanguage* with the following words:

“Furthermore, we focus our analytical attention upon the only observable data to which we can relate theoretical predictions: *the utterances which are produced when the learner attempts to say sentences of a TL*. This set of utterances for *most* learners of a second language is not identical to the hypothesized corresponding set of utterances which would have been produced by a native speaker of the TL had he attempted to express the same meaning as the learner. [...] This linguistic system we will call ‘interlanguage’ (IL).” (Selinker, 1974, pp. 34-35)

He then continues on to identify three sets of observable data that are relevant to interlingual identifications: “(1) utterances in the learner’s native language (NL) produced by the learner; (2) IL utterances produced by the learners; and (3) TL utterances produced by native speakers of the TL.” (Selinker, 1974, p. 35) Selinker also discusses the *fossilization* of Interlanguage in which “linguistics items, rules and subsystems” (Selinker, 1974, p. 36) become a fixed part of the interlanguage and will generally be impossible to remove despite TL exposure or tuition. In many ways his work represents the early school of Interlanguage, which held the belief that interlanguage was a construction consisting entirely of a combination of influences from the NL and TL. Hopkins demonstrates this in the form of a diagram:

(Hopkins, 1982, p. 36)

McKay and Hornberger explained this early stance with relation to the term *interlanguage* “the *inter*-prefix refers to the notion that the linguistic system that any given learner or community of learners or users had at any particular moment is quantitatively and conceptually somewhere between the first language and the target.” (Mackay & Hornberger, 1996, p. 80)

In the works of later linguists, we see a new school of thought forming, which could be seen as the call to approach interlanguage from a Chomskyian view point, ie. as a unique form of idiosyncratic dialect, or a language in its own right. Many later researchers favoured the Chomskyian approach, that the interlanguage was not merely an attempt at a target language that is heavily influenced by the learner’s native language, instead it should be viewed as a linguistic system in its own right, a language of which the individual learner is the only true native speaker. This could be seen as a form of multi-monolingualism. This approach underpins Corder’s paper *The Elicitation of Interlanguage* (1973) in which he wrote: “We must attempt to describe his language in its own terms, at least in the first instance, and not in those of any other language.” (Corder, 1973, p. 36) This leads us to Tarone’s more description of the basic assumption that underlies interlanguage, that “the language produced by second-language learners is systematic – that is, that there is some organization to be found in the language produced by learners.” (Tarone, Variation in Interlanguage, 1988, p. vii) It is this approach that provides the basis for this research. This research does not focus on the causes of, or the influences on interlanguage. Instead it is interested in documenting the general trends within the English of native German speakers learning English and contrasting this to native English authors. Therefore basing this research on the findings of Tarone and Corder means that this research is not restricted to only finding interlingual markers that can be explained by NL or TL influence. Corder also raises a point that should be emphasised with relation to this research, that if we are investigating interlanguage then we cannot speak of errors in the sense that the interlanguage is a ‘mal-formed’ version of the TL as this negates the principle that it is a language form in its own right. Wode supported Corder’s views by saying that “the developing learner language, whether L1 or L2, cannot be measured as grammatically correct or incorrect according to the standard of the target language (s) involved.” (Wode, 1981, p. 53) This research tries to emphasis this approach throughout, and focuses on how the language is most frequently used. However, occasionally for clarity of language it is sometime unavoidable to use the word *error*, it should be noted that whenever such a word is used in this project it relates to a deviation from the norm and not a critique of the author or their writing.

As German is a widely taught language in the English speaking world, there has therefore been a lot of research into German-English interlanguage. Yet again most of this has a pedagogical or SLA focus, there have been numerous researchers into interlanguage who have based their studies on L1 German learners of English (Nemser, 1967; Hopkins, 1982; Ebert, 1982; König, 1982). However, this is frequently to demonstrate a particular concept or theory and therefore, while of interest to this study, it will not have a direct impact on it or its methods. König and van der Auwera’s book entitled *Germanic Languages* (1994) contains very useful documentations and descriptions of the structures of a wide range of Germanic languages, but most notably of German and English. While not specifically referred to in much of this project it deserves recognition as providing an exceedingly useful background for this research.

## 2.2 Methodology Literature

There are numerous different approaches to authorship analysis. Chaski (2001), MacMenamin (2001 & 2002) and Olsson (2001) all advocate a predominantly statistical approach, which does have the advantage of producing what appear to be very convincing, easy to understand results. In contrast there the much more content based analysis which can be seen in the works of Donald Foster (2001) (anonymous author). It was decided that a multivariate approach, using both qualitative and quantitative methods, would be best suited to this topic. As statistics alone cannot represent the changeable essence of language, yet qualitative analysis alone risks producing little more than speculations about trends within the language. Grant (2008) advocates the multivariate approach for authorship analysis, concluding: “Authorship is itself not a singular activity but has diverse functions and questions with forensic interest can arise out of all of these functions and in many different ways.” (Grant, 2008, p. 227). Tarone relates this to interlanguage, stating that: “*linguistic context* may have a variable effect on the learner’s use of related phonological and syntactic structures,” (Tarone, On the Variability of Interlanguage Systems\*, 1983, p. 142).

Corder states that even if we assume the learner to be the only native speaker of his own idiosyncratic dialect, then we must assume him to make no errors, despite this he may still produce what he calls “slips of the tongue or pen” (Corder, 1973, p. 36) While there are clear instances of such slips in the texts being analysed in this research, it is also important to note the danger of analyst interpretation (it is for this reason that all features are marked, whether they are believed to be slips or not). Hopkins (1982) explained that through asking an author to back translate the text being analysed, an analyst can discover not only considerably more about the author’s intention, but also can discover that what may initially have appeared to be correct, may have been intended to have a different meaning than its actual one and hence be an error. This degree of analysis would not be possible in the real-life situation that this research intends to build a model for, and would therefore have no benefit as part of this research. However, the observation does serve to highlight the assumptions under which an analyst operates under. In order to minimise such assumptions the analyst for the authorship analysis model being designed, cannot risk anymore assumptions as to why certain features occur, but must mark them all, regardless of whether they deem an error to be merely a slip. Corder (1973) also discusses the role of the analyst though speaks of native intuitions and the positive role they play in a linguist (or teacher) performing error analysis. While this research realises that the analyst will always influence the analysis, it is hoped that it is in the positive way of native intuitions helping to identify unique features, rather than assumptions masking what is really happening in the construction of the IL. Corder’s positive opinion can be seen reflected in the method for the analysis. Not only is a close reading of the texts incorporated as the initial stage of each section of analysis, but Chapter Six also evaluates the effectiveness of the analyst’s intuitions.

Tarone, Cohen and Dumas (1983) created a table of communication strategies (Tarone, Cohen, & Dumas, A Closer look at some Interlangugae Terminology: A Framework for Communication Strategies, 1983, p. 6). This is a very useful table for better understanding SLA and has was born in mind during the initial analyses of the student texts, which lead to it contributing two of the markers that were used for the full analysis. Although it is a very useful table, the implications for this research are limited, as it is very difficult to quantify the strategies it identifies and this research focuses more on the recurring interlingual markers rather than the strategies that have been employed to create them. It also generalises interlanguage systems, which Corder (1973) and Wode (1981) said should be unique to each author. However, without some degree of generalisation, it would be impossible to conduct this research, as this research if looking for trends across different German learners of English.

The majority of my quantitative analysis was conducted using SPSS, a computer program for statistical analysis. In order to ensure that my research was conducted with a correct and informed method of using this program I frequently referred to several key texts in the field along with the SPSS help section that is part of the program itself. As there are very few contradictory approaches of beliefs within literature on SPSS, there is very little that can be discussed in this respect. The main text book that was referred to was *Discovering Statistics using SPSS 3rd Edition* this is a very accessible book with clear and concise guidelines on using SPSS for different forms of analysis. I also referred to *SPSS for Psychologists* 4th Edition, although written predominantly for psychology students it is easily transferable to linguistics.

Labov noted that a person’s style of speech varies in register and that the less attention a person pays to their speech the less formal their language is and the “vernacular, in which the minimum attention is paid to speech, provides the most systematic data for linguistic analysis.” (Labov, p. 29) This a very important observation for this research and it underpins the motivation for looking at different genres of text. David Crystal has written extensively about internet language, and the different ways it affects language. One prevailing theme is that the internet is more informal, yet different aspects do have different conventions. This supports the idea that internet language would provide a very appropriate data set for this research. Crystal dedicates an entire chapter of one book to what he terms *Chatgroups* (Crystal, 2006) in which he discussed the linguistic systems that can be seen, and how the chatgroup situation affects the language being produced. He surmises that each group has its own conventions and that there is a great deal of creativity surrounding the language and how it is used and declares that chatgroups are one of the only domains in which one can find language that has not been interfered with: “it provides a domain in which we can see written language in its most primitive state.” (Crystal, 2006, p. 176). A statement that clearly relates to the Labovian principles above.

# Chapter 3. – Methodology

## 3.1 Academic definition of ‘Interlanguage’

*Interlanguage* is a linguistic term that was first introduced by Selinker (1972). The concept builds on existing research in contact linguistics, and particular the works of Weinreich. It is commonly used with relation to second language acquisition (SLA) (Winford, 2003; Lee McKay & Hornberger, 1996; Hopkins, 1982; Selinker, 1992). McKay and Hornberger (1996) explain the term *interlanguage* with the following observation: “The *inter*-prefix refers to the notion that the linguistics system that any given learner or community of learners or users had at any particular moment is quantitatively and conceptually somewhere between the first language and the target.” (Lee McKay & Hornberger, 1996). To a degree this reflects a language ideology of multi-monolingualism.

## 3.2 Structure of this Project

This project is divided into several sections. The first section of analysis focuses on student essays, Chapter 3 covers the methodology, and Chapter 4 the analysis and discussion. This section of analysis is designed to produce some preliminary results to indicate what kind of interlanguage variables may indicate authorship by a native German speaker of different abilities. The second section, Chapter 5, extends the analysis to cover internet texts; to investigate whether the same interlingual variables can be seen and how the change in genre would affect the ability to determine authorship by a native German speaker. The third section of analysis evaluates the ability of the identified interlingual markers to determine authorship by a native German speaker. The limitations will be discussed along with the results of a blind test, designed to quantify the reliability of such analysis.

## 3.3 Methodology of Analysis of Corpora of Student Texts

### 3.3.1. Gathering of Data

The student essays were divided into three groups; Beginner, Intermediate and Advanced (it should be noted that in this research Beginner does not signify German authors that are complete Beginners at English, rather they are the lowest ability from whom it is possible to gather 200 words of writing). The aim was to gather a section of writing from a range of authors over the range of abilities. The intention was to collect approximately 200 words for each author. This will be a sufficient volume of text to establish recurring errors in contrast to one-off mistakes. For each group I gathered essays from twenty different authors per each group, this quantity would make it possible to establish which interlingual markers recurred over several authors in each group or over the groups (in reality I only used five authors from each group, as these rendered acceptable results, and the marginally improved accuracy of results from analysing all the authors would not justify the time it would require). This number of authors will diminish the effect of each individual author’s style or idiolect, yet still be a manageable amount of text to analyse. It was estimated that 2 pages of the students’ exam text book would render approximately 200 words of text.

The texts were donated by a *Gymnasium* (high school) in Hamburg, Germany (for the sake of confidentiality this school will be referred to as *Schule*). The permission letter can be seen in Appendix B. In order to ensure anonymity the texts were all coded according the group, author number and page number. Any overtly identifying items in the text were also blanked out. The authors were selected at random from school groups which their teacher deemed to be representative of each group. The Beginner group was taken from students in year three i.e. the students were thirteen years old, the Intermediate group from year five, i.e. fifteen years old, and the Advanced group from year eight, i.e. eighteen years old. The distribution of gender was as follows:

|  |  |  |
| --- | --- | --- |
| Group | **Male** | **Female** |
| Beginner | 11 | 9 |
| Intermediate | 6 | 14 |
| Advances | 15 | 5 |

It is apparent that there is an uneven distribution of male and female authors, yet this is unavoidable if the authors are to be randomly selected. Artificially choosing the authors would create an unnatural representation of the language of these groups of people. If one were to extend this study later then texts could be gathered that form a representation of the wider distribution and ratios of English learners throughout Germany and other German speaking countries. The texts were all sent electronically by a teacher from *Schule*, they were received in a .jpeg format, the analysed versions can be seen in Appendix C.

### 3.3.2 Initial analysis

The first stage of analysis was a simple read through of five texts from each group, marking any features that appeared unusual or unique to the German-English Interlanguage present. From this basic initial analysis, it was possible to build the following crude list of some features that were present:

|  |  |
| --- | --- |
| * Spelling error
 | * Wrong word / plural error
 |
| * Capitalization error
 | * Error in verb construction / wrong tense
 |
| * Wrong verb
 | * Present tense instead of past
 |
| * Wrong pronoun
 | * Word order
 |
| * Missing article
 | * Superfluous article
 |
| * Missing “that”
 | * Error constructing plural
 |
| * Use of “something”
 | * Starting a sentence with a conjunction
 |
| * Over use of just
 | * Over use of elision
 |
| * Wrong preposition
 | * Appeal to authority
 |
| * Avoidance
 | * Punctuation
 |
| * Non-standard letter style/formatting
 | * Unusual style of writing the letter *I*
 |

The most predominant of which appeared to be; spelling errors, capitalisation errors, the use of a wrong verb, a superfluous article or a lack of article where one would expect one, and the over use of elision. Starting with a close read of the text does not render accurate results, but does render a strong base on which to model the main analysis. As a linguist there is little comment that can be made about graphological features (such as the unusual style of writing the letter *I*). However, it should be noted that I am not being selective as to the type of marker, instead I am looking at the full range of levels of language description at this point.

### 3.3.3 Model for full analysis

From the initial analysis it was possible to build a preliminary list of intlerlingual markers of authorship by an L1 German speaker writing in English. This list was then used to categorise the unique elements in the student essays. It is important to note that this is much more than an error analysis; the aim of this research is not a pedagogical one, it is an investigative one, therefore, it is important to document any element of the language that appears unusual or unique, whether it is technically correct or not. Hypercorrectness could itself be a notable feature. The list of markers that was built is as follows (the full table with definitions can be seen in Appendix R):

|  |  |  |
| --- | --- | --- |
| **Marker** | **Abr.** | **Example.** |
| **GRAMMATICAL** |  |  |
| Lack of Capitalization | -cap | “British museum” (beg01-p1)“what i’m seeing” (int04-p2) |
| Over capitalization | +cap | “I opened the door: They were children” (beg04-p2)  |
| Wrong tense | Ten | “they have been married with another guy” (adv01-p2) |
| * Present tense instead of past tense
 | Pre.pst | “he said he is ill” (beg01-p1) |
| Error in Verb construction | V.con | “they eated it for breakfast” (int01-p1) |
| * Plural verb instead of singular
 | Pl | “A big thing, an UFO, have landed here in New York!” (int04-p1) |
| * Missing Verb part
 | -vp | “we [] shocked” (beg02-p2) |
| Wrong verb part | W.vp | “the American Dream which can’t get true” (adv04-p2) |
| Word order | W.o | “here were crowned kings and queens.” (beg04-p1) |
| Punctuation | Punct | “The woman [...] lives at “Shark Bay”” (int05-p1) |
| ! marker after a greeting | ! | “Hi! [J]!” (beg05-p1) |
| **LEXICAL** |  |  |
| Spelling error | Sp | “thier” (int02-p2)  |
| Noun singular not plural  | -s | “There are lots of stories about shark” (int03-p1) |
| Noun plural instead of singular | +s | “the people who take it out will be our new king” (int02-p1) |
| Error making a noun plural | Pl.er | “many storys about people” (int01-p1) |
| Wrong word | ww | “they have the size of a bear” (int04-p1) |
| Wrong temporal word | w.temp | “We hadn’t seen each other three years ago” (beg04-p2) |
| Wrong pronoun | w.pn | “when he says that this low payed legal or illegal” (adv04-p1) |
| Missing pronoun  | -pn | “but he doesn’t tell\_.” (adv01-p1) |
| Wrong article | w.art | “they eated it for breakfast” (int01-p1) |
| Wrong verb | w.vb | “I came with them” (beg03-p1) |
| Wrong preposition | w.prep | “waited before the museum” (beg01-p2) |
| Missing preposition | -prep | “\_First I didn’t want to go with them” (beg01-p1) |
| Extra preposition | +prep | “they had spend time for walking and talking together.” (adv02-p2) |
| Extra word | +w | “[name] came to me and asked me if...” (beg02-p2) |
| Missing word | -w | “In Australia\_are many sharks” (int02-p1) |
| Superfluous definite article | +art | “we went to the Big Ben” (beg04-p1) |
| Lack of article | -art | “things from\_Roman time” (beg05-p1) |
| Missing “that” | -tht | none |
| Missing “and” | -nd | “wanted to go\_play video games” (beg01-p1) |
| Over use of word “something” | +sth | “tell you something about yesterday!” (beg01-p1) |
| Over use of word “just” | +jst | “we are just ready” (beg01-p2) |
| **STYLE.** |  |  |
| Lack of capital after letter greeting | L.cap | “Dear diary,now Athur is the king,” (beg02-p1\_) |
| Style of writing “I” | I |  Beg01-p1  |
| Starting sentence with conjunctions | Conj | “And I were very bored.” (beg01-p1) |
| Unconvential letter greeting  | Gr | “Hi! [Name]!” (beg05-p1) |
| Unconvential letter ending | End | “Dear [Name]” (beg04-1) |
| Excessive use of elision | El | “what’s” (beg03-p1) |
| Stilted word order | St.wo | “in his jacket was a big hole” (beg01-p1) |
| Stilted word use | St.wu | “we went out the museum” (beg03-p1) |
| Excessive “and” | +nd | “was gone and very broken and in his jacket was a big hole”(beg01-p1) |
| American spelling | US | “realized” (int02-p2) |
| **HOPKINS’ INTERLINGUAL TECHNIQUES** |  |  |
| Appeal to authority | ap.auth | “One of his \* was gone [...]\*=Rad” (beg01-1) |
| Avoidance | Av. | “He didn’t even go into the water\_” (int05-p1) |
| **Other** |  |  |
| Unsure | ? |  |

The markers have been split into rough categories, these hold no significance for the results, but serve to make the analysis easier as the markers can be colour coded for ease. In the case of some of the markers the uniqueness of the feature is not the appearance alone, but the frequency of the appearance. In the case of over use of the words ‘just’ (+jst) or ‘something’ (+sth) it is at the analysts discretion when the word is being over used, but in cases of elision or the use of conjunctions at the start of a sentence the only option is to mark all the occurrences and later compare this to native speaker writings.

These markers were then used to categorise every unique feature of the student texts. Initially the first five texts from each group were analysed, the plan was later to expand this to cover all the texts, but it became apparent that this would not generate enough new information (and the 15 texts already analysed were enough to support any findings already found) to justify the time it would require. A section of analysed text can be seen below. This demonstrates the method used to code the text. The unique section was underlined with the corresponding colour and the abbreviation of the particular marker that relates to it was written in brackets in the margin, as close to the same line as possible.

All the analysed texts can be seen in Appendix C. As the texts are all taken from students’ exam books the essays have already been marked. It is impossible to remove these markings and it would be naive to think they would have no influence on an analysis, yet it is important that they are ignored as much as possible, and not allowed to influence the analysis. They are written from a pedagogical perspective, possibly by a native German speaker and not to the same criteria as the analysis being conducted. Once the analysis was finished the results were tabulated into a frequency table, this can be seen in Appendix D. The results are discussed in the next chapter.

### 3.3.4 Statistical analysis

The next stage was a statistical analysis to ascertain whether it was possible to distinguish which group a text belonged to according to the interlingual features it contained, and if so, which of the markers were the most significant to determining which group a text belonged to. The previous initial analyses were carried out using Microsoft Excel, but SPSS is much more appropriate for statistical purposes, and in particular the analyses I was planning; logistic regression. The first step was to convert the frequency table into a format suitable for SPSS. This involved swapping the table, it was also decided to convert the table to indicate presence (1) and non-presence (0) of each marker within each author’s writing, this can be seen in Appendix E.

During the conversion of the data to SPSS the abbreviations used for the analysis of the texts were automatically altered. This does not affect the results, but in order to fully understand the data in SPSS it is useful to know how the abbreviations are related, the table in Appendix S shows how the abbreviations correspond to each other.

The best test to determine if it is possible to ascertain group membership from the markers, is logistic regression. This test is particularly useful as it is based on linear regression and predicts which category an object is likely to belong to according to certain variables. In this case the variables are the markers I have chosen. However, logistic regression does not count the non-appearance of a marker as being significant. This is particularly useful when one considers that many of the markers are highly content based. One would not find an unconvential letter ending in a literature essay, yet the lack of this marker is not significant. It is also highly likely that given more text a marker that is not present in a short essay, would be present in a longer essay or a different selection of text. Logistic regression is therefore a very appropriate statistical method here, as it does not count the lack of a marker as significant. The statistical program SPSS has the capacity to automatically analyse data in many different ways, this includes a function to do both binary and mutinomial logistic regression. For this stage of the analysis I will use the multinomial logistic regression function.

Multinomial logistic regression is very similar to binary logistic regression, but where binary logistic regression is used when there are just two categories of variable, multinomial regression can be used with considerably more variables. As there are 45 variables initially in this analysis, it clearly requires multinomial logistic regression. Multinomial logistic regression can also be called polychotomous logistic regression, like SPSS this project will refer to it only as multinomial, but it should be noted that they are the same thing. The importance and usefulness of multinomial logistic regression as an investigative and intelligence tool is recognised by Weiseburd and Britt in their book *Statistics in Criminal Justice* (Weisburd & Britt, 2007, p. 581).

The process of the analysis will be further discussed in the following chapter in conjunction with the results.

# Chapter 4 – Discussion and Analysis

## 4.1 Details of Analysis

Through doing the analysis, it rapidly became clear that another analyst would most likely interpret and categorise the texts in very different ways. I tried to make the definitions as precise as possible, where necessary modifying them as the analysis progressed; despite this it is virtually impossible to eliminate all uncertainty. It is also impossible to eliminate analyst error there are a number of markers that the analyst is bound to miss, or not considered to be a marker at that time. However, the effects of these two issues on the analysis can be greatly diminished by maintaining consistency throughout all of the analyses. The most effective way of doing this is to ensure that the same analyst performs all of the analyses, with the same method and conditions, this will eliminate interrater error. It is impossible to eliminate intrarater reliability, though constancy can be improved through having precise definitions and trying to perform the analyses under the same conditions within a reasonably short space of time. In cases where the analyst is unsure of a particular convention in English (for example whether a certain word should be capitalised), then Google is to be used as a basic corpus. Although this is not a precise corpus, as the results change daily, it does demonstrate if one feature is a more prevalent convention than another.

The analysis also highlighted that many of the markers overlapped with each other. This is to be expected to some extent, as it is a balance between being able to classify all the markers present in the text and having so many labels that it is impossible to identify trends of markers. In order to preserve the integrity of the analysis is it important to minimise assumptions about the author’s intentions. This sometimes means that an oddity in the text may have to be categorised as a marker that the analyst does not feel accurately represents what happened. E.g. “Sealines” (int01-p1) instead of *Sea-lions*, this is classified as a spelling error, as sealine is not a standard word. However, it is most likely that this is not a random spelling error, but that the confusion was between the word *lion* and *line*. It should be borne in mind when one is looking at the rankings that many of the markers overlap and are interrelated. For example when looking at the total frequencies of the markers (this is discussed fully later), one can see that there were 78 occurrences of the wrong word being used and therefore this is the fourth most frequently occurring marker. However, there are also 29 occurrences of the wrong preposition being used and 8 of the wrong verb, which would result in 115 occurrences of the wrong word being used. There were also instances during the analysis where what initially looked like a spelling error had to be classified as a verb construction error, for example “ley” (beg02-p1). Without eliciting more information about the author it is impossible to ascertain whether this is an error made while trying to decline the verb, or a spelling error. As it is a verb, it was decided to mark it as a error in verb construction.

 Spelling error is ranked as the highest, with a total of 89 occurrences and with an average of 2.97 occurrences per author, and error in verb construction is ranked as the fifth highest with 53 total occurrences per author and an average of 1.77 occurrences per author. It is important to bear in mind that these two markers can be very similar, during analysis it was frequently very hard to distinguish between the two (see above). One word would fit in the classification for one marker, yet experience would indicate that the error was motivated by another marker.

As discussed in the previous chapter the texts have already been marked by teachers at *Schule*. I stated in the previous chapter that it would increase the validity of the analysis to ignore these corrections as much as possible. However after the analysis it was interesting to notice that these corrections often differed significantly from my markings. The difference in motivation for marking was frequently demonstrated, my motivation being to observe and document what was happening, the teachers being a pedagogical motivation. This difference is highlighted through the numerous instances in which the teacher’s marking consisted of writing the preferred version next to an error, or assuming what the pupil meant and writing that next to the text. One could also attribute this difference in motivations as the reason why the teachers hardly ever corrected stylistic features that would commonly be seen as wrong in English, such as the unconventional greetings in letters.

## 4.2 Results and discussion of initial analyses

The observations from the initial close-read through have already been noted in Chapter 3 and indicated that there were indeed recurring trends across the authors. The next stage of analysis was the full analysis of the student texts. The analysed essays can be seen in Appendix C, the frequency of each marker was then tabulated. The full table of frequencies can be seen in Appendix D and a graph displaying the total frequencies for each group can be seen below.

The first observation is that many of the markers identified in the initial read through frequently occur in the texts as suspected. There are however, also other markers with a very high frequency, that were not predicted by the initial read through. This supports the concept that an initial read through is a useful analytical tool, which creates a strong base on which to plan a fuller analysis, yet it does not render accurate results on its own.

It is key to note that every text contains some markers, the range minimum number of markers in a page is six, the maximum found was 55. This relates to a minimum of 24 markers for one author, i.e. in approximately 200 words, and a maximum of 94 markers for one author in approximately 200 words. This equates to between 0.12 to 0.47 markers per word. This indicates that in order to perform an adequate analysis, it would be necessary to have a significant volume of text. As is frequently the case in forensic linguistics: the more text, the more accurate the analysis. The presence of just a few markers could clearly not indicate authorship by a native German speaker, as many of the markers are frequently found in texts by native English speakers. One can regularly find authors that do not just regularly start sentences with conjunctions, but also frequently start paragraphs with conjunctions. For example John Olsson starts a paragraph with the words: “But first let us consider the role of the emergency operator” (Olsson, 2004, p. 144) and a recent BBC online article started a paragraph : “And he denied rumours about his health and insisted his eyesight is not "deteriorating".” (News, 2009) Therefore, as one can expect many of the markers to be present to some extent across all the groups and in the writings of native speakers, one would need a significant volume of text in order to generate enough markers to determine which of the groups an author belongs to.

The frequencies for each group can also be displayed in a graph format:



This graph displays the vast differences in frequencies between the markers. It also demonstrates the point that has been previously made, that this analysis is not just error analysis. If it were the one would expect the Beginner line to be consistently higher than the Intermediate and Advanced. This is obviously not the case, instead they all fluctuate independently.

One interesting feature that was most prevalent in the Beginner texts, was the style of writing the letter I. There are two main styles of writing the letter *I* in the Beginner texts these are:

 (beg01-p1)

(beg02-p1)

These can be seen throughout several of the authors in the Beginner group. In contrast we have a much more traditional English style of writing the letter *I* being used consistently in the more Advanced texts:

(int04-p2) 0 (int05-p2)

The distribution chart clearly shows the evolution throughout the texts. The importance is not the number of occurrences per author, as the majority of authors either have a unique style of writing the capital letter I or they do not, the number of occurrences within an author’s texts is dependent on the content of the text. However, the number of author’s in each group that do not use the traditional British style of writing the capital letter I is noteworthy. The following table clearly shows the evolution, with all of the Beginner authors using a distinctive style of writing the letter I, but only one of the Intermediate authors and two of the Advanced authors.



Although this is an interesting observation and demonstrates a clear evolution and difference between the groups, there is very little more information that can be drawn from this, as it is a graphological observation not a linguistic one.

During the analysis it became clear that there were some very interesting motivations and trends behind the spelling errors and related markers (eg. V.con). I compiled a table of all spelling errors and some of the other markers that were of interest, this can be seen in Appendix F. Due to time and length constraints it was impossible to examine this properly within this study, so these observations can only be considered indications at this point.

As expected there were a few errors that appeared to be influenced by the German language, e.g. communistic being spelt: “kommunistic” (adv04-1). There are also spelling errors where one does not just see the misplacing of letters, but the mistake revolves around a cluster that occurs many other English words, but not the one it is used in e.g. “coughtn’t” (beg02-p1) being used instead of *couldn’t*. This would indicate that the authors therefore have an understanding of how letters cluster in English words (the concept of consonant clusters within a language is discussed by Pinker, 1994). Interestingly there are also errors that appear to be influenced by a Germanic pronunciation of the English word, e.g. crazy being spelt “crasy” (beg04-p1), as well as errors that appeared to be influenced by German spelling systems, e.g. interview spelt “interwiev” (int05-p2). It would be very interesting to study the extent of interlingual and intralingual, orthographical and phonological influences further.

There was also one author who demonstrated a possible French influence in their work. Beg05 demonstrated two features in particular that seemed to be caused by a French influence: “favorit” and “ballon”. It could be that this author does in fact have a French linguistic background, through either being bi-lingual or also studying French, however this can only be speculation as I do not have access to a detailed linguistic background for this author.

It is useful to see which are the most frequent markers. Although we do not have the data to compare the marker frequencies with a native English author, this can give us an idea of which markers are common among a native German speaker writing in English. The twenty most frequent markers are as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Marker** | **Abr.** | **Total Occurrences** | **Average Number of Occurrences per author** | **Overall Rank** |
| Spelling error  | Sp | 89 | 2.96667 | 1 |
| Stilted word use | St.wu | 79 | 2.63333 | 2 |
| Style of writing “I” | I | 78 | 2.6 | 3 |
| Wrong word  | ww | 61 | 2.03333 | 4 |
| Error in Verb construction  | V.con | 53 | 1.76667 | 5 |
| Wrong tense | Ten | 44 | 1.46667 | 6 |
| Excessive use of elision | El | 39 | 1.3 | 7 |
| Punctuation | Punct | 38 | 1.26667 | 8 |
| Starting sentence with conjunctions | Conj | 32 | 1.06667 | 9 |
| Wrong preposition | w.prep | 29 | 0.96667 | 10 |
| Missing word | -w | 20 | 0.66667 | 11 |
| Lack of Capitalization | -cap | 17 | 0.56667 | 12 |
| Unsure | ? | 17 | 0.56667 | 12 |
|  Missing Verb part | -vp | 15 | 0.5 | 14 |
| Word order | W.o | 14 | 0.46667 | 15 |
| Noun singular not plural  | -s | 12 | 0.4 | 16 |
| Lack of article | -art | 10 | 0.33333 | 17 |
| Over capitalization | +cap | 9 | 0.3 | 18 |
| Stilted word order | St.wo | 9 | 0.3 | 18 |
| Wrong verb | w.vb | 8 | 0.26667 | 20 |

It is very interesting to note how quickly the total number of occurrences drops. There are a total of 45 markers, one would assume therefore that the top twenty occurrences would all occur relatively frequently. Yet the table above clearly indicates that only the top ten most frequently occurring markers occur at least once per author. There are only eight occurrences of the wrong verb being used in a total of approximately 3000 words, this gives an average of 0.27 occurrences per author. This is so low that the chance of an author demonstrating this marker in their work is nearly equal to random chance. Later we will use SPSS to more accurately determine which markers are significant to deciding which group an author belongs to. However, at this stage it is interesting to note that these results appear to be following a log-normal distribution curve. This is a very common pattern for elements of language, as Zipf demonstrates in his 1932 word entitled *Selected Studies of the Principle of Relative Frequency in Language*.

## 4.3 SPSS Analysis

The next stage of analysis was a statistical one to determine how different the styles of each group are, and which markers are the significant distinguishing ones. For this analysis the program SPSS (version 15) was used. This is a statistical program capable of analysing large amounts of data quickly and accurately in many different ways. The data was manipulated as discussed in Chapter 3. Then the automatic function within the program was used to complete the multinomial logistic regression. The full results chart can be seen in Appendix G. The program warns that there are too many variables (i.e markers) and that this may be producing results that are too accurate in that they may be too tailored to this set of texts and could not be expanded to include any new or unknown texts.

 As it is the intention of this research to create a model that can be expanded to analyse new texts as part of the author analysis, this warning indicates the need to reduce the number of variables that we are focusing on. As a part of the analysis, under the section called “parameter estimates” the program calculated the *Wald chi-square* statistic, which indicates the significance of each marker for determining how valuable it is on its own for determining which group (Beginner, Intermediate or Advanced) a text fits into. The following table from the SPSS analysis indicates that using all the markers selected the program can accurately classify which group an author fits in 100% of the time. This is disturbing as it indicates that the markers are too specific to this group of texts and might not be generalisable. Due to the laws of probability it is also impossible to achieve 100%, or total accuracy when predicting something.

**Classification**

|  |  |
| --- | --- |
| Observed | Predicted |
| 1 | 2 | 3 | Percent Correct |
| 1 | 5 | 0 | 0 | 100.0% |
| 2 | 0 | 5 | 0 | 100.0% |
| 3 | 0 | 0 | 5 | 100.0% |
| Overall Percentage | 33.3% | 33.3% | 33.3% | 100.0% |

(1 stands for the Beginner group, 2 for Intermediate, and 3 for Advanced)

 When one rearranges the results according to which markers have the highest *Wald chi-square* statistic we get the following results in the table below. Multinomial logistic regression requires that one group be a reference category. In this case SPSS automatically chose the Advanced group as the reference category (this has no bearing on the results), it then compared the Beginner group to the Advanced group, then the Intermediate group to the Advanced group. Therefore, when SPSS calculates the Wald chi-square statistic it calculates the significance of each marker in the Beginner group for differentiating a text as Beginner not Advanced, it then does the same for the Advanced group. It does not need to calculate this for the Advanced group as this is the reference category. Therefore, we end up with *Wald chi-square* statistics for each marker for the Beginner group then the Intermediate group as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| **Abbreviation (SPSS)** | **Wald (Lower Bound)** | **Abbreviation (SPSS)** | **Wald (Lower Bound)** |
| **Beginner Group** |  | **Intermediate Group** |  |
| [s\_A=0]  | 0.00000133423848112014 | [Ten=0] | 0.00000242721601361770 |
| [Ten=0]  | 0.00000060520474489534 | [Pre.pst=0] | 0.00000204449041567073 |
| [cap=0]  | 0.00000051752215937747 | [s=0] | 0.00000201937677354143 |
| [vp=0]  | 0.00000048857867374321 | [cap=0] | 0.00000091123800626597 |
| [s=0]  | 0.00000046781834818093 | [Pl=0]  | 0.00000073264532673834 |
| [Pre.pst=0] | 0.00000043282713656441 | [vp=0] | 0.00000050867844906593 |

Choosing the markers that have this highest *Wald chi-square* statistic means that we will have the best determiners (even though they are not significantly better than the other markers when one their own). The fact there is not a big difference in the *Wald chi-square* statistic and each marker has an exceedingly low *Wald chi-square* statistic, combined with the high prediction rate means that it is not just the presence of a few markers that indicates different group memberships, but that the combination of all the markers is a very powerful predictor. In order to reduce the number of variables as SPSS suggested, I focused on the six highest ranked markers from each category. It is interesting to note, that the five of the six most significant markers were the same for the Beginner group as the Intermediate group, giving a total of seven different variables that were significant over the two groups.

Once we have reduced the number of variables (markers) we can then re-run the multinomial logistic regression using only these seven variables. In the hope that this will eliminate the initial error in the previous analysis, that was caused by having too many variables. Although this did not eliminate the warning, we have reduced the number of variables from 45 to seven and reducing it anymore would not be wise when each marker has such a low *Wald chi-square* statistic. The reduced variables did mean that it was slightly less difficult to predict group membership, with the markers indicating that one of the Beginner texts should be in the Intermediate group. Fourteen out of fifteen being correctly predicted though is still an exceedingly high percentage of accuracy.

**Classification**

|  |  |
| --- | --- |
| Observed | Predicted |
| 1 | 2 | 3 | Percent Correct |
| 1 | 4 | 1 | 0 | 80.0% |
| 2 | 0 | 5 | 0 | 100.0% |
| 3 | 0 | 0 | 5 | 100.0% |
| Overall Percentage | 26.7% | 40.0% | 33.3% | 93.3% |

It should also be born in mind here that the grouping of the authors is rather imprecise, as one would expect a proficient Beginner to be creating text very similar to a less proficient Intermediate author.

## 4.4 Chapter Conclusion

There are certainly recurring markers throughout all of the author’s texts. This supports the initial hypothesis that there are interlingual markers that can aid in authorship analysis, by determining whether an author is a native German speaker or a native English speaker. The results do prove that it is possible to ascertain which proficiency level an author fits into. Though it is uncertain how this will transfer to other genres, as the markers might be too precisely fitted to these particular texts.

It is clear that the markers selected hold a high predictive power in combination for predicting whether a text belongs to a Beginner, Intermediate or Advanced L1 German speaker writing in English. This indicates that the same markers will also be able to separate when investigated in another genre, will be able to differentiate whether a text belongs to a native German speaker writing in English or a native English speaker. The only limitation with differentiating between the three different levels of German speakers, is that there are too many variables. Rather than discarding some of the variables now, it would be better to investigate all the variables in the internet genre, and then determine which markers are the best predictors of an L1 German speaker as opposed to an L1 English speaker. This is because one cannot assume that the markers that are the most significant at distinguishing which proficiency level an author is, will also be the best for determining an author’s mother tongue.

These analyses indicate that the markers will be successful predictors of whether an author of an English text is a native German or English speaker, however, this can be no more than an indication until there is data for a native English speaker writing in English. The fact that many of the chosen markers are occur widely in native speaker’s writings has already be discussed, the next stage would therefore be to collect data of a native English speaker writing in English and compare this to writing from a native German speaker writing both in English and the same context. As discussed in the introduction, student texts differ greatly from ‘real-life’ texts that may require authorship analysis. Therefore, it is also necessary to investigate these interlingual elements as indicators of authorship by a native German or English speaker in a different genre.

One limitation of this study is that the analyst, me, is mildly dyslexic. This may therefore affect the intra-rater reliability, however, all precautions were taken as detailed above to minimise the affects of this. The possibilities of some minor inconsistencies are not a major problem as the aim of this research is to prove that in principle, it is possible to build a model for analysing whether an anonymous author is native German speaker.

# Chapter 5 – Application

## 5.1 Introduction

As already discussed the majority of the research on interlanguage focuses on student or academic texts, and its primary aim is normally a pedagogical one. While such research has uncovered some very interesting findings, it does not have much of an impact for the world of authorship analysis, or for investigative linguistics. This is because (with the exception of plagiarism) there is seldom reason to complete authorship analysis on an academic text written by a student. This raises the question of what type of texts would it be valuable to base this research on? There is no simple answer for this, but what is clear is that texts cannot be artificially produced for this purpose of this research. The internet is the biggest source of readily accessible texts from many genres, including what Crystal (2006) calls *chatgroups*. It has already been demonstrated that a non-real time discussion forum would elicit data that would

As previously discussed in order to create a model to determine whether it is possible to predict whether an anonymous author is a native German speaker or not, it is important to base this research on ‘real-life’ data, which is more likely to actually require analysis. The first stage of this section therefore, was to select texts that could both conceivably require analysis in a real-life scenario, and also elicit reliable data. The trends that were identified in the previous chapter can then be analysed in this different genre, to see in to what extent they are transferrable, and how the change in genre affects the ability to predict if an anonymous author is an L1 German speaker.

There were some significant differences between the student corpora and the internet corpus. The main difference is the difference in genre; one being written during an exam with the knowledge that the quality of the English used will later be marked, the other written on the internet with the purpose of sharing ideas, comments and questions with a group of people all on an online white supremacist web forum. This change in genre means a change in content, reason for communication, formality, standard style, among other things.

The second major difference between the corpora is the mode of writing; the student texts were all written by hand, whereas the internet texts were all written through a computer. It should not be assumed that all the authors using a computer are necessarily using a keyboard, there are other methods of writing through a computer (e.g. voice recognition, or notebook technology which recognises handwriting), however there is evidence that many of the authors were using keyboards, due to spelling errors where it would appear that the key next to the desired one was pressed, or errors in key order. E.g. “you wil lall say” (Michael Wittmann). There are also some deliberate ‘spelling errors’ that have evolved due to conventions of technology (or within the field of White Nationalism), e.g “women kick a$$”(Michael Wittmann). One must also consider the impact of the software. Many programmes that involve writing now come with automatic spell checking software – this can highlight when you have made an error, suggesting how you should correct it, or it will also sometimes autocorrect an error. While one would expect all these things to affect the results slightly, the analyst cannot assume that certain features are purely due to the technology, as this would lead to the analyst’s assumptions influencing the results to a greater extent. It would be another study entirely to ascertain the extent to which computer use affects the language and hence the interlingual identifiers discovered in this study.

The extremist website *Stormfront* (<http://www.stormfront.org/forum/>) is a white supremacist discussion forum, for users based all around the world. It was founded in 2000 by former Klu Klux Klan “Grand Dragon”, Donald Black. The slogan for the website is “White Pride Worldwide”  (STORMFRONT, 2000) and it contains different sections for users from different parts of the world to discuss different topics. It is the focus on bringing members together from around the world that makes it such a useful site for this research, as the predominant language is English, even though members are from a variety of different linguistic backgrounds. It should be noted that the views and opinions expressed on the site are incidental and in no-way supported by this research. All posts are in a public domain and therefore, they do not need to be anonymised, and you do not need to be a member to access them.

By focusing on members of Stormfront and their posts, it ensured a degree of commonality. There was a recurring theme of white supremacy throughout the posts (though this was not the only focus of the posts), focusing on the one website also means that there is only one set of forum conventions to influence the works.

## 5.2 Methodology

The aim was to build two corpora of internet texts, one of native German speakers writing in English and one of native English speakers writing in English. The methodology for the analysis of the internet corpus was designed to be as similar as possible to the methodology for the student corpus analyses. This is to ensure that there are as few variables as possible to affect the results of the analyses. For this reason the same set of markers was used when analysing the internet corpus. Like the student corpus, all the analyses were done by me, and to under the same conditions (e.g by hand, with the same coding system). It was decided that ten L1 German members and ten L1 English members would provide sufficient inter-author variation across this new genre to identify inter-author trends, and that approximately 1,000 words of text would demonstrate a range of intra-author variation, and enable slips to be differentiated from recurring features.

The limitation of choosing data from a public forum, where no contact is made with the initial author, is that there is no way to validate what the author’s native tongue is, or whether they are operating under an on-line persona and pretending to be from a different linguistic background. During the collection of the texts I noticed many members of the forum were keen to emphasis their Germanic roots, sometimes stating that they were German, when other information they supplied appeared to contradict this statement and indicate they did not fit into the definition of ‘German’ for this research (i.e. someone who’s mother-tongue is German). One cannot take geographical information into consideration, as there are many native German speakers who live in English speaking countries. It was decided that if an author states that they are German, and do not explicitly contradict this (for example by claiming to be of several different nationalities) then they must be treated as a German author. For this reason a search was performed for all posts in which a member of Stormfront wrote the phrase “I am German”, a more extensive search for all the posts from chosen members was then conducted and 900 - 1,000 words of text collected per member (the link to the post in which they stated the phrase *I am German* is included directly under the name of each member in the corpus). The same method was used for building the native English speakers corpus. As the site is an American one, with an American founder and prevalence of American members (many of the members selected for the German corpus indicated living in America) it was decided that the corresponding search would be conducted with the phrase “I am American”. Texts were actually collected and analysed for eleven American authors, but the frequencies of the second author usinfantryMOS11C, were not included in the statistical analysis as not enough texts could be gathered to reach the approximate 1000 words required. So although he texts are apparent in the appendix, they were not part of any of the further analyses.

Many members extensively quote other members, these quotes were removed and replaced with “[Quote]”, as the linguistic background of those quoted was unknown and would hence, have a detrimental impact on any findings. There were also several authors who had an automatic ‘signature’ that appeared at the end of each post. The following is Norsewulf’s ‘signature’:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
Adolf Hitler: April 20, 1889 – April 30, 1945.

David Lane: November 2, 1938 – May 28, 2007.

Heil Odin! The Allfather.

As the posts were frequently shorter than the ‘signatures’, which it can be presumed were automatically generated (as each author’s signature is identical across all of their posts), it would unduly influence the findings if every occurrence was included in the total word count of the corpus. Therefore, the signature was included as part of the corpus for the first post collected from each author, but in all the recurring instances is was replace with “[repeated signature]”.

 Although it became apparent that there were new markers appearing in the internet corpus, that had not been noted in the student corpus, I avoided adding any new categories to the list I was looking for, as this would have altered the results unless I re-analysed the student corpus. It was decided that it would be best to fit the new markers into the already decided groups, then investigate later if the markers take different forms in the two different genres.

The texts were analysed and annotated following exactly the same method used for the student texts, this is to minimise the number of variables that will impact of the results and findings. The analysed texts can be seen in Appendix J and Appendix K. The results were then tabulated (the table can be seen in Appendix L), the results then converted to an acceptable format for SPSS and inputted into SPSS. The multinomial logistic regression function was then used, the results of which can be seen in Appendix M and Appendix N and are discussed below.

## 5.3 Analysis and Discussion

As one would expect the same warnings appeared when performing the logistic regression on the internet texts as the student texts. This is not surprising, as there was no alteration to the variables used for the text analysis. However, like with the student analysis, the next stage was to reduce the number of variables to the most reliable ones according to the Wald chi-squared statistic (the table with the variables ranked according to their Wald chi-squared statistic can be seen in Appendix O). As with the student texts there was not a significant difference between which markers were the most predictive indicators of group membership. The ten most significant variables were: pn, w.art, v.con, ten, w.vp, w.vb, vp, w.prep, pl.er, w. The classification table from the logistic regression with all the variables was as follows (1 stands for L1 German speaker, 2 for L1 English speaker):

 **Classification**

|  |  |
| --- | --- |
| Observed | Predicted |
| 1 | 2 | Percent Correct |
| 1 | 10 | 0 | 100.0% |
| 2 | 0 | 10 | 100.0% |
| Overall Percentage | 50.0% | 50.0% | 100.0% |

Yet again there is a hundred percent accuracy – an indicator that the variables are too closely fitted to these texts and not generalisable. As expected the accuracy of predicting group membership decreased with the reduced number of variables, though we can see that it is still very accurate:

**Classification**

|  |  |
| --- | --- |
| Observed | Predicted |
| 1 | 2 | Percent Correct |
| 1 | 7 | 3 | 70.0% |
| 2 | 0 | 10 | 100.0% |
| Overall Percentage | 35.0% | 65.0% | 85.0% |

It is interesting to note that none of the native English speakers were mistaken for being a native German speaker, though the reverse did happen. There are two main explanations for this; the first is that the German speaker was an exceedingly competent learner of English, the second relates to the already mentioned issue of online personas, in particular related to the noted tendency of some Stormfront members to emphasise their German heritage. It does indicate that a native English speaker is exceedingly unlikely to be identified as a native German speaker, which means a false identification as an author as being an L1 German speaker is very unlikely.

The variables focused on during the student analysis were slightly different to the ones highlighted during the internet analysis as being significant. The difference is the significance of the markers can most likely be attributed to the difference in genre. However, it is interesting that two of the markers overlap: (ten) the use of the wrong tense, and (vp) the lack of a verb part.

There were an interesting number of L1 English authors that used the word *then* instead of the word *than*. This may possibly relate to American pronunciation of these words. Although this does not affect these results, it is an interesting observation, especially when taken into consideration with the earlier observation of the influence of Germanic pronunciation on spelling in the student texts.

## 5.4 Conclusions

This section of analysis does indicate that it is possible to determine whether an anonymous author of a English text is a native German of a native English speaker. The results indicate that while you are less likely to attain a false result that an author is a native English speaker, it is exceedingly unlikely that the result will indicate authorship by a native German speaker, when in fact the author is a native English speaker.

It would be interesting to investigate how the markers overlap across a wider range of genres. It could be coincidence that the use of the wrong tense and the lack of a verb part are significant markers in deciding whether authorship is by a native German or English speaker in both student texts and internet texts, though it could be that these markers are important determiners across a wider range of genres. It would also be a valuable study to expand this study across different types of internet forums, to see if different forums and their conventions impact upon which markers are the most significant and on the reliability of the test.

As the results of this test indicate that it is possible to predict whether an author is a native German speaker or not, the next logical step would be to test the reliability of such analysis through a blind test. This next section of analysis will be detailed in the next chapter.

# Chapter 6 – Evaluation of Methods.

## 6.1 Introduction and Method.

The simplest way of evaluating these markers, is to test their accuracy through a blind test. In order to do this is was necessary to elicit the help of a colleague who, using the same search method I had employed, found five native German speakers writing in English and five native English speakers. They cross checked these against the members I had already used in my corpora, ensuring that none of them overlapped. They anonymised the texts and mixed the authors so that there was no discernable order relating to the native language. Due to a mixture of time constraints and the fact that authorship analysis rarely has the luxury of long texts, it was decided that there should be approximately 200 words per author, this can be made up of either one, or multiple posts. This would be a more realistic representation of a ‘real-life’ analysis in which such analysis may be necessary. It would also indicate whether the analysis is still reliable with a reduced volume of text.

As with the previous analyses, the initial stage of the blind test is a close-reading of the texts. In some of the cases intuition that has been honed through previous analyses, indicated whether an author was a native German speaker or not. I then made a note of this (the full results are tabulated below).

In the previous chapter I identified which were the key markers for determining which group (native German speaker or native English speaker) an author belonged to. Therefore, the next stage was to focus purely on these markers and how often they occur within the texts.

## 6.2 Analysis and Discussion

During the analysis it became apparent that when one reduces the markers in this manner and then analyses the texts it affects the results. For example if a word was missing in the student or internet corpora, then it could be classified as a missing word, but only if it didn’t fit into one of the subsections of missing words (eg. –prep, -pn, -vp) which were made markers in their own rights, so that we could ascertain whether one frequently occurring type of word had more indicating power than another.

The following table shows the results of this analysis.

|  |  |  |  |
| --- | --- | --- | --- |
| Author | Analyst’s Intuition | Markers  | Author States |
| 1 | L1 English | 4 | American |
| 2 | L1 German | 2 | American |
| 3 | L1 German | 8 | German |
| 4 | L1 English | 1 | American |
| 5 | L1 English | 2 | German |
| 6 | L1 English | 0 | American |
| 7 | L1 German | 1 | German |
| 8 | L1 German | 0 | German |
| 9 | L1 German | 3 | German |
| 10 | L1 English | 0 | American |
|  | 8/10 = 80% |  |  |

The answers under the heading “analyst’s intuition” relate to what I thought the native tongue of the author was after my initial read through of the texts, each text was considered on its own merits and not by process of elimination. While the previous analyses will have served to improve my intuitions, there was no structured analysis undertaken to reach these answers. The number under “markers” relates to how many of the significant markers were present in the texts of each author. The nationality under author states relates to whether the stated “I am American” or “I am German” (when used to indicate their L1 as discussed in the previous Chapter). A higher number of markers indicates a higher likelihood that the author is a native German speaker.

While it is apparent that the analyst’s intuition is relatively affective at predicting whether or not an anonymous author is a native German speaker, there is no evidence to show whether or not this is more effective than a lay person’s intuition.

We noted in the previous chapter that although a L1 German speaker may sometimes be falsely identified as an L1 German speaker, the reverse did not occur at during the logistic regression analysis. However, in this analysis we can see that the markers do indicate that some authors may be native German speakers when in fact they are not. For example author 1 displays the second highest number of markers, four within the approximately 200 words, however they proclaim themselves to be American and not a native German speaker. The initial read through also indicated that the author was a native English speaker. There are also occurrences of both L1 English and L1 Germans authors, not displaying any of the discriminating markers at all within their texts. If we look at the total number of markers for the L1 German authors we can see that this is considerably higher than the total number of markers for the L1 English authors, thirteen markers to seven. However, when we take into consideration that 8 of the markers for the L1 German texts were supplied by one author alone, then the difference between the groups does not seem so significant.

## 6.3 Conclusions

The results of this analysis indicated that searching for the most significant markers in smaller anonymous texts was not a reliable method for ascertaining whether or not the author was an L1 German or L1 English speaker. In contrast to this it does indicate that an analyst’s intuition alone is a good indicator. It would be very interesting to expand this research by comparing the intuitions of linguists with backgrounds in this field, to the intuitions of lay-people.

The texts were deliberately chosen to be short texts, to represent the length more likely to require analysis in a real-life scenario. However, there was no clear distinction between the frequency of the selected markers for the two groups of authors. I would venture to attribute this lack of distinction to the brevity of the texts, as this is the only significant difference between this analysis and the analysis of the previous chapter. This research could be furthered by performing the analysis on anonymous texts of different lengths, to ascertain how the number of words correspond to the accuracy of the anonymous testing result. It would also be an interesting continuation to do a more formal statistical analysis using the same method as used in the previous two chapters for the student and internet corpora (logistic regression based testing using SPSS).

A further limitation was the number of authors. Due to time constraints it was decided to only focus on five L1 German speakers and five L1 English speakers. There is a possibility that with a greater number of authors a slight trend in the frequency of the selected markers may become visible.

Chapter 7 – Conclusions

## 7.1 Summary of Findings

This research was intended to discover whether it was possible to use interlingual identifiers to determine whether an anonymous author of a text written in English was a native German speaker or a native English speaker. The frequency chart shows that there are numerous markers which recur consistently across different L1 German speakers writing in English. This leads us to conclude that there indeed markers that can predict an author’s native language (NL). The analysis of the student texts shows that these markers vary according to the level of the learner’s English ability and that in this context it is possible to ascertain which level a student is. It is uncertain how this will transfer (to other genres, groups of authors, or texts with different topic areas), especially as the levels of the student are rather arbitrary and were primarily related to the school year of the author.

The results from the internet analysis were a primary indicator that the original hypothesis was to some extent true. There was a clear distinction in the markers between the group of L1 German and L1 English speakers. However, the blind test indicated that the markers were not good identifiers when transferred to the texts selected for the blind test. As the only significant change in the texts, from the previous analysis, was the size of the texts for each author, this indicates that the indicators are not reliable for texts with such a low word count. This does not mean that it is not possible to determine if an author is a native German or English speakers, as the two previous analyses (student and the full internet analyses) show that the interlingual markers are very reliable for determining between groups. The warning generated by SPSS that there were too many variables, even after the variables were reduced, and they may be too specific to the texts would appear not to be detrimental to the model, as the markers still hold a powerful predicting power across genres. However, this can only be a tentative observation as the markers are only analysed in two genres; student exam texts, and *Stormfront* posts.

## 7.2 Limitations and Indications

One limitation of the initial student analyses is that it is impossible to accurately separate learners into groups of level ability, as language ability is a sliding scale and dependant on many different factors. Although the grouping is appropriate in this scenario, it would be very difficult to transfer this to real life contexts. I quoted Selinker in the Literature Review as having written: *“*This set of utterances for *most* learners of a second language is not identical to the hypothesized corresponding set of utterances which would have been produced by a native speaker of the TL had he attempted to express the same meaning as the learner.” (Selinker L. , 1974, pp. 34-35) It is interesting that he italicised the word ‘most’. We can see in the results from the SPSS reduced variable analysis of the internet corpora, that some of the non-native English speakers were demonstrating a sequence of markers that are close to the occurrences of markers for the group of native English speakers. One could conclude that these three L1 German authors would be the type of learners that Selinker had in mind when he italicised the word *most*.

It was also impossible within the realms of this project to acquire detailed linguistic backgrounds for all the authors analysed. This is highlighted by the student author, beg05, who demonstrated spelling errors that appeared to be consistent with French spelling conventions. There were also German students who used American spelling conventions. In the same way that it is impossible to group learners into accurate groups, it is becoming completely difficult to separate people into groups, that have been influenced by just one language. With increasing globalisation, society is becoming more and more ‘multi-kulti’ or influenced by different languages and cultures.

At the moment this process is not widely applicable due to the lack of trained analysts, if this were to be become a widely used analysis, then it would be necessary to conduct analyst training and testing in order to increase inter-rater consistency and reliability. It may also be necessary to investigate the possibility of a more computerised analysis, through the creation of a specially designed program, though it would be impossible and irresponsible to turn this model into a fully computerised process.

While the internet texts were chosen to represent ‘real-life’ or naturally occurring texts, they were all taken from one website. This had a benefit that it reduced many of the variables (such as site conventions, or the input of the site moderators) that may affect the results and disguise or falsely indicate trends however; it also means that the results are less generalisable. It is unknown if we would get the same positive results when looking at different *chatgroups* or a different genre of internet text.

The results of the blind test showed that the markers were not immediately transferrable to such a short volume of words. However, this does not mean that it is not possible to perform authorship analysis to determine the L1 of an anonymous author. It does indicate the need for further investigation into the reliability of the test in relation to the length of text. It also demonstrated the importance of intuition in such analyses. This supports Corder’s positive comments on the native intuitions of the linguist (or teacher) (1973).

## 7.3 Implications

This research indicates some positive findings, yet also demonstrates a need for further analysis and research within this field. One area that demonstrated a need for more analysis was the motivations of spelling errors (and related markers). Some impact of German pronunciation on spelling could be seen in the student essays, further research in this area could lead to a classification of the spelling errors, which could in turn be investigated as an indicator of an author’s native language.

Before this research can be turned into a fully functioning model for wider authorship analysis it needs to be investigated further, looking at more genres, a wider range of sources, with a wider range of authors from different places, ages and social backgrounds. It would also be beneficial to widen the research to include different analysts using the same method, so that the significance of inter-analyst variation can be measured. It would also be imperative that the impact of text length is further investigated, as this has a clear and significant impact on the reliability of the analyses. It would also be a valuable study to conduct similar analyses focusing on identification of author’s with a different L1. This research shows that there are clear trends, and further research is defiantly warranted in these areas, as a working model to identify an anonymous author’s L1, is within our reach.

# Bibliography

Source Texts:

(2000-2009). *Stormfront* -*Stormfront Discussion Forum*. Retrieved June - September 2009, from Stormfront: http://www.stormfront.org/forum/

Bibliography:

Álvarez, E. G. (2004). *Interlanguage Lexcial Innovation.* Munich: Lincom GmbH.

Barron, A. (2001). *Acquisition in interlanguage pragmatics: learning how to do things with words in a study abroad context.* Amsterdam: John Benjamins Publishing Company.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus Lingusitics: Investigating Language Stucture and Use.* Cambridge: Cambridge University Press.

Brace, N., Kemp, R., & Snelgar, R. (2009). *SPSS for Psychologists 4th Edition.* Bodmin: Palgrave Macmillan.

Calda-Coulthard, C. R., & Coulthard, M. (eds) (. (1996). *Texts and Practices. Readings in Critical Discourse Analysis.* London: Routledge.

Chambers, J. (2003). *Sociolinguistic Theory. 2nd Edition.* Oxford: Blackwell Publishing.

Chambers, T. S.-E. (2002). *The Handbook of Language Variation and Change.* Oxford: Blackwell Publishers Ltd.

Chaski, C. (2001). Empirical Evaluations of Language-based Author Identification Techniques. *Forensic Linguistics: The International Journal of Speech Language and the Law 8(1)* , 1-65.

Conley, J., & O'Barr, W. (2005). *Just Words: Law, Language, and Power (2nd ed.).* Chicago/London: University of Chicago Press.

Corder, P. (1973). The Elicitation of Interlanguage. In J. Svartvik (ed.), *Errate: Papers in Error Analysis* (pp. 35-54). Stocholm: Rotbeckman.

Cotterill, J. (2007). *Language in the Legal Process.* Eastbourne: Palgrave Macmillan.

Coulthard, M. (2000). Whose Text is it? On the Linguistic Investigation of Authorship. In S. Sarangi, & M. Coulthard, *Discourse and Social Life* (pp. 271-87). London: Pearson.

Crystal, D. (2006). *Language and the Internet (2nd ed.).* Cambridge: Cambridge University Press.

Ebert, K. (1982). The Definite Articles with Inalienables in English and German. In W. Lohnes, & E. (. Hopkins, *The Contrastive Grammar of English and German* (pp. 64-75). Michigan: Karoma Publishers, Inc.

Faerch, C., & Kasper, G. (1983). *Strategies in Interlanguage Communication.* Harlow: Longman Group Limited.

Field, A. (2009). *Discovering Statistics Using SPSS 3rd Edition.* London: SAGE Publications Ltd.

Foster, D. (2001). *Author Unknown: On the Trail of Anonymous, 2nd ed.* London: Macmillan.

Gee, J. P. (1999). *An Introduction to Discourse Analysis: Theory and Method.* Oxon: Routledge.

Gibbons, J. (2003). *Forensic Linguistics: An Introduction to Language in the Justice System.* Oxford: Blackwell Publishing.

Gibbons, J. (1994). *Language and the Law.* Harlow: Longman Group UK Limited.

Grant, T. (2008). Approaching questions in forensic authorship analysis. In J. Gibbons, & M. T. Turell (eds), *Dimensions in Applied Linguistics* (pp. 215-229). Amsterdam & Philadelphia: John Benjamins Publishing Company.

Grant, T. (2003). Review of McMenamin, G. 'Forensic Linguistics: Advances in Forensic Stylistics. *Forensic Lingustics 10.1* , 154-156.

Heine, B., & Kuteva, T. (2005). *Language Contact and Grammatical Change.* Cambridge: Cambridge University Press.

Hopkins, E. (1982). Contrastive Analysis, Interlanguage, and the Learner. In W. Lohnes, & E. Hopkins, *The Contrastive Grammar of English and German* (pp. 32-48). Michigan: Karoma Publishers Inc.

Kasper, G., & Blum-Kulka, S. (1993). *Interlanguage Pragmatics.* Oxford: Oxford University Press US .

König, E. (1982). Scalar Particles in German and Thier English Equivalents. In W. Lohnes, & E. Hopkins, *The Contrastive Grammar of English and German* (pp. 76-101). Michigan: Karoma Publishers, Inc.

König, E., & van der Auwera, J. (ed.s) (1994). *The Germanic Languages.* London: Routledge.

Labov, W. (1984). Field Methods of the Projecct in Lingusitic Change and Variation. In J. Baugh, & J. (. Sherzer, *Language in Use* (pp. 28-53). Engelwood-cliffs: Prentice-Hall.

Labov, W. (2001). *Principles of Lingusitic Change: Social Factors. Volume 2. .* Oxford: Blackwell Publishers Ltd.

Labov, W. (1972). *Sociolinguistic Patterns.* Pennsylvania : University Pennsylvania Press, Inc.

Lado, R. (1957). *Linguistics Across Cultures.* Ann Arbor: University of Michigan Press.

Lee McKay, S., & Hornberger, N. (1996). *Sociolingusitics and Language Teaching.* Cambridge: Cambridge University Press.

Lucy, D. (2005). *Introduction to Statistics for Forensic Scientists.* Chichester: Wiley.

Macaro, E. (2002). *Learning Strategies in Foreign and Second Language Classrooms: The Role of Learner Strategies.* London: Continuum International Publishing.

Mackay, S. L., & Hornberger, N. (. (1996). *Sociolinguistics and Language Teaching.* Cambridge: Cambridge University Press.

McArthur, T. (2003). *Oxford Guide to World English.* Oxford: Oxford University Press.

McMenamin (Ed.), G. (2002). *Forensic Linguistics: Advances in Forensic Stylistics.* Boca Raton, Florida: CRC Press.

McMenamin, G. (2001). Style markers in Authorship Studies. *Forensic Lingusitics 8(2)* , 93-97.

Nemser, W. (1974 [article first published 1967]). Approximative Systems of Foreign Language Learners. In J. (. Richards, *Error Analysis. Perspectives on Second Language Acquisition* (pp. 55-63). London: Longman Group Limited.

News, B. (2009, September 27). *'I do not roll over,' says Brown .* Retrieved September 27, 2009, from BBC News: http://news.bbc.co.uk/1/hi/uk\_politics/8277131.stm

Olsson, J. (2004). *Forensic Linguistics: An Introduction to Language, Crime and the Law.* London & New York: Continuum.

Pinker, S. (1994). *The Language Instinct.* Oxford, Massachusetts: Harvard University Press.

Pinna, A. (2007). Exploiting LSP Corpora int eh Study of Foreign Languages. In D. Gálová, *Languages for Specific Purposes: Seaching for Common Purposes* (pp. 146 - 162). Newcastle: Cambridge Scholars Publishing.

Richards, J. (25). A Non-contrastive Approach to Error Analysis. *English Language Teaching 25* , 204-219.

Richards, K. (2003). *Qualitative Inquiry in TESOL.* Basingstoke: Palgrave Macmillan.

Sankoff, G. (2002). Lingusitic Outcomes of Language Contact. In T. S.-E. Chambers, *The Handbook of Language Variation and Change* (pp. 638-668). Oxford: Blackwell Publishers.

Selinker, L. (1974). Interlanguage. In R. (. Jack, *Error Analysis: Perspectives on Second Language Acquisition* (pp. 31-54). London: Longman Group Limited.

Selinker, l. (1992). *Rediscovering Interlangugae.* Harlow: Longman Group UK Limited.

Solan, L., & Tiersma, P. (2005). *Speaking of Crime: The Langugae of Criminal Justice.* Chicago and London: The University of Chicago Press.

Swales, J. (1990). *Genre Analysis: English in Academic and Research Settings.* Cambridge: Cambridge University Press.

Tarone, E. (1983). On the Variability of Interlanguage Systems\*. *Applied Linguistice 4(2)* , 142-164.

Tarone, E. (1988). *Variation in Interlanguage.* London: Edward Arnold.

Tarone, E., Cohen, A., & Dumas, G. (1983). A Closer look at some Interlangugae Terminology: A Framework for Communication Strategies. In C. Faerch, & G. Kasper, *Strategies in Interlanguage Communication* (pp. 4-14). Harlow, Essex: Longman Group Limited.

Thompson, G. (2004). *Introducing Functional Grammar (2nd ed.).* London: Arnold, Hodder Headline Group.

Weisburd, D., & Britt, C. (2007). *Statistics in Criminal Justice (3rd ed.).* New York: Springer Science+Business Media, LLC.

Willis, D. (2003). *Rules, Patterns and Words. Grammar and Lexis in English Language Teaching.* Cambridge: Cambridge University Press.

Winford, D. (2003). *An Introduction to Contact Linguistics.* Oxford: Blackwell Publishing.

Wodak, R., & Meyer, M. (2001). *Methods of Critical Discourse Analysis.* London: SAGE Publications.

Wode, H. (1981). *Learning a Second Language. I. An Integrated View of Language Acquisition.* Tübingen: Gunter Narr Verlag.

Zipf, G. (1932). *Selected Studies of the Principle of Relative Frequency in Language.* Cambridge, Massachusetts: Harvard University Press.

# Appendices.