Candidate Knowledge? Exploring epistemic claims in scientific writing: A corpus-driven approach.

Garry Plappert

Abstract

In this article I argue that the study of the linguistic aspects of epistemology has become unhelpfully focused on the corpus-based study of hedging and that a corpus-driven approach can help to improve upon this. Through focusing on a corpus of texts from one discourse community (that of genetics) and identifying frequent tri-lexical clusters containing highly frequent lexical items identified as keywords, I undertake an inductive analysis identifying patterns of epistemic significance. Several of these patterns are shown to be hedging devices and the whole corpus frequencies of the most salient of these, *candidate* and *putative*, are then compared to the whole corpus frequencies for comparable wordforms and clusters of epistemic significance. Finally I interviewed a 'friendly geneticist' in order to check my interpretation of some of the terms used and to get an expert interpretation of the overall findings. In summary I argue that the highly unexpected patterns of hedging found in genetics demonstrate the value of adopting a corpus-driven approach and constitute an advance in our current understanding of how to approach the relationship between language and epistemology.

Keywords: epistemology, hedging, corpus-driven, concordance, collocation, genetics

1. Introduction

Whilst the identification of hedging devices has proven to be a very useful and successful enterprise within applied linguistics, it has been argued that the study of these devices has become concentrated onto a small group of the 'usual suspects' (Groom, 2007; 2010; Plappert, 2012) of words and structures that are known to have an epistemic effect in a claim or proposition. As such linguistic markers of modality such as modal verbs (eg: *may*, *might*, *can*, *could*), modal adjectives (eg: *possibly*, *probably*) and n-grams identified as functioning as *hedges* (such as *it is possible that* and *it is likely that*) often form the starting place for analysis of the linguistic aspects of epistemology. This impasse has been compounded by a plethora of corpus-based studies (eg: Hunston, 1995; Noguchi *et al*., 1996; Thompson and Ye, 1996; Williams, 1996; Chi-Hua, 1999 and cf. Hyland 1998), which, whilst providing excellent empirically-based descriptions of known epistemic structures, are unlikely to contribute to the discovery of additional or unknown epistemic devices. In this paper I will argue, in agreement with Groom (2007; 2010) that the answer to this impasse is to explore corpus-driven methods of analysis in order to uncover new or unexpected epistemic devices in English. Through a close analysis of four clusters, I demonstrate that it is possible to discover a number of additional strategies for nuancing claims, which are not typically mentioned in seemingly exhaustive studies such as Hyland (1998). I also argue that the peripheral presence of the ‘usual suspects’ in the cotext of nodes such as *tumor suppressor gene*, *mutations in the gene encoding* and *loss-of-function mutations* raises the possibility that the epistemic devices of which we are already aware may be far more marginal phenomena than we currently assume. Whilst many researchers have sought to study hedging devices in academic writing, few if any have asked whether hedging is as central and as ubiquitous as is assumed and still further few (none, to my knowledge) have attempted to discover whether academics typically use the devices we assume that they do when forming claims. Indeed, we might even be accused of a certain naivety in taking such a focus in Applied Linguistics. If non-hedged claims or subtler forms of hedging are what are actually typical in a discipline we will not find this out by starting with known hedging devices and this is why I argue that a corpus-driven approach can still be a useful and coherent one.

In what follows I discuss the relevant literature related to the study of hedging in academic writing (section 2) and argue that a corpus-driven approach would provide a useful supplement to the plethora of corpus-based studies that have been carried out. I then discuss the corpus chosen for this study and the process by which this was analysed (section 3) before presenting the most salient results (section 4). Finally in section 5 I discuss the ramifications of the study and the limitations inherent in the approach taken.

2. Literature Review

2.1 Hedging and the linguistics of epistemology in Applied Linguistics

The study of epistemology within Applied Linguistics has focused on the linguistic devices used to mitigate claims (cf. Hyland 1998), though the term used for this phenomenon has varied considerably. Thus Hyland (1998) is able to identify studies of hedges (Lakoff, 1972) as well as '*compromisers* (James, 1983), *downtoners* (Quirk et al, 1972), *weakeners* (Brown & Levinson, 1987), *downgraders* (House & Kasper, 1981), *softeners* (Crystal & Davy, 1975), *backgrounding terms* (Low, 1996) and *pragmatic devices* (Stubbe & Holmes, 1995)' (1998:9, my italics) as constituting what he wishes to call *hedging*. This subject, then, has undoubtedly received plentiful coverage in Applied Linguistics and work focused on identifying or analyzing hedging in academic discourse has become so common that Groom (2007) has identified (rather despairingly) the 'usual suspects' of corpus study on this subject:

'A glance at the recent literature identifies report clauses and other attributive forms […] modal verbs and other hedging devices […] and extraposed complement clauses and other kinds of that- clause […] as being amongst the usual suspects' 2007: 40)

Implicit in this list is the study of semi-fixed phrases known variously as lexical bundles (eg. Biber, 2009, Cortes, 2004), fixed collocation patterns (Oakey, 2008) and structures named partly according to which piece of software was used to identify them, such as *clusters* in WordSmith Tools (Scott, 2004) and *c-grams* in W—Matrix (Rayson, 2009). Although the exact definition of these structures varies from linguist to linguist and from software to software, what is consistent is that in each case the linguistic form of the item to be studied is preselected. The advantage of this approach for the large scale analysis of written academic discourse is that the seemingly exhaustive lists of hedging devices provided by works such as Hyland (1998) and (2009) provide a clear and labour-saving basis for selecting and analyzing items from wordlists, allowing the analyst to proceed with collocation or concordance line based description. Studies such as these can confirm and deepen our understanding of the functioning of a specific item of hedging and the typical structure and findings of these will be discussed below. However, such studies of known hedging devices are by their very nature unlikely to widen or extend the very list from which they are chosen: the list of known hedging devices. If we wish to take seriously the challenge of extending this list of 'usual suspects', or at least investigating whether there are any forms of hedging not covered by this list, it would be helpful to take an approach that makes no assumptions at the outset as to which features are epistemologically key.

In what follows I will therefore argue that we can expand our understanding of hedging in academic writing through a corpus-driven approach. The terms *corpus-based* and *corpus-driven* (Tognini-Bonelli, 2001) have received a great deal of discussion within corpus linguistics (cf. Biber, 2009), and this is a useful distinction particularly when what is being considered is how the research aims of any given study are supposed to be being met by the use of corpora. In her original formulation Tognini-Bonelli defines the corpus-based approach as being one that uses corpora ‘as a repository of examples to expound, test or exemplify given theoretical statements’ (2001:10). Each of the studies of hedging mentioned above therefore constitutes a ‘corpus-based’ study (and often explicitly so) since the linguistic item to be studied is pre-selected before any actual analysis begins. The use of the corpus is then often limited to counting and investigating tokens of this item with a view to understanding how it functions in the particular genre being studied. When seeking to meet the research aims of the corpus-based study of known hedging devices this is not a problem, since the aim of the study is simply to extend our knowledge of a certain device, not hedging devices *in toto*. However, when the aim is to assess or extend our current list of ‘usual suspects’ and our model of how these apply, this approach will clearly not help, and this is where the corpus-driven approach becomes useful. This is because it is an *a posteriori* approach, where ‘a theoretical statement can only be formulated in the presence of corpus evidence and is fully accountable to it’ (Tognini-Bonelli, 2001: 11). Thus whilst the corpus-based approach builds on our current theoretical understanding of hedging, the corpus-based approach can challenge, or at the very least test whether the hedges discovered in a given discourse community or genre are consistent with our list of ‘usual suspects’. Moreover, if they are not, it can be used to add to that list and even replace it. Whilst the coherence of the corpus-based/corpus-driven distinction has been criticized (cf. McEnery and Hardie 2012) I argue here that it is still a useful way of drawing our attention to the relationship between theory and method in corpus linguistics. I also contend that whilst previous corpus-driven approaches such as that of Biber (2009) have been highly successful in establishing the frequency and distribution of (in that case) formulaic language they still involve some pre-selection of the linguistic device (a multi-word unit) and that this type of preselection may not allow us to see the full picture of how claims are typically made in a discipline. If an approach like this is taken in order to study hedging there is a clear assumption that the linguistic features functioning as hedges will appear as frequent n-grams and I will argue that it can be worthwhile to take an even more radically corpus-driven approach to hedging than this. As such this study takes as its starting point not highly frequent items of hedging identified through corpus methods (as a corpus-driven study might typically be described as doing) but rather starts with items of terminology. In doing so it asks the following questions:

1- what types of epistemic claim are highly frequent items typically involved in making?

2- Are these claims typically hedged or not?

3- If they are, do we find the ‘usual suspects’ of hedging or not?

It is radically corpus-driven in that it starts not with the assumption that we know what we want to look for (e.g., hedges) but rather with the intention of creating an inductive description of the typical claims in a discipline from the data. Through this method I will try to demonstrate that it is possible to discover previously unidentified hedging devices, before finally concluding that the ‘usual suspects’ of study such as modal adjectives and phraseological chunks such as *it is probable that* may have a more peripheral role in the hedging in scientific writing, and particularly in the discipline of genetics, than might currently be assumed.

Perhaps the key text amongst corpus-based approaches to hedging is Hyland (1998), which reports on an investigation of a small corpus of 26 research articles ‘in the field of cell and molecular biology’ (p.96). Whilst corpus size and scope has increased considerably since this work was carried out it remains an influential text in describing hedging in scientific research articles. Hyland's approach is of course focused on a relatively small number of texts but results in an impressive if somewhat predictable list of devices that can be used to hedge a scientific claim. These include modal auxiliaries, what he calls 'epistemic lexical verbs' such as *suggests* and *indicates*, and 'epistemic adjectives, adverbs and nouns' with wordforms such as *essentially*, *relatively*, *generally*, *most*, *slightly* and *presumably* and their various frequencies being presented and compared via normalized frequency to their occurrences in the *JDEST* corpus (a corpus of 2,000 texts of approximately 500 words each totaling around 1 million words and comprising English texts from ten scientific disciplines) and the more familiar *Brown/LOB* corpus. Hyland also provides a discussion of the hedging of numerical data and what he terms 'non-lexical hedges'. This latter category is of particular interest in that whilst he presents these as fairly abstract 'strategies' (the frequency of which he also attempts to judge), the actual linguistic details of these are far from obvious or predictable and include phrases such as 'one cannot exclude the possibility that', 'cannot presently be ruled out' and the perhaps more predictable 'it is not known whether'. Hyland sub-categorises these strategies as 'reference to limiting experimental conditions', 'reference to a model, theory or methodology' and 'admission to a lack of knowledge' and provides plentiful examples of these from a corpus of just 26 research articles. In his more recent work Hyland has used corpus methods to contribute to work in disciplinary discourse (e.g., Hyland, 2004; Hyland and Tse 2008) but it is perhaps this earlier work on hedging in scientific articles that has most influential on the study of hedging in scientific writing, in particular, and academic writing more generally.

We can consider Cortes' (2004) pedagogically motivated study into student writing as a typical example of such approaches. The overall aim of the study is disciplinary comparison, with the pedagogic motivation coming from an EFL perspective where students writing in a language other than their first language are being taught academic writing. By pre-selecting lexical bundles for study Cortes assumes that the construction of 'target bundles' (which are derived from professional writing in the relevant fields, as represented by published research articles) is what is needed for improved student writing. The study then proceeds from the identification of these bundles using automatic corpus methods. While this is a sensible approach, it is not one that is interested in extending our list of hedging devices but rather seeks to exploit what we already know for a particular pedagogic application.

Studies of this type are plentiful and have extended well beyond the identification and description of hedges to a range of formally identifiable wordforms that are judged to be rhetorically significant in scientific writing. The study of, for instance, personal pronouns (Noguchi *et al.*, 1996); Chi-Hua, 1999), verbs (Hunston, 1995; Thompson and Ye, 1991; Williams, 1996) and indeed almost all of the forms that Hyland (1998) has previously identified have been executed successfully, exploiting the benefit of being able to identify the objects of study quickly and to provide a thorough empirical examining of their functioning.

However, if we accept Sinclair's view (1991) that the most important potential gain from a corpus approach is the enhanced ability to discover facts about language that are not immediately obvious or even available to or achievable through intuition, it seems unwise to focus only upon the capacity of corpus methods to measure and deepen our understanding of linguistic items of which we are already well aware.

2.2 The linguistic study of genetics

There has been considerable interest in both the public and professional discourses surrounding genetics. The essentialist nature discourses surrounding genetics (cf. Nelkin and Lindee 1995), the inherently hierarchical nature of a focus on the proper functioning (or otherwise) of our genes and the potential for genes and other hereditary aspects to become the focus of discrimination (Hubbard and Wald 1993) have all been raised as concerns that have been inadequately addressed as the way that we encode our understanding of genetics into language has struggled to keep pace with scientific advances. Most worryingly, previous research has repeatedly argued that knowledge about genetics is often expressed within a *deterministic frame* (Hubbard and Wald 1993; Nelkin and Lindee 1995) and that this constitutes a misrepresentation of the nature of genetic causation that may lead to the adoption of a fatalistic attitude to personal health amongst the general public (cf. Shen and Condit (2012) for a discussion of this issue). Indeed, even research focused on an audience of 'undergraduate students at a major southern research university' in the United States and seeking to conclude that the deterministic view of genetics is *not* the predominate one found that 39 out of 137 participants (28.8%) 'expressed a deterministic conception of genetics' (Condit 1999). This research has led to repeated identification of the deterministic frame in the mass media (cf. Carver et al. 2008) and despite repeated identification of this supposed miscommunication of causation in genetics this set of meanings appears to be pervasive and persistent. Carver et al. (2008) helpful identify what they regard as the 'key words and phrases' that constitute the deterministic frame within media discourses surrounding genes and genetics and they list these as being *gene for*, *cause*, *control*, *culprit*, *disease gene*, *responsible for*, *wired in* and *born with*. This list however was arrived at somewhat intuitively since they were identified from a small corpus of news media texts as constituting a less scientific encoding of the causal processes (etiology) of genetics. Whilst previous studies have enlisted a geneticist to comment upon and interpret media discourses, few if any have analysed the professional discourse of genetics in order to provide a sound empirical basis for any claims as to what does or doesn't constitute 'scientific' discourse on genetics.

In order to address this I attempt to show in what follows how a method focused on entities rather than on hedging devices can allows us to identify the typical patterns of claims in a discipline, as well as the preferred methods of nuancing those claims. I also follow Groom (2007; 2010) and Plappert (2012) in arguing that it is possible to extend our list of known hedging devices through the corpus-driven approach.

3. Methodology

3.1 Corpus Compilation

The leading journal in the field of genetics is *Nature Genetics* (29.648 Thomson Reuters 2014, accessed February 2014) and it was decided that the corpus for this study would be comprised of texts from this journal. This choice was made in order to remove the potential variables introduced by including a range of journals, whether through 'house style', editorial idiosyncrasies or even the inclusion of multiple and conflicting paradigms. The texts for this study came from a ten-year period (1999-2008 inclusive) and were collected together in a corpus that I have called *genecorp*. In order to be maximally representative of this, the most prestigious work in the field, *genecorp* contains 2,979 texts from the journal *Nature Genetics*, spanning from 1999-2008. These texts are labeled by the editors of *Nature Genetics* as being of nine different text types including *articles, brief communications*, *letters* and *news and views*. Despite this apparent range of text types it was decided that all of the texts found would be used in *genecorp*. Indeed, it is consistent with the corpus-driven approach to avoid any *a priori* assumptions about the distribution of linguistic items according to text type or genre.

In order to individuate the texts each text file was labeled with part of the name of the first named author (since there are typically multiple authors in scientific writing). In addition to this each text was also labeled with the year it was published, so that the filename has in effect three parts, as shown by these examples: 'LAN01_A', 'RAG04_L', and 'NUS06_L'. This format has the additional benefit of allowing the analyst to see both the genre of the text and the year the text was published whilst looking at concordance lines. In addition to this, the corpus was also organised into nine separate folders, corresponding to the nine different text types, giving the analyst the opportunity to quickly isolate all of the texts of a particular type for any subsequent work focused on genre. The files within each folder were then also subdivided into folders based on the calendar year that they were published, again allowing for the automatic selection of texts on the basis of publication date rather than genre.

3.2 Analysis of *genecorp*

In order to carry out a 'bottom-up' analysis of claims made in *genecorp* the following procedure was adopted:

1. Generation of keywords using *BNC World* as reference corpus
2. Generation of clusters containing the ten most key keywords
3. Selection of all clusters containing three lexical items from (2)
4. Collocation analysis of tri-lexical clusters from (3)
5. Concordance line analysis of tri-lexical clusters from (3)
6. Form generalisations about geneticists epistemic practices based on the evidence of (4) and (5)
7. Inspect whole corpus frequencies where possible to check the plausibility of (6)

The use of keywords analysis to extract items for further study in an entirely bottom-up way has proven to be highly successful in corpus-based approaches to discourse analysis (cf. Gledhill, 1995; 1996; Tribble, 2000; Scott, 2000; Baker, 2006; Gabrielatos and Baker, 2008). Using keywords also has the advantage of removing researcher bias, with the concomitant advantage of making the choice replicable, as it is not reliant upon the choices or intuitions of the analyst. In this case, it also yielded words that are of very high frequency, as is demonstrated in the following table:

Keyword	Raw frequency
cells	35,961
gene	29,058
genes	28, 230
mice	24, 532
expression	28, 409
cell	22, 381
DNA	19,999
protein	17, 732
mutations	14, 895
genome	12, 959

Figure 1: Raw frequencies for the ten most key keywords from *genecorp* using the BNC as reference corpus

The high frequency of these words suggests that they are likely to be present across the whole corpus, allowing for detailed investigation of the salient patterns in which any node item occurs. Where a list of keywords contains much lower frequency items it might not be possible to carry out inductive analysis (since a certain minimum amount will of course be required before patterns can be identified). Most importantly, high frequency implies good coverage across the corpus, providing a stronger basis on which to identify patterns and draw conclusions. Nevertheless, whilst whole corpus keywords can provide a number of very high frequency lexical items that are likely to be central to the discipline of genetics, the problem of 'too much data' (Hunston, 2002) remains at this stage. Frequencies of between twelve and thirty thousand for the ten keywords identified in figure 1 (above) clearly constitute an excess of what the analyst can realistically deal with when seeking to take a fine grained 'bottom-up' approach and therefore what was needed at this stage was to refine this further. The second step, then, that was taken was to select clusters containing each of the ten keywords in order to focus the analysis on smaller subsets of frequently occurring strings containing the keywords (2). Since this approach requires terminological items that would be constitutive of common means of forming propositions in genetics, a third step was taken to select from these clusters those containing three lexical items (3). This was done because previous work has identified lexical chains as being of particular use in identifying terminological items (e.g., Rogers, 2007:17) and high frequency lexical items would appear to be a plausible starting point for the discovery of terminology for further investigation.[1] The result of this was the following list of the most frequent clusters that contain at least three lexical items at least one of which is a keyword from *genecorp*:

genecorp keywords	Tri-lexical clusters
cells	wild type cells, embryonic stem cells, cos 7 cells, bone marrow cells, stem es cells, embryonic stem es cells, cd8 t cells, cos 1 cells
gene	**gene expression data**, **gene expression patterns**, **mutations in the gene encoding**, **gene expression profiles**, tumor suppressor gene, **changes in gene expression**, **analysis of gene expression**, **variation in gene expression**, **gene expression profiling**

[1] Credit is also due to Paul Rayson who suggested to me in conversation that focusing on frequent multi-word units containing several lexical items would enable me to identify terminological items.

genes	tumor suppressor genes, X linked genes, **protein coding genes**, differentially expressed genes
expression	**gene expression data**, **gene expression patterns**, **mutations in the gene encoding**, **gene expression profiles**, **changes in gene expression**, **analysis of gene expression**, **variation in gene expression**, **gene expression profiling**
cell	cancer cell lines, lymphoblastoid cell lines, es cell lines, cell cycle arrest, es cell clones, cell cycle progression, whole cell extracts, cancer cell line, planar cell polarity, breast cancer cell
analysis	western blot analysis, northern blot analysis, southern blot analysis, RT PCR analysis, RTA PCR analysis, blot analysis using, RNA blot analysis, **analysis of gene expression**
DNA	DNA binding domain, DNA copy number
protein	green fluorescent protein, protein protein interactions, **protein blot analysis**, fluorescent protein GFP, green fluorescent protein GFP, protein protein interaction, wild type protein
mutations	**mutations in the gene encoding**, loss of function mutations, disease causing mutations
genome	genome wide association, wide association study, wide association studies, genome wide linkage, genome wide significance, human genome project, human genome research, human genome sequence, the human genome project

Figure 2: tri-lexical clusters containing the ten most key keywords from *genecorp;* clusters that contain more than one of the ten most key keywords are in bold

Since each of the key tri-lexical clusters occurred no more than roughly two hundred times it was felt that there was a manageable amount for detailed concordance line analysis. However, it is worth noting that Hunston (2002:52) advises that 100 concordance lines is roughly what an analyst can cope with when attempting to identify *general* patterns, with 30 lines given as around the limit for *detailed* patterns. In order to avoid being overwhelmed by the level of detail around the node phrases two methods were used. Collocation data for the node cluster was generated (step 4) using collocates generated using raw frequency in order to be as consistent as possible with the corpus-driven approach. This has the benefit of providing objective data as to the most frequent patterns around the node and providing an element of triangulation for the concordance line analysis (step 5), which, taking up the suggestion of Hunston (2002:52) and following Sinclair (2003), was based on sets of 30 random lines with patterns being identified until no further patterns appeared. The attempt was then made to form generalisations about the claims which geneticists typically make and about the linguistic processes used to nuance these claims (step 6), with these then being checked against whole-corpus frequencies in order to assess their plausibility wherever possible (step 7).

4. RESULTS

The results of this analysis can be divided into two broad categories: routine reports of methods which do not require hedging of any sort (these will be discussed in section 4.1) and what I call 'epistemic nodes'- clusters which are found because they have become a typical way of encoding a claim which may exhibit hedging, and which will be analysed in detail in sections 4.2 and 4.3.

4.1 'Normal Science'- the unhedged reporting of routine methods

The majority of the clusters analysed were involved in the reporting of methods and routine findings that required no hedging whatsoever. In an apparent demonstration of what Kuhn (1970) calls 'Normal Science', the propositions containing these clusters do not require hedging because they merely report the (definite) results of established methods (in the case of example 1, below) or describe those methods, as in example 2:

1. Plk4+/- embryonic fibroblasts had increased centrosomal amplification, multipolar spindle formation and aneuploidy *compared* with *wild-type cells*. (ros05_l)

2. We generated Cdc25b-deficient mice by homologous recombination in *embryonic stem cells* (Fig. 1a,b)13. (lin02_l)

Routine statements such as these require no hedging and yet are clearly a central part of the process of scientific writing. Of the 63 clusters examined 59 were of this type, revealing the most common unhedged claims present in the discourse; they represent statements that build up to and support wider and more contentious claims. The remaining clusters, which I call epistemic nodes, do exhibit hedging and despite being comparatively rare are crucial in revealing the typical claim patterns made in the discipline. These were found to fall into two principal patterns, the use of left-side modifiers such as *candidate* and *putative* and the variation of verb phrase choices, from those involved in unhedged claims such as *X causes Y* to those involved in more nuanced claims such as *X is associated with Y*. The variation surrounding these epistemic nodes will be presented throughout the remainder of the results section.

4.2 *Putative* and *Candidate* as hedging devices: *tumor suppressor gene*

The principal finding surrounding the node *tumor suppressor gene* is that it is usually hedged with *candidate* and *putative*, as demonstrated by the following collocation data:

word	no.	L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4	R5
the	88	11	6	6	27	19	0	1	6	4	7	1
of	86	10	8	17	23	8	0	0	0	6	9	5
a	85	3	3	5	27	39	0	0	5	1	1	1
in	61	1	3	4	3	1	0	18	13	2	7	9
is	45	1	2	7	7	0	0	6	12	4	2	4

that	34	12	5	2	1	0	0	9	2	0	2	1
and	28	2	1	3	1	1	0	5	7	3	3	2
as	26	2	3	10	8	1	0	2	0	0	0	0
to	26	4	2	2	1	0	0	1	2	8	3	3
putative	16	0	0	1	0	15	0	0	0	0	0	0
for	15	2	0	3	1	2	0	3	3	0	0	1
cancer	13	0	0	1	0	1	0	0	1	5	2	3
candidate	11	0	1	0	1	9	0	0	0	0	0	0
which	11	1	2	0	0	0	0	2	4	0	1	1
with	10	1	2	0	2	0	0	1	0	1	1	2
mutations	10	1	6	2	0	0	0	0	1	0	0	0
IDB4	9	1	5	1	0	0	0	2	0	0	0	0
1	9	1	0	2	1	1	0	1	0	1	1	1
inactivation	8	0	1	6	0	0	0	0	1	0	0	0
function	8	0	1	0	0	1	0	5	0	0	0	1

Figure 3: The twenty most frequent collocates of *tumor suppressor gene* in *genecorp*

The list of the twenty most frequent collocates of *tumor suppressor gene* is striking in containing a number of lexical items that may have an epistemic function, most notably *candidate* and *putative* which would appear to both mark possibility within the span of the node. Examination of the concordance lines made this relationship much clearer and indeed identified a range of epistemic strategies; some of which could be formally identified (including established hedging devices such as *known* and *may*) and others which were instantiated by semantic sequences or other structures that would be more difficult to spot in either wordlists or lists of multi-word units. The principal patterns of epistemic significance within the expanded contexts of *tumor suppressor gene* were as follows.

4.2.1 Named *tumor suppressor gene*

The most common feature of epistemic significance found in the concordance data surrounding *tumor suppressor gene* was striking in being not a hedging device but rather an unhedged claim: realized as a nominalization with the name of the gene being referred to, and hence being labeled as a *tumor suppressor gene*, as can be seen in the following examples:

> **3**. Inactivation of **the *tumor-suppressor gene*** PTEN and lack of p27KIP1 expression have been detected in most advanced prostate cancers1, 2. (cri01_l)

> **4**. This cell line also lacks the von Hippela Lindau (VHL) ***tumor suppressor gene*** (der01_pro)

> **5**. The protein RB1CC1 (retinoblastoma 1 (RB1)-inducible coiled-coil 1) has been identified as a key regulator of the **tumor-suppressor gene** RB1 (ref. 1). (cha02_l)

> **6**. Mutations in the TP53 **tumor-suppressor gene** are found in 70-80% of BRCA1-mutated breast cancer but only 30% of those with wildtype BRCA1 (ref. 3). (har02_l)

What is epistemically significant in each of these cases is that the status of the gene as a *tumor suppressor gene* is apparently already established and therefore does not require any form of hedging. Whilst it is probably significant that it is deemed necessary to mention in each case that the gene is a *tumor suppressor gene*, the remainder of the proposition functions to construct new knowledge in each sentence, as something in addition to this. Thus in example 3 above it is presented as a given that *PTEN* can be accorded the status of *tumor suppressor gene* '**the *tumor-suppressor gene* PTEN'** and the new knowledge being presented is that 'inactivation of this' as well as 'lack of p27KIP1 expression' has 'been detected in most advanced prostate cancers'.

4.2.2 *putative tumor suppressor gene*

As was seen in the collocation data (figure 3) the most frequent lexical collocate of *tumor suppressor gene* is *putative*, which occurred 16 times in the 5:5 span of the node *tumor supressor gene*, and in nine different texts. The use of *putative* as a hedging device can be seen in examples such as the following:

7. Results of transfection studies in experimental animal systems <u>support the idea</u> that Idb4 is a ***putative tumor-suppressor gene*** in hematologic malignancies (liu05_a)

8. Global assessment of promoter methylation in a mouse model of cancer <u>identifies</u> ID4 as a ***putative tumor-suppressor gene*** in human leukemia (liu05_a)

It seems clear that the adjective *putative* is acting as an epistemic marker here, expressing possibility. Indeed, *putative* appears in one example in Hyland (1996) though it is not included in lists of epistemic markers or hedging devices such as those provided by Hyland (1998) and (2009). As such *putative* can be seen as a hedging device instantiating the lexical expression of modality in *genecorp*, and one that can be added to our list of the 'usual suspects'. In addition to this, further evidence of epistemic signaling is present in the surrounding context of *putative tumor suppressor gene*, as in the following example:

9. Evidence for Idb4 as a ***putative tumor-suppressor gene*** in the pathogenesis of cancer, such as shown here for both murine and human leukemia, has, to our knowledge, not been previously reported. (liu05_a)

10. We used this system to identify a new ***putative tumor-suppressor gene***, Idb4 (liu05_a)

In a number of these examples rhetorical devices can also be observed in the cotext of *tumor suppressor gene* and examples such as 9 above where it is stressed that the finding 'has not previously reported' whilst in example 10 (above), the authors stress that they have identified a '***new*** *putative tumor suppressor gene*', rather than merely stating that they have identified a *putative suppressor gene*. There are also explicit examples of the expression of uncertainty in examples 11 and 12 (below), which express the propositions that the role of the relevant *tumor suppressor gene* is '*uncertain*' and even that a *putative tumor suppressor gene* cannot yet be found:

11. the role of the ***putative tumor-suppressor gene*** H19 is uncertain 3,4 (spa04_bc)

12. Although 17p deletions occur in 50% of cases, a ***putative tumor suppressor gene*** remains unidentified7 (mac01_a)

Each of these examples surrounding the initial cluster *tumor-suppressor gene* exemplifies a tendency of epistemic talk to cluster around contested epistemic nodes. Whilst for the majority of the clusters there is no range of epistemic devices (because they are concerned with routine statements of method or findings) the four nodes discussed in detail here illustrate the ranges of possible claims in the discipline. Uncertainty surrounding the causal role of *tumor suppressor gene* is initially signaled through the adjective *putative*, and the writers also go on to express this further in explicitly stating that it is new, unknown or that its role is uncertain. Example 10 above is also noteworthy in that it points to an epistemic stage prior to the identification of a *putative tumor suppressor gene* where some conditions are fulfilled (17p deletions occur in 50% of cases) and yet this is not enough to warrant the identification of a *tumor suppressor gene*. Finally the adjective *uncertain* is of interest in this context, being a clear epistemic marker.

4.2.3 *candidate tumor suppressor gene*

When *candidate* appears as a collocate of *gene* it indicates the possibility of a particular named *gene* being a *tumor suppressor gene*. The form *candidate tumor suppressor gene* occurs nine different times and in seven different texts and interestingly the concordance data suggests that a *candidate tumor-suppressor gene* can be both a starting hypothesis for a piece of research and the conclusion of that research, as in the following examples:

13. we hypothesized that IDB4 may be a candidate tumor suppressor gene in cancer (liu05_a)

14. We conclude that HIC1 is a ***candidate tumor-suppressor gene*** for which loss of function in both mouse and human cancers is associated only with epigenetic modifications. (che03_l)

In example 13 above an initial hypothesis is cited as having been the identification of a *candidate tumor-suppressor gene*, whilst conversely in example 14 it is the conclusion that HIC1 is a *candidate tumor-suppressor gene* of a modified form, where 'loss of function in both mouse and human cancers is *associated only with* epigenetic modifications'.

4.2.4 *X is a tumor suppressor gene*

A further though much less frequent means of expressing the status of a gene as a *tumor suppressor gene* was the unhedged use of the copula, and this occurred five times, as in the following examples:

15. TSLC1 is a ***tumor-suppressor gene*** in human non-small-cell lung cancer (kur01_l)

16. SUFU is a newly identified ***tumor-suppressor gene*** that predisposes individuals to medulloblastoma by modulating the SHH signaling pathway through a newly identified mechanism. (tay02_a)

These examples again label a given gene as being a *tumor suppressor gene*, though through a slightly different form from the named *tumor suppressor gene* strategy seen above (4.2.1). Though there are not enough examples here to be able to make any confident generalisations, it would appear that this unhedged use is found at or around the point of discovery; in example 15 above *TSLC1* being a *tumor suppresor gene* is the main finding of the paper, whilst in example 16 the status of *SUFU* as a *tumor suppressor gene* is explicitly marked as being *newly identified*. However, the following further example of this type indicates that the use of the copula can still be associated with a more hedged claim:

> **17**. The observation of bi-allelic alterations in TCF1 in human liver tumors meets the criteria of the classical two-hit recessive model of oncogenesis 23, 24 and supports the hypothesis that TCF1 is a ***tumor-suppressor gene*** that is altered early in carcinogenesis, leading to adenoma formation. (blu02_l)

In this example *tumor suppressor gene* occurs in the copula construction *TCF1 is a tumor-suppressor gene* but in this case that construction itself occurs in a that- clause within ‘*supports the hypothesis that*’ *TCF1 is a tumor suppressor gene*. Whilst this still appears to be contributing to a claim of the type that *X is a tumor suppressor gene* this positioning within a that- clause constitutes a modification and slight hedging of the claim, indicating that the copula form may still be positioned within a hedged claim. It should be noted that ‘supports’ is a known hedging device but the string *supports the hypothesis that* is still worth adding to our list of n-grams with hedging functions in academic writing.

4.2.5 classic/classical tumor suppressor gene

The use of the label *classic* or *classical* was found to be a further strategy relating to the ontological status of a gene as a *tumor suppressor gene*. In this case it appears to have a strengthening effect on the claim, and seems to constitute an even stronger claim than either of the previous forms discussed above since the use of CLASSIC can be understood in the sense that what has been found is a prototypical example where the evidence is exactly and ideally in accordance with the ontological criteria. The following examples illustrate this phenomenon:

18. Thus, VHL acts as a classic ***tumor-suppressor gene*** that is inactivated according to Knudson's two-hit hypothesis1. (cor03_a)

19. This classical ***tumor-suppressor gene*** is completely inactivated in HCT116 cells by a frameshift mutation of one unmethylated allele and hypermethylation of the other allele7. (tin04_bc)

Indeed, this connection is explicitly made in example 18 above, where the writer states that *VHL* meets *Knudson's two-hit hypothesis*. However this description is complicated by the writer's use of '*acts as'* as the process in this clause, rather than, for example, the copula. It would appear that '*VHL acts as a classic tumor-suppressor gene*' falls somewhat short of the proposition 'VHL is a classic tumor suppressor gene' and yet the use of the word *classic* appears to indicate that the classification criteria have been (ideally) met.

4.2.6 The frame X the X of + *tumor suppressor gene*

Another context for *tumor suppressor gene* is the frame *X the X of + a tumor suppressor gene*, which occurred four times. The similarity in meaning expressed by strings instantiating this pattern can be seen in figure 4 below:

X	**the**	**Y**	**of**	**a tumor suppressor gene**
implying	the	existence	of	a tumor suppressor gene
indicating	the	presence	of	a tumor suppressor gene
suggesting	the	presence	of	a tumor suppressor gene
in agreement with	the	inactivation	of	a tumor suppressor gene

Figure 4: Table illustrating the use of the frame X *the* Y *of* + *a tumor suppressor gene* in *genecorp*

In each of the examples *tumor suppressor gene* appears to occur in the context of a hedged claim. Each of these four examples connects some evidence with the possibility that the conclusion to be drawn is that there is a *tumor suppressor gene* present. The wordforms in the X position appear to have a shared meaning of ‘suggests’ whilst the wordforms in the Y slot seems to have a shared meaning of ‘presence’. Whilst no one word is always present in either of these two slots, the frame itself can be glossed as carrying the meaning of ‘suggests the presence of’ a *tumor suppressor gene*.

4.2.7 functions as a *tumor suppressor gene*

The form *functions as a tumor suppressor gene* can also be found four times in *genecorp*. This again appears to be a further example of a lexical expression of a hedged ontological status, since it again fall short of constituting a form such as '*X is a tumor suppressor gene'*, as in the following examples:

20. We conclude that SUFU functions as a ***tumor-suppressor gene*** in a subset of desmoplastic medulloblastomas. (tay02_a)

21. NF1 functions as a ***tumor-suppressor gene***, and loss of heterozygosity in somatic tissues has been associated with tumor formation3. (git03_l)

22. Our results indicate that Notch1 functions as a ***tumor-suppressor gene*** in mammalian skin. (nic03_l)

4.2.8 *tumor suppressor gene* and the lemma *KNOW*

A far more predictable strategy is the use of the lemma KNOW though interestingly this occurs only five times within the examples of *tumor suppressor gene,* and only once is it relevant to an epistemic claim about a *tumor suppressor gene*, in the following example:

23. can act as a tumor-suppressor gene in paraganglioma genesis but is not known to be a breast *tumor suppressor gene* (kur02_bc3)

Whilst the lemma KNOW would therefore not appear to be a frequent strategy for signaling epistemic status around the string *tumor supressor gene* it is of course likely to be a frequent (explicit) device for epistemic signaling in the corpus more widely (see 5 below).

4.2.9 *may*

Finally a predictable example of grammatical modality emerged as a further means of epistemic signaling around the categorisation of a named gene as a *tumor suppressor gene* with the wordform *may* occurring five times as a hedging device, as in the following examples:

24. suggests that PTPRJ **may** be a *tumor suppressor gene* acting in human colorectal cancer. (rui02_l)

25. our findings indicate that RB1CC1 **may** be a *tumor-suppressor gene* in breast cancer. (cha02_l)

26. The clinical significance of hypermethylation across chromosome 2q14.2 is unclear, but the fact that it is a common event suggests that regions within the cytogenetic band **may** encode possible *tumor suppressor gene(s)*. (fri06_a)

4.3 Is caused by or is associated with? Claims containing the keyword *mutations*

This section describes the epistemic patterns surrounding the nodes containing *mutations*. Three clusters containing *mutations* were identified (figure 2 above) and these fell into two broad categories: two of the clusters containing *mutations* (*loss-of-function mutations* and *mutations in the gene encoding*) exhibited very similar epistemic patterns both in terms of the types of claim that they were typically constituent of and the way in which these claims were nuanced linguistically. These claims were again not typically nuanced through use of the ‘usual suspects’ but rather through variation in the verb group used to characterize the claim. In the third cluster (*disease causing mutations*) the claims were of a slightly different type, presumably because a causal claim is already encoded in the string *disease causing mutations*. The principal patterns found in the cotext of the strings containing *mutations* will be presented in the remainder of this section.

4.3.1 Unhedged causal claims

The most common type of claim containing both *loss of function mutations* and *mutations in the gene encoding* were unhedged causal claims. These most often involved the lemma CAUSE, with inflections of this base being present in 62 of 174 concordance lines for *mutations in the gene encoding* and 40 of the 219 instances of *loss-of-function mutations*. In the spirit of corpus-driven analysis it was not assumed that the separate wordforms associated with the lemma CAUSE would function consistently but analysis of these showed that they operated in very similar ways, giving agency in causal claims to *mutations* in given genes and connecting this to specific deleterious observable effects, as in the following examples:

27. *Loss-of-function mutations* in Tub **cause** late-onset obesity, retinal degeneration and hearing loss in tubby mice4, 5, 6. (mak06_l)

28. *Mutations in the gene encoding* 3bold beta-hydroxysteroid-Delta 8,Delta7-isomerase **cause** X-linked dominant Conradi-Hunermann syndrome (bra99_l)

29. Tangier disease **is caused by** *mutations in the gene encoding* ATP-binding cassette transporter 1 (rus99_l)

Whilst the lemma CAUSE was the most frequent means by which causal claims were made, a range of other verb groups expressing causal meaning was also found, with 37 instances for *mutations in the gene encoding* and 62 for *loss-of-function mutations*.

30. ***Loss-of-function mutations*** in the cathepsin C gene ***result in*** periodontal disease and palmoplantar keratosis (too99_l)

31. ***Mutations in the gene encoding*** B, a novel transporter protein, **reduce** melanin content in medaka (fuk01_l)

It should be noted that the most common type of claim found around this node phrase is unhedged. Thus whereas previous studies (cf. Carver et al. 2008) have suggested that 'deterministic' discourse about genetics is a feature of 'unscientific' mass media writing we see here that professional geneticists themselves express findings which are unhedged and causal in nature. What appears to differ from the forms found in the media in the findings of previous studies is that here it is the *mutations* in the gene which are given the causal role, not the gene itself.

4.3.2 The identification of *disease causing mutations*

The principal pattern discovered in the concordance lines containing *disease causing mutations* was the discovery of epistemic claims regarding the existence of *disease causing mutations*, as in the following example:

32. Next, we sequenced MKS1 in 22 non-Finnish MKS families available to us and **identified** *disease-causing mutations* in four of them (Table 1). (kyt06_bc)

This is again an unhedged claim that *disease causing mutations* have been found. In one of the few examples where this claim is hedged, the writers again use *putative* as the hedging device:

33. Our analysis of DNA samples from Alexander disease patients **has identified *putative*** *disease-causing mutations* in four amino acids in the rod and tail domains of GFAP (Fig. 3). (bre01_l)

This leads to a rather incongruous statement where the process is not hedged (*has identified*) and yet the object of the discovery is. What is in fact uncertain is presented as something that has been ‘identified’. Although it is difficult to assess the rhetorical function of such forms intuitively, my geneticist informer felt that this form would be seen as preferable to claiming that ‘we have possibly identified’ *disease causing mutations*. There seems to be some empirical evidence to support this intuition: *may* occurs as a left side collate of *identify* just 37 times out of 3,987 examples (ie. just over one percent of the time) whilst from 2346 examples of *identification* in the corpus there are few examples of hedging features in the L1 position with *suggested, probably*, *predicted* and *potentially* the most frequent occurring just three times each.

4.3.3 Variation of verb group to fall short of a causal claim: ‘predispose’ group

Claims falling short of these outright causal or identification claims were principally found to be nuanced by the use of alternative verb patterns to characterise the relationship between mutations and various syndromes. Thus in the following examples *mutations* do not *cause* but are merely *associated with* the named features:

34. ***Loss-of-function mutations*** in TRPM6 **are associated with** hypomagnesemia with secondary hypocalcemia, a rare autosomal-recessive disorder8. (gud05_nav)

35. Human mitochondrial DNA deletions **associated with *mutations in the gene encoding*** Twinkle, a phage T7 gene 4-like protein localized in mitochondria (spe01_a)

Again what is interesting here is that geneticists seem never to report that mutations *possibly* cause X. Whilst this might be thought of as being an ontological issue rather than an epistemological one (in the sense that the geneticists, it might be argued, are not claiming a possible cause, but merely an association), the patterning around these nodes should give us pause for thought. If geneticists present findings that *mutations* are *associated* with X and *cause* X why should they not also speculate that certain *mutations in the gene encoding* may *cause* X? In terms of the linguistic realisations present, there seems to be an obvious ‘jump’ from claiming an association to claiming a cause. Moreover, given that geneticists are so often seeking to make unhedged causal claims when they identify *mutations in the gene encoding* or *loss-of-function mutations* it seems reasonable to assume that findings of *associations* between *mutations* and various effects imply a *possible* relation of causation without expressing this linguistically.

4.3.4 Removal of verb group to create epistemic implicature

Finally cases were found where not even *associations* between *mutations* and disorders are being claimed. Instead, the mere existence of specific *mutations* in groups of individuals with a particular feature is regarded as a finding, as in the following examples:

36. Here we report ***mutations in the gene encoding*** RANKL (receptor activator of nuclear factor ligand) in six individuals with autosomal recessive osteopetrosis whose bone biopsy specimens lacked osteoclasts. (sob07_bc)

37. We **identified** ***loss-of-function mutations*** in ATP6V0A2, encoding the a2 subunit of the V-type H+ ATPase, in several families with autosomal recessive cutis laxa type II or wrinkly skin syndrome. (kor08_bc)

In such cases there is clearly the implication that there might be a significant relationship between the *mutations* and the disorder and presumably the hope is that this will eventually be proven to involve a causal link. Again, what is interesting here is that the lack of a finding of a causal link does not stop this constituting a publishable result. It should be noted that these examples are both ‘brief communications’ and can thus be understood as an earlier stage in the process and yet the geneticists do not provide any explicit characterization of the possible nature of the link between the *mutations* and the disorder. This is a particularly interesting type of example in terms of demonstrating the usefulness of the corpus-driven approach. Searching for hedging devices in a corpus would not give us examples such as these because they do not contain hedging devices, or any explicit linguistic realisation of the underlying claim. Rather, it is the inspection of the concordance lines of many examples of *mutations in the gene encoding* and *loss-of-function mutations* that shows us that when these types of *mutations* are juxtaposed with a specific named disorder, geneticists are usually seeking a causal link. Thus, to a geneticist, simply juxtaposing the two creates an implied speculation- that there may be a link. But this is not made explicit, and therefore cannot be the object of a corpus search.

4.3.5 Summary of corpus driven findings for *mutations*

The following table summarises the verbal patterns found in relation to node phrases containing *mutations* and attempts to describe the epistemic function of these:

Pattern	Epistemic Function	Examples of forms identified
CAUSE group	To make a causal claim involving *mutations*	CAUSE*; LEAD* to; IMPAIR*; are due to; PRODUCE*;RESULT* + in; RESULT* + from; STOP*; TRIGGER*; UNDERLIE*
PREDISPOSE group	To posit a causative connection between mutations and a disorder that falls short of a full causative claim	PREDISPOSE*; INVOLVE*
ASSOCIATED group	To express an association between *mutations* and a disorder without expressing a causal connection	ASSOCIATE*; LINK*;
COPULA group	To identify *mutations*	is; are
IMPLICATURE group	To juxtapose *mutations* with a disorder without characterizing the connection between the two linguistically	have; has
EFFECTS and **CONSEQUENCES** group	To discuss the effects of *mutations*; To assess the effects of *mutations*; To speculate as to the effects of *mutations*	EFFECT*; CONSEQUENCE*

Figure 5: Summary of results of corpus-driven analysis for *loss of function mutations* and *mutations in the gene encoding*

5 Discussion

The inductive analysis of the node phrases in this article has demonstrated a range of different epistemic claims. The prevalence of unhedged claims was clear in all nodes discussed but when geneticists sought to limit their claims they only rarely used the 'usual suspects' to do so. Most strikingly, the wordforms *putative* and *candidate* emerged as being the most frequent hedging devices found in the case of *tumor suppressor genes*. It is perhaps worth just dwelling on the significance of this finding for a moment. What has been shown is that, in ten years of publications in *Nature Genetics*, no geneticist ***ever*** reported that:

X is a *possible* tumor suppressor gene

or

X is a *probable* tumor suppressor gene

or

X is a *likely* tumor suppressor gene

Rather, in almost all hedged cases they stated that they had discovered a *putative* or *candidate tumor suppressor gene*. This speaks directly to what Gledhill (2000) has called 'the preferred ways of saying things' in a discourse community and might well be a startling revelation to an academic attempting to publish a possible discovery of a *tumor suppressor gene* in *Nature Genetics*, to whom we would of course immediately recommend that they dub it either a *candidate* or *putative tumor suppressor gene*. In order to get a sense of the significance of this finding, and of the success (or otherwise) of this inductive corpus-driven method in developing an accurate picture of hedging in genetics from a ‘bottom-up’ perspective, it is instructive to now view whole corpus frequency figures for the fifty most frequent wordforms carrying epistemic meaning found in *genecorp*:

rank	item	frequency	rank	item	frequency
1	shown	13892	26	indicates	2557
2	identified	9804	27	indicating	2532
3	showed	7741	28	suggests	2312
4	known	5775	29	probably	2217
5	indicated	5525	30	correlation	2021
6	detected	5282	31	statistical	2021
7	should	4479	32	detect	1997
8	evidence	4334	33	estimated	1855
9	indicate	4038	34	putative	1756
10	possible	3285	35	support	1691
11	suggest	3187	36	significance	1678
12	absence	3118	37	revealed	1635
13	shows	2983	38	often	1618
14	due	2864	39	few	1599
15	resulting	2858	40	unknown	1577
16	significantly	2850	41	hypothesis	1559
17	expected	2828	42	suggested	1544
18	potential	2816	43	responsible	1524
19	suggesting	2815	44	report	1489
20	confirmed	2793	45	possibility	1476

21	candidate	2761	46	causes	1440
22	approximately	2750	47	seems	1430
23	negative	2674	48	associations	1406
24	cause	2657	49	established	1387
25	likely	2614	50	provides	1382

Figure 6: The fifty most frequent wordforms carrying epistemic meaning in *genecorp*

Candidate appears twenty-first in the list, whilst *putative* appears thirty-fourth, and it seems to be clear from both of these figures and from the evidence from the concordance line analysis that *candidate* and *putative* are amongst the most frequent hedging devices in the field of genetics. Most striking perhaps is that *candidate* is more frequent than *approximately*, *suggest* and *probably*, whilst both words are more frequent than *few*, *significance*, *hypothesis, possibility* and *report*.

When we compare the frequencies of *putative* and *candidate* to the most frequent clusters that function as hedges, this claim looks even more compelling. The most frequent three word cluster which carries epistemic meaning, *we found that,* occurs 1400 times, meaning that both *putative* and *candidate* are more frequent than any three word cluster carrying epistemic meaning, as figure 7 demonstrates:

rank	cluster	frequency
1	we found that	1400
2	consistent with the	1300
3	the basis of	1272
4	in the absence	1208
5	on the basis	1204
6	the role of	1034
7	based on the	1014
8	is associated with	1000

9	the effect of	968
10	the identification of	861

Figure 7: The ten most frequent three word clusters in *genecorp*

Finally when we consider four-word clusters of the type that Biber (2009) investigates when studying formulaic language, we again see ten examples of the ‘usual suspects’ which are again considerably less frequent than *candidate* and *putative*, with seemingly ubiquitous phrases such as *these results suggest that* and *these results indicate that* occurring just several hundred times, as illustrated by figure 8:

rank	cluster	frequency
1	the presence of a	381
2	has been shown to	377
3	as a result of	365
4	is consistent with the	363
5	these results indicate that	338
6	these results suggest that	335
7	have a role in	328
8	we found that the	319
9	we did not detect	311
10	presence or absence of	310

Figure 8: The ten most frequent four word clusters in *genecorp*

This set of results raises the possibility that hedging strategies are far more specific to particular academic disciplines that previously assumed. Whilst *putative* and *candidate* might be common in other scientific disciplines they might also be distributed in very specific areas or even (particularly in the case of *candidate*) be unique to the field of genetics. These possibilities threaten the usefulness of general academic wordlists proposed in works such as Coxhead (2000); Simpson-Vlach and Ellis (2010) and Gardner and Davies (2014); and lend support to previous critiques of such lists (cf. Hyland and Tse 2007) which have argued that considerable disciplinary variation is being glossed over in the attempt to produce a universally usefully ‘general’ list of academic words or structures. At the very least they suggest that it would be advisable to supplement such general academic wordlists with discipline-specific lists if those seeking to use them are going to be able to access the ‘preferred way of saying things’ in their chosen field. Given that it is now a fairly straightforward task to produce a disciplinary-specific wordlist, such supplementary work should not prove prohibitively onerous and indeed the production of disciplinary-specific wordlists is already becoming fairly common (e.g., Mudraya 2006; Martinez, Beck and Panza 2009; Liu and Han 2015). It is the suggestion of this study that these lists might also take into account possible local variations of hedging patterns and that these, as well as the more common disciplinary ontologies, should form part of this type of supplementary work.

The material presented here has focused on the findings surrounding just the node phrases *tumor suppressor gene*, *mutations in the gene encoding*, *loss-of-function mutations* and *disease causing mutations*. Investigation of the patterns surrounding these strings revealed that the main epistemic issue surrounding these node phrases was one of ontological categorisation; a process where what is at issue scientifically is whether or not a given entity is to be classed in a particular way or given a specific label. What has proven particularly interesting about this process is that in *genecorp* the linguistic means of nuancing claims around this process is not a good fit with what our typical expectations of what the most frequent hedging devices might be, as suggested, for example, by Hyland (1998). Whilst the concordance lines featuring these phrases provide plentiful examples of geneticists falling short of making outright claims such as X **is** a *tumor suppressor gene* or *loss-of-function mutations* **cause** Y, rare indeed are the examples of modal adjectives, grammatical modality or what Hyland (1998) calls ‘epistemic lexical verbs’ in these concordance lines. Indeed, one might say that the ‘usual suspects’ hardly seem to be present at all. Rather, geneticists prefer to use the wordforms *candidate* and *putative* or verb phrase patterns such as *are linked to* and *are associated with* to signal modal meaning. Given that a prime objective for the writers is gaining acceptance for publication for their finding in what is the most prestigious journal in the field it is of course tempting to speculate as to what the rhetorical function of a claim such as:

X is a *putative tumor suppressor gene*

Especially when compared to what might be considered more congruent forms such as:

X *might be* a *tumor suppressor gene*

Or

X is a *possible tumor suppressor gene*

One of the ways on which this study might be followed up would be to carry out some structured interviews with geneticists to try to establish why *putative* seems to them to be a preferable expression of modality than *possible* and of course the suspicion must be that the rhetorical effect of *putative* is somehow one that is less obviously uncertain than *possible*.

In a sense there is a tension between the corpus-driven method and the findings since (arguably) the most interesting finding (that *putative* and *candidate* frequently function as hedging devices) is probably identifiable from collocation data. Contrary my findings elsewhere about the relationship between *mutations* and causative language (Plappert: 2012) when it comes to identifying L1 hedges of a node item, collocation data is both useful and accurate in assessing the frequency of the pattern; indeed, it will identify every example. Nevertheless, detailed analysis of the concordance lines containing *tumor suppressor gene* provides us with a rich and thick description of the functioning of these items and comparison to the other patterns found leaves us confident that *candidate* and *putative* do indeed perform a hedging role. In this sense the combination of the corpus-driven concordance line analysis, the viewing of collocation data and the comparison to other forms through whole corpus frequencies provides a triangulation of methods that can leave us with a hypothesis of which we can be extremely confident; that geneticists frequently use *candidate* and *putative* as hedging devices.

It should be noted that there were also five instances of the modal auxilliary *may* in the collocation data for *tumor suppressor gene*, confirming that this predictable epistemic signaling device is used to hedge the status of an object as a *tumor suppressor gene*, as in the following examples:

24. suggests that PTPRJ **may** be a *tumor suppressor gene* acting in human colorectal cancer. (rui02_l)

25. our findings indicate that RB1CC1 **may** be a *tumor-suppressor gene* in breast cancer. (cha02_l)

Another infrequent form identified was the use of the label *classic* or *classical*. This was also found to be a linguistic strategy relating to the ontological status of a gene as *tumor suppressor gene,* with three such examples being identified. In such cases the modifying adjective appears to be functioning as an intensifier marking the given object as an archetypal example of a *tumor suppressor gene*, as in the following example:

19. This classical ***tumor-suppressor gene*** is completely inactivated in HCT116 cells by a frameshift mutation of one unmethylated allele and hypermethylation of the other allele7. (tin04_bc)

In addition to these examples the form *functions as a tumor suppressor gene* was found four times in *genecorp*. This again appears to be a further example of a lexical expression of a hedged ontological status, since it again fall short of constituting a form such as '*X is a tumor suppressor gene'*.

Finally the frame 'X the Y of + *tumor suppresor gene*' was identified as a further pattern of epistemic significance, as illustrated by figure four above. This provides a fascinating link with previous corpus-driven work since it seems to be a similar example to Hoey's (2004) famous example of the time + distance + journey pattern in that it appears to constitute an underlying semantic regularity which would again be difficult to identify automatically, especially when the words instantiating both the X and Y positions can vary. All four examples appear to function similarly in appraising that the available evidence supports the involvement of a *tumor suppressor gene,* whilst falling short of an outright assertion that a *tumor suppressor gene* is present. As such this frame seems to provide a yet further means of expressing the possibility of the presence of a given object. My conclusion is that the corpus-driven method remains a valuable source of unexpected findings for the corpus linguist and that it remains a *candidate* method for any linguist wishing to approach the data with fresh eyes.

Acknowledgment.

The research reported in this study was fully funded by the Arts and Humanities Research Council (AHRC) of the United Kingdom.

References

Baker, P. 2006. Using corpora in discourse analysis. London: Continuum.

Biber, D. 2009. ‘A corpus-driven approach to formulaic language in English’, International Journal of Corpus Linguistics 14 (3), pp. 275-311.

Brown, P. and S. Levinson. 1987. Politeness: Some universals in language usage. Cambridge: CUP.

Carver, R., R. Waldahl, and J. Breivik. 2008. ‘Frame that Gene: A tool for analysing and classifying the communication of genetics to the public’, EMBO reports 9: pp. 943-947.

Chih-Hua, K. 1999. ‘The use of personal pronouns: Role relationships in scientific journal articles’, English for Specific Purposes 18, pp. 121-138.

Condit, C. M. 1999. ‘How the Public Understands Genetics: Non-deterministic and Non-discriminatory Interpretations of the "Blueprint" Metaphor’, Public Understandings of Science, 8 (3), pp. 169-180.

Cortes, V. 2004. ‘Lexical bundles in published and student disciplinary writing: Examples from history and biology’, English for Specific Purposes 23 (4), pp. 397-423.

Coxhead, A. 2000. ‘A new academic wordlist’, Tesol Quarterly 34 (2), pp. 213-238.

Crystal, D. and D. Davy. 1975. Advanced Conversational English. London: Longman.

Gabrielatos, C. and P. Baker. 2008. 'Fleeing, Sneaking, Flooding: A corpus analysis of discursive constructions of refugees and asylum seekers in the UK press, 1996-2005', Journal of English Linguistics 36, pp. 5-38.

Gardner, D. and M. Davies. 2014. 'A new academic vocabulary list', Applied Linguistics 35 (3), pp. 305-327.

Gledhill, C. 1995. 'Collocation and genre analysis: the phraseology of grammatical items in cancer research abstracts and articles', Zeitschrift fur Anglistik and Amerikanistik 43, pp.11-29.

Gledhill, C. 1996. 'Science as collocation: phraseology in cancer research articles' in S. Botley, J. Glass, T. McEnery and A.Wilson (eds) Proceedings of the Teaching and Language Corpora 1996. UCREL Technical papers 9, pp. 108-126.

Groom 2007. Phraseology and epistemology in humanities writing: a corpus-driven study. Unpublished PhD thesis. University of Birmingham.

Groom, N. 2010. 'Closed-class keywords and corpus-driven discourse analysis' in M. Bondi and M. Scott (eds) Keyness in Texts. Amsterdam and Philadelphia: Benjamins.

Hoey, M. 2004. Lexical Priming. A new theory of words and language. London: Routledge.

House, J. and G. Kasper. 1981. 'Politeness markers in English and German' in F. Coulmas (ed.) Conversational routine, pp. 157-185. The Hague: Mouton.

Hubbard, R. and E. Wald. 1993. Exploding the gene myth: How genetic information is produced and manipulated by scientists, physicians, employers, insurance companies, educators, and law enforcers. Boston: Beacon.

Hunston, S. 1995. ‘A corpus study of some English verbs of attribution’, Functions of Language 2, pp. 133-158.

Hunston, S. 2002. Corpora in Applied Linguistics. Cambridge: Cambridge University Press.

Hyland, K. 1996. ‘Writing without conviction? Hedging in scientific research articles’, Applied Linguistics 17 (4), pp. 433-454.

Hyland, K. 1998. Hedging in scientific research articles. Amsterdam: John Benjamins.

Hyland, K. 2008. ‘As can be seen: Lexical bundles and disciplinary variation’, English for Specific Purposes, 27(1), pp. 4-21.

Hyland, K. 2009. Academic Discourse. London: Continuum.

Hyland, K. and P. Tse. 2004. Metadiscourse in academic writing: a reappraisal’, Applied Linguistics, 25 (2), pp. 156-77.

Hyland, K. and P. Tse. 2007. ‘Is there an “Academic Vocabulary”?’, Tesol Quarterly 41 (2), pp. 235-253.

James, A. 1983. ‘Compromisers in English: A cross-disciplinary approach to their interpersonal significance’, Journal of Pragmatics 7, pp. 191-206.

Kuhn, T. S. 1970. ‘Logic of discovery or psychology of research’ in I.A. Lakatos (ed) Criticism and the growth of knowledge. Cambridge, Cambridge University Press.

Lakoff, G. 1972. ‘Hedges: A study of meaning criteria and the logic of fuzzy concepts’, Chicago Linguistic Society Papers 8, pp. 183-228.

Liu, J. and L. Han. 2015. ‘A corpus-based environmental academic wordlist building and its validity testing’, English for Specific Purposes 39, pp. 1-11.

Low, G. 1996. ‘Intensifiers and hedges in questionnaire items and the lexical invisibility hypothesis’, Applied Linguistics 17 (1), pp. 1-37.

Martinez I.A., S. C. Beck, and C.B. Panza. 2009. ‘Academic vocabulary in agriculture research articles: A corpus-based study’, English for Specific Purposes 28, pp. 183-198.

McEnery, T. and A. Hardie. 2012. Corpus Linguistics: Methods, Theory and Practice. Cambridge: Cambridge University Press.

Mudraya, O. 2006. ‘Engineering English: A lexical frequency instructional model’, English for Specific Purposes 25, pp. 235-256.

Nelkin, D. and S. Lindee 1995. The DNA Mystique: The Gene as Cultural Icon. New York: W. H. Freeman.

Noguchi, J., T. Orr, and Y. Tonio. 2006. Using a dedicated corpus to identify features of professional English usage: What do 'we' do in science journal articles? In A. Wilson, D. Archer, and P. Rayson (eds) Corpus Linguistics around the world, pp. 155-166. Amsterdam and New York, Rodopi.

Oakey, D. 2008. The form and function of fixed collocational patterns in research articles in different academic disciplines. Unpublished PhD thesis. Leeds University.

Plappert, G. 2012. Phraseology and epistemology in scientific writing: A corpus-driven approach. Unpublished PhD thesis. University of Birmingham.

Quirk, R., S. Greenbaum, G. Leech, and J. Svartik. 1972. A grammar of contemporary English. Harlow: Longman.

Rayson, P. 2009. Wmatrix: a web-based corpus processing environment. Computing Department, Lancaster University. http://ucrel.lancs.ac.uk/wmatrix/

Rogers, M. 2007. ‘Lexical chains in technical translation: A case study in indeterminacy’ in. B.E. Antia (ed.) Indeterminacy in Terminology and LSP, pp. 15-35. Amsterdam: John Benjamins.

Scott, M. 2000. ‘Focusing on the text and its key words’ in L. Burnard and T. McEnery (eds) Rethinking Language pedagogy from a corpus perspective, pp. 103-122. Frankfurt: Peter Lang.

Scott, M. 2004. WordSmith Tools version 7, Stroud: Lexical Analysis Software.

Shen, L. & C.M. Condit. 2012. ‘Addressing fatalism with health messages’ in H.Cho (ed.) Communication in public health, pp. 191-208. Los Angeles: Sage.

Simpson-Vlach, R. and N.C. Ellis. 2010. ‘An academic formulas list : New methods in phraseology research’, Applied Linguistics 31 (4), pp. 487-512.

Sinclair, J. M. 1991. Corpus, concordance, collocation. Oxford, Oxford University Press.

Sinclair, J. M. 2003. Reading Concordances. London: Pearson Longman.

Stubbe, M. and J. Holmes. 1995. ‘You know, eh and other ‘exasperating expressions’: an analysis of social and stylistic variation in the use of pragmatic devices in a sample of New Zealand English’, Language and Communication 15 (1), pp. 63-88.

Tognini-Bonelli, E. 2001. Corpus linguistics at work. Amsterdam: J. Benjamins.

Thompson, G. and Y. Ye. 1991. ‘Evaulation in the reporting verbs used in academic papers’, Applied Linguistics 12, pp. 365-382.

Thomson Reuters. 2015. Journal Citation Report, Science Edition.

Williams, I.A. 1996. ‘A contextual study of lexical verbs in two types of medical research report: clinical and experimental’, English for Specific Purposes 15, 3: pp. 175-197.

Tribble, C. 2000. ‘Genres, keywords, teaching: towards a pedagogic account of the language of project proposals’ in L. Burnard and T. McEnery (eds), Rethinking language from a corpus perspective, pp. 75-90. Frankfurt: Peter Lang.