# 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:

| 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 1: Summary of results of corpus-driven analysis for *loss of function mutations* and *mutations in the gene encoding***

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


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

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

Coxhead, A. 2000. 'A new academic wordlist', *Tesol Quarterly* 34 (2), 213-238.

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

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

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.
House, J. and G. Kasper. 1981. 'Politeness markers in English and German' in F. Coulmas (ed.) *Conversational Routine*, 157-185. The Hague: Mouton.
Hunston, S. 1995. 'A corpus study of some English verbs of attribution', *Functions of Language* 2, 133-158.
Hyland, K. 1998. *Hedging in scientific research articles*. Amsterdam: John Benjamins.
Hyland, K. 2009. *Academic discourse*. London: Continuum.
Hyland, K. and P. Tse. 2004. Metadiscourse in academic writing: a reappraisal', *Applied Linguistics*, 25 (2), 156-77.
Lakoff, G. 1972. 'Hedges: A study of meaning criteria and the logic of fuzzy concepts', *Chicago Linguistic Society Papers* 8, 183-228.
Low, G. 1996. 'Intensifiers and hedges in questionnaire items and the lexical invisibility hypothesis', *Applied Linguistics* 17 (1), 1-37.
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.
Plappert, G. L. *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.
Simpson-Vlach, R. and N.C. Ellis. 2010. 'An academic formulas list: New methods in phraseology research', *Applied Linguistics* 31 (4), 487-512.
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), 63-88.
Thomson Reuters. 2015. *Journal Citation Report,* Science Edition.
Thompson, G. and Y. Ye. 1991. 'Evaluation in the reporting verbs used in academic papers', *Applied Linguistics* 12, 365-382.
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), 175-197.