Candidate Knowledge? Exploring epistemic claims in scientific writing: A corpus-driven approach
Garry Plappert (University of Birmingham, United Kingdom)

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., 2006; 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 an inductive 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.

Background

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). 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 analysing items from wordlists, allowing the analyst to proceed with collocation or concordance line based description. 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.

Methods

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. The texts for this study came from a ten-year period (1999-2008 inclusive) and were collected together in a corpus known as genecorp. In total genecorp contains 2,979 texts from the journal Nature Genetics, spanning from 1999-2008.

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)

Results

The following table exemplifies the results of this study by summarising the verb phrase patterns found in relation to node phrases containing mutations and attempts to describe the epistemic function of these:

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Epistemic Function</th>
<th>Examples of forms identified</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAUSE group</td>
<td>To make a causal claim involving mutations</td>
<td>CAUSE*; LEAD* to; IMPAIR*; are due to; PRODUCE*;RESULT* + in; RESULT* + from; STOP*; TRIGGER*; UNDERLIE*</td>
</tr>
</tbody>
</table>
The inductive analysis of the node phrases in this article demonstrates 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. Rather, when they did seek to nuance claims they tended to use either much less familiar explicit hedging devices (such as putative or candidate) or otherwise typically modified the verbal group used to form the claim, leading to claims such as is linked to or is associated with instead of is caused by or other unhedged causal claims. 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.

References


