

Collocation and the rhetoric of scientific ideas. Corpus linguistics as a methodology for genre analysis.

Chris Gledhill

Aston University, LES, Birmingham B4 7ET, UK

KEYWORDS: cancer research, corpus linguistics, genre analysis

AFFILIATION: Languages and European Studies, Aston University.

E-MAIL: C.J.Gledhill@aston.ac.uk
PHONE NUMBER: 359 3611 ext 4232

Abstract

This paper attempts to integrate the ethnographic approach of genre analysis (Swales 1990) with the large scale computational analysis of phraseology in the field of corpus linguistics (Sinclair 1987a *inter alia*). In particular the author attempts to describe how language is used in hard science, how scientists create new science in their writing and how language functions in extremely specialised circumstances. The paper describes the working context of cancer research articles at Aston University's Pharmaceutical Sciences Department and uses the statistical analysis of lexis in rhetorical sections of research articles to characterise these sections in terms of their collocational and discoursal properties.

Extended abstract

One hypothesis is that new science is actually embodied in research articles by a process of reformulating concepts within the text. To test this, the scientific claims of a sample of ten texts are analysed in terms of reformulation of grammatical metaphor, discourse signalling and posture (Halliday 1985, Sinclair 1981). A second hypothesis is that new science is founded on a system of preferred expressions, and that collocation is a fundamental mechanism that allows for new formulations to take place throughout the text. A corpus analysis of 150 cancer research articles (The pharmaceutical sciences corpus: 500,000 tokens) is undertaken to characterise the phraseology of grammatical items in research articles and in the various rhetorical sections of research articles namely Titles, Abstracts, Introductions, Methods, Results and Discussion sections.

The paper finds that research articles use language to create new science by reformulating data as research models and by altering the established

patterns of phraseology. Collocation is seen to vary systematically in rhetorical sections, and the concept of phraseology is postulated as a preferred way of expressing a delimited set of semantic and communicative roles. Science should therefore not be seen as a body of facts transmitted via language, but as a special linguistic construct, mediated by the mechanisms of textual reformulation and phraseological innovation.

Abstracts written by authors have been characterised in terms of morpho-syntactic features, especially verb tense and modality (Hanania and Akhtar 1985, Malcolm 1987, Gunawardena 1989, Salager-Meyer 1992). From a more rhetorical approach, discourse and analysis of abstracts has involved comparison of rhetorical moves between abstracts and articles (Nwogu 1989, Endres-Niggemeyer 1985, Salager-Meyer 1992) and thematic choice between successful and non-successful abstracts (Gibson 1992, Drury 1991). This paper applies a phraseological methodology to an area that has been relatively well documented in information science and text linguistics, but less so in genre analysis: the collocational properties of abstracts as they compare to those of the research article. Some phraseological features, such as explicit discourse signals, have already been identified in the genre analysis of scientific articles (Oster 1981, Tadros 1985, Master 1987, Brett 1994) and to a lesser extent in abstracts (Diodato 1982, Zambrano 1987), but to my knowledge there has been no general phraseological comparison of articles and abstracts as yet (see Gledhill 1995a for a more balanced picture).

In this article I shall exploit the notion of phraseology, defining it as: a system of preferred expressions differentiated by the rhetorical aims of a discourse community. I shall treat phraseology as a lexical and a discoursal phenomenon. In terms of lexis, the concept of collocation has been used in the analysis of the intermediate level of language between syntax and lexis (the lexico-grammar, Halliday 1993). Recurrent word patterns have also been instrumental in recent developments in lexicography and the description of English, as in the Cobuild project (Sinclair 1987, Francis 1993). On the level of discourse, phraseology plays an important role in rhetorical choice, and idioms have been claimed to constitute important stages in the rhetorical development of texts (Moon 1992, McCarthy and Carter 1993). Another aspect of phraseology involves tracing the development of expressions within texts where deviation from the norm implies innovation and neology in the scientific community (Pavel 1993). This aspect of textual development touches on the concept of logogenesis which is the subject of ongoing research (Gledhill 1995 and forthcoming).

Biber and Finegan (1986) have been primary ex-

ponents of computer-based register analysis, an approach that measures variation in texts by the occurrence of linguistic features. They identify dimensions such as 'abstractness' and 'explicit information' that emerge from the co-occurrence of grammatical features such as clause complexes, *it*-clefts, adverbials, and more recently, lexical chains and deictic anaphora (Biber 1992). Kretzenbacher (1990) uses a similar methodology in his analysis of academic abstracts and articles. Essentially, this register-analysis approach maintains that text-types exist on a continuum, and that their differences can be explained by the analysis of internal linguistic properties of texts as opposed to external social and rhetorical features. Typically, register studies depend on computer corpora that are grammatically marked up or 'tagged'.

Genre analysis, on the other hand, works on a smaller scale of language than register, putting more emphasis on the specific context of a professionally or socially recognised discourse type. As such, genre analysis attempts to find patterns of conventional formulations which are accounted for by processes of use and production. Thus these formulations are not necessarily grammatical or cohesive features. Instead, they involve the norms of a community which shares certain values, and preferred textual structures expressed in rhetorical moves such as: 'establishing a territory', 'establishing a niche' and 'occupying the niche' (Swales 1990:141). An important difference is that while the register approach claims that a set of co-occurrent features has similar functions throughout the language, the genre approach assumes that grammatical forms have different functions for different discourse settings, and that within a genre a rhetorical move may be realised differently depending on what linguistic features can be adapted according to the practices of the discourse community. Swales (1990:42) therefore distinguishes between genre as a conventionally recognised instance of language in a discourse community, and register as the 'language of' a certain field, such as science or journalism.

While Atkinson (1992) has joined Biber's approach to the genre analysis of research articles, this paper attempts to contextualise corpus analysis, firstly by setting out the context of use of a particular genre and secondly by analysing a large corpus of the genre in collaboration with its users. In this paper we limit our analysis to typical writing strategies in cancer research abstracts. For this purpose, 150 papers (500,000 words) were collected with the collaboration of 15 expert informants from Aston University's Pharmaceutical Sciences Department. The papers were scanned by an electronic optical reader and placed on an IBM PC hard disk for automatic analysis. A decision was made to analyse only high frequency grammatical items

which were significantly more frequent in abstracts than in research articles. These items were then analysed for collocational properties using OUP's concordancer *Microconcord* (Johns-Scott 1991). The results suggest that when computational analysis of high frequency grammatical items is carried out with a view to taking discourse features and the context of production into account, the computational approach provides the genre analyst with a replicable, powerful tool of analysis.