THE USE OF SPECIALISED CORPORA: IMPLICATIONS FOR RESEARCH AND PEDAGOGY

Luz Gil Salom
Universidad Politécnica de Valencia

1. Introduction
Corpus-based genre analysis usually employ specialised corpora with the analytical objective of providing a linguistic description of a particular genre. The linguistic identification of a given type of discourse aims to offer insights into teaching a particular genre in the field of EAP/ESP. Corpora are invaluable resources that are needed to raise learners’ awareness of language and textual patterning of genres (Tribble 2002). And we might add, it would provide an incomparable motivating tool by making second language learning relevant to students’ disciplinary fields of study. Corpus-based genre analysis “pushes ESP toward empirically based understanding of language used for specific purposes” (Belcher 2006: 142).

The teaching of language for specific purposes has focused mainly on the analysis and use of authentic texts. At the Universidad Politécnica de Valencia, this has generally meant that teachers need to analyse engineering texts for pedagogical purposes. In this specific context, our research interests have been focused on the analysis of specialised corpora. Our first research project ACIA (Análisis de un Corpus de Inglés Académico), financed by the Universidad Politécnica de Valencia, aimed to analyse academic texts using software tools to obtain linguistic, rhetorical and genre patterns from the different disciplines that make up the academic curricula of the UPV. The project was initially involved in the creation and organisation of a corpus of academic technical and scientific texts written in English and its computerised treatment. The organisation and design of the corpus included the classification of the texts according to genre in: research articles, technical reports, abstracts, theses and conference papers, and also belonging to the most representative disciplines taught at the UPV: Agriculture, Biology, Electronics, Computer Science, Telecommunications, Civil Engineering, Industrial Engineering, Mechanical Engineering, Physics and Chemistry. The main reasons for choosing the journals in the corpus were that they are cited in the Science Citation Index (SCI®), they are read by university lecturers and students, and it is in these journals where lecturers and postgraduate students try to publish their research. The project also dealt with the design and development of language engineering products, such as TextWorks, a software application for text analysis, which includes a concordancer and a tagger. The programme we developed tried to complement a predominantly lexical approach within the existing software with a more grammatical focus. The programme allows the researcher to realise not only an analysis of texts at the word level but also at the lexico-grammatical level.

In the frame of our second research project RETEX: Redes Textuales: Análisis Intratextual e Intertextual del Discurso de una Comunidad Científica, financed by the Generalitat Valenciana, we created a specialised corpus of 21 PhD theses written in Spanish by students and teaching staff from the Universidad Politécnica de Valencia, all native speakers. The theses were published in an online library of theses and dissertations, ProQuest Information and Learning. According to its classification, the field is computing, but the departments involved are: Systems Engineering and Control,
Based on these two specialised corpora we developed different analyses for elucidating generic aspects of the text. The studies analysed (1) the evaluative patterning of RAs and technical reports; (2) the macrostructure of RAs and PhD thesis introductions and (3) the phraseology specific to certain sections of a RA, such as discussion sections. The aim of this review is then, to examine the research findings from two specialised corpora. Our research has an underlying pedagogical motivation as it aims to help Spanish undergraduates and academics write effective research documents in a way that meets the expectations of the international scientific community.

2. Evaluative patterning in RAs

The result of an analysis of 50 research articles and 15 technical reports revealed that, in computer science and electronics, evaluative language is an essential lexico-grammatical resource that carries important interpersonal meanings, including persuading the reader’s point of view and the validity of the research presented (Gil-Salom, Soler-Monreal & Stuart 2001). The study proposed a taxonomy of lexico-grammatical resources of evaluation with the idea that this could be useful for both academic reading and writing.

Another study describes some appraisal resources expressing evaluation, attitude, modality and certainty in the discussion sections of engineering texts. The research is based on a corpus of 46 RAs in the disciplines of telecommunications, computer science, nanotechnology and robotics. In discussion sections of academic papers the writer seeks to persuade the academic community of the validity of the research presented and the convenience of accepting new knowledge claims. Recent research on lexico-grammatical means of imposing attitudes, interpretations and assessment of truth value on readers has been carried out in specific disciplines (Thompson & Ye 1991, Stotesbury 2003, Koutsantoni, 2004).

Lexical, grammatical and discourse markers are presented with examples, in which we show how writers estimate the achievement of research goals and interact with the reader (Gil-Salom, & Soler-Monreal forthcoming). Due to the complexity and importance of RA discussion sections, we consider that the linguistic markers identified could be used in EAP reading/writing courses so as to familiarise future engineers with the rhetorical conventions that are considered adequate by the disciplinary community.

Effective writing depends on critical reading. This involves developing the ability to make judgements about how a text is argued. Arguments, claims, evidence and even the organisation of the text will be realised through evaluative language. Recognition of evaluative language will not only help the learner to read critically but also to avoid making mistakes such as showing lack of academic modesty through the use of overstatement. Expert academic writers will perceive both overstatement and understatement as strange. The novice academic needs tools to be able to come to grips with meanings which are not overtly stated but are built up through a logical argument which is often overlaid with evaluative discourse. Through critical reading, learners
become familiar with models of evaluative language, which will help them in their writing to master the discourse modes of their academic field.

3. Generic macrostructure

In the following two studies we investigated the corpus data following a top-down type of analysis. The starting point is with the macrostructure of the text with a focus on larger units of text rather than sentence-level lexico-grammatical patterning. In the first study we analysed the section headings and subheadings of 40 RAs in the fields of computing, robotics and telecommunications to identify their macrostructure. Our findings revealed that the standard IMRD model cannot account for all the specific choices regarding organisation and headings of individual RAs (Soler-Monreal, Gil-Salom, Carbonell-Olivares 2006a, 2006b). Scientific writers split the information into a variable number of sections and subsections and combine generic headings with unconventional ones.

Sections tell us how the writer sees the structure of his/her text. For this reason, the terminology employed in the title of each section should serve as a guide to the reader throughout the text. Writers combine generic and partially generic titles with topic-specific headings, particularly in the sections and subsections devoted to describing methods and presenting results or proposing models. Their purpose when organising information in their RAs seems to be to indicate not only communicative functions but also explicit elements involved in their research. One explanation to this may be that the reader does not usually read the text thoroughly unless he/she is especially interested in it. A way of helping him/her decide whether the contents of the RA will be of any use to him/her is to indicate clearly what every section is about, not only what its role is.

Our next study analysed the structure of a corpus of introductions drawn from 21 PhD theses written in Spanish on a variety of computing topics. The study was carried out under the move-step tradition initiated by Swales and his CARS model for RA introductions and revised by Bunton for PhD thesis introductions in English. Central to the notion of genre in the ESP domain is the move-structure analysis, which classifies segments of text according to their prototypical communicative purpose for a particular genre. The study shows that the construction of Spanish thesis introductions is complex due to the nature of the research topic and the various objects under study, and to the need the graduate writer feels to contextualise the research and explain the organisation of the PhD thesis (Carbonell-Olivares, Gil-Salom & Soler-Monreal in press). This explains the number of steps and sub-steps that describe the rhetorical functions of text segments in Spanish PhD thesis introductory chapters when establishing the territory (Move 1) and announcing the research carried out (Move 3). Since the academic community shares some research procedures irrespective of the country or the language used to disseminate research results, it is interesting to assess how well the model built on English texts works for texts in other languages. Our study shows that although most of the steps in Bunton’s model are present in the move structure of Spanish PhD thesis introductions, some new steps and sub-steps are also identified. By studying Spanish texts, we aim to contribute to the study of an academic genre that has not been the object of empirical analysis in Spanish studies.
We hope these findings will help Spanish graduate students raise awareness of the schematic structure of PhD thesis introductions. Research on actual texts provides realistic information that can be used for academic writing courses allowing graduate students to appreciate the complexity and variation that is involved in the process of writing PhD thesis introductions. When teaching their students a new genre, instructors can make them aware of features that are obligatory or optional in that genre, such as the moves and steps in the introductory chapters of Spanish PhD theses. Whether the conventions of the genre are typical of the Spanish academic community or the discipline of computing is likely to be answered after a contrastive study using a comparable corpus of PhD theses written in English, which will be the subject of further research.

4. Collocational patterning in RAs

A part of corpus-based research has centered on the exploration of lexical phrases (Sinclair 1991; Gledhill 2000a, 2000b; Stubbs 2002) and has presented language as a series of choices determined by the context in which it is employed. Native speakers use recurrent lexico-grammatical patterns when communicating in particular registers. This is especially relevant in scientific academic discourse, where the conventions of genres are interwoven with their linguistic realisations.

The study of collocational patterns has direct pedagogical applications. Learners are not usually taught collocations explicitly. However, we believe that the acquisition of phraseological competence is necessary for effective and precise communication. In the area of English language teaching, the works of Howarth (1993, 1996), Oakey (2002) and Tribble (1990, 2002) point out the importance of collocations in academic writing. Other studies propose the teaching of grammar prioritising the behaviour of individual lexical units (or pattern grammar), i.e. taking account of the lexical patterns of a given register (cf. Hunston 1995, 2002).

Collocations in native speaker academic writing have an effect on the precision of what is written. This can be expected to cause problems for non-native speakers of English, who are not exposed to large numbers of academic collocates and thus use few word combinations in their written work. Spanish researchers are often faced to the difficulty of finding some of the common patterns of lexical combinations used in formal, academic English when they present their research results in this language.

Collocations are of use to both the writer and the reader in terms of speed and ease of processing (Howarth 1996). The writer is able to express his/her meaning precisely while the reader is able to understand the phrase, as one lexical unit, requiring no additional ‘looking up’ of the item in the mental lexicon. Idioms do not cause problems in processing due to their idiosyncrasy, but rather aid the mental processes through their ‘familiarity and fixedness’. Such fixed phrases make use of humans’ large memory capacity by being stored as wholes, thus reducing computation demands and increasing fluency.

We developed two corpus-based studies of RAs in three engineering disciplines (computing, robotics and nanotechnology) to analyse collocational patterning. The aim of our first analysis was to explore the collocational patterns of three semitechnical and
specialised words used in a corpus of 54 engineering research articles in the fields of computing, robotics and nanotechnology: robot, performance and lattice. The analysis showed that, although these words can be found in general English, their collocates contribute to restrict and precise their meaning in a specialised corpus (Soler-Monreal, Gil-Salom & Carbonell-Olivares 2007).

The aim of the second study was to explore the collocational patterns of the noun result/s colligating with nouns, adjectives, prepositions and verbs in engineering English. The second aim was to illustrate different restrictive patterns when reporting and commenting on research results. In order to find out the collocates of result/s we have analysed a 500,000 word corpus of scientific RAs from leading journals. It has been processed by TextWorks (Gil et al. 2004), which has provided the concordance lines that have enabled us to identify the lexical items that usually collocate with the noun result/s in the following positions: L1 (the first position to the left of the node or search word), L2 (the position immediately to the left of L1), R1 (the first position to the right of the node) and R2 (the position immediately to the right of R1). Our data illustrates different restrictive patterns when reporting and commenting on results. Making learners aware of these patterns should arouse their consciousness of the use of language in specialised contexts and help them to improve their academic writing as regards accuracy and fluency.

Students' lack of familiarity with the phraseologies of a given genre distinguishes their writing from the professionals'. This results in unnatural and unidiomatic or inappropriate idiomatic expressions. To become an expert writer in ESP, students should be taught to use “appropriate lexico-grammatical patternings with consideration of various contextual and situational features of the discourse for the notions and functions one wishes to convey” (Flowerdew 2008: 133). This statement reinforces the argument that ESL learning may be enhanced by tailoring it to the students' disciplinary needs. Following up on that idea, Flowerdew argues for the importance of integrating corpus-based analysis into the development of textbook materials, and she suggests training students to 'appropriate' and 'authenticate' the corpus data for their own purposes.

The research findings of genre-based corpora have not influenced ESP materials or specialist dictionaries in the same way that the research findings from generalised corpora have heavily influenced general ELT pedagogic materials, dictionaries and grammars. Such ESP corpus-based materials tend to be used in localised contexts for particular groups of learners. We hope that the research findings from the specialised corpora will find wider applications in ESP materials, and will inform specialist dictionaries, so that they better reflect the actual language use of particular subject disciplines.

References


Gil-Salom, L., C. Soler-Monreal (forthcoming): “Appraisal resources in scientific research article discussions”.


