A Lexical Database of Collocations in Scientific English: Preliminary Considerations

Isabel Verdaguer, Eva González
University of Barcelona
Dept. Filologia Anglesa i Alemanya
Gran Via de les Corts Catalanes, 585
08007 BARCELONA
SPAIN
i.verdaguer@ub.edu
evagonzalez@ub.edu

Abstract
The aim of this paper is to present the theoretical principles underlying the making of a lexical database of English collocations of non-specialized words used in scientific language. This project was prompted by the shortage of reference tools providing information about the use and combinatorial properties of general words in specific registers. A case study will illustrate that in scientific texts, words, especially polysemous verbs, have a distinct semantic and combinatorial behaviour. Following the assumption that the meaning and the grammatical and collocational patterns of words are interrelated, we suggest that context-specific information should be included in specialized reference tools to facilitate the written production of scientific texts by non-native speakers of English.

1. Introduction: Collocational patterns in scientific discourse

The expansion of the international use of English in science and scholarship has contributed to the diffusion of scholarly work on discourse analysis (Swales 1990; Flowerdew 2002) and phraseology and collocations in academic English (Howarth 1996; Williams 1999; Gledhill 2000; Oakey 2002) - now mostly based on the analysis of corpora -, and to the publication of monolingual and bilingual/multilingual dictionaries of scientific terms. The purpose of these dictionaries is to provide information about the meaning of specialized words and their equivalents in other languages respectively. With a few exceptions[1], they do not include information about general words used in scientific register or references to grammatical and collocational patterns. The project which we present here, a lexical database of collocations in scientific research papers, was prompted by the shortage of reference tools giving information about the use and combinations of general words in scientific English.

Clarity and precision of expression is essential in academic writing and this is to a large extent achieved by the correct combination of terms. Scientists who are not native speakers of English usually know the meaning of scientific words, which are frequently morphologically and semantically similar to those of their own language, and have little difficulty in using them in their written discourse. However, in order to use a word correctly, the writer of scientific English has to know the grammatical and lexical patterns which are associated with a particular word. Thus, information about the way words typically combine in scientific discourse would help scientists to write research papers effectively.

2. Lexicology and corpus linguistics

In most branches of formal and functional linguistics there has developed a general trend towards lexically oriented approaches to language. The study of lexis has been dramatically transformed in that what used to be considered syntactic phenomena are now regarded as projections of lexical properties (Altenberg and Granger 2002). Thus, the lexicon, from which syntactic as well as semantic information can be projected, has acquired a relevance which it did not have in the past.

A linguistic approach which integrates syntax and semantics allows a systematic description of word meaning due to the interdependence between the meaning of a lexical item and the syntactic pattern in which it occurs. Modern linguistics (Sinclair 1991; Levin 1993) has shown that the different meanings of a word occur in different contextual patterns. Evidence from cognitive psychology (Lee 2001) and corpus studies (Sinclair 1991; Gledhill 2000; Oakey 2002; Stubbs 2002) has proved that pre-fabricated word strings are memorized as wholes and that native speakers frequently use recurrent word combinations in their linguistic production.

Technological advances have contributed to analyzing language from a new perspective; computers have allowed us to compile large quantities of authentic texts and describe actual language use. The analysis of large corpora with computer-assisted methods has made it possible to identify and describe these recurrent patterns of word co-occurrence known as collocations which are not observable through introspective analysis only.

3. A lexical database

A project consisting in the making of a lexical database of English collocations[2] of non-specialized words used in scientific language is currently carried out in the Spanish universities of Barcelona, Illes Balears and León. This lexical database (meant for Spanish speakers) gives information about the meanings and the grammatical and collocational patterns of general words used in scientific writing.

Due to the lack of publicly available corpora of scientific English - the Professional English Research Consortium (PERC) is developing a 100 million word corpus of Professional English, but it is still not publicly available - we have been compiling a 1 million+ word corpus of scientific research articles from prestige journals[3] from the areas of biology, biochemistry, and biomedicine. Only articles signed by at least one native speaker of English have been selected to guarantee the analysis of native competence.

The software used to process the data (WordSmith Tools) provides a list of concordance lines that makes it possible to identify the patterns of word co-occurrence and their frequencies. The figure below shows some concordance lines analyzed for the case study that we present here.

N	Concordance
13	P-2 that is essential fordevelopment or bone morphogenesis. of BMPs raise the possibility that each BMP molecule possesses distinct roles at d
14	and the unexpected finding of two PCS homologs in the C.elegans genome raise the possibility that phytochelatin synthase overexpressionmay be s
15	the phosphorylation of more than one oligosaccharide per polypeptide chain raise the question of whether phosphorylation is the result of a single or mul
16	ted by asterisks. Underlined residues indicate the peptide used to raise the FBA19 antiserum. To determine the expressionpatterns of G
17	s in RNA recombination, and their possible interaction with he TCV RdRp, raise the question of whether these hairpins play cis-acting roles in standa
18	H.-Y. Lin, un- published work).However, he present studies involving TAT3 raise the possibility that the activities of other cytokines or growthfactors th
19	rs [13], p21 ras [11] and G-proteins 16]. However, oncerns have been raised about the inferencethat caveolarlocalization can be assumed solely
20	tain non-specificantibodies against MTP or PDI. Middle panel: antibodies raised after immunization recognizedexpressedMTP and endogenousPDI
21	YK; mouse monoclonal 7E10 and rabbit polyclonals FBA30 and FBA31 all raised against recombinant GRASP65; a rabbit polyclonal antisera FBA32,
22	GRASP65; a rabbit polyclonal antisera FBA32, and a sheep antisera FBA34 raised against recombinant GRASP55; the H-7 monoclonal to the HA-epito
23	he complete cross-reactivity of both mutants with polyclonal rabbit antisera raised against the wild-type enzyme reported here indicates that the mutan
24	multisubunit complex in the chloroplast stroma Polyclonal antibodies were raised against a recombinant CRP1 fragment encompassing approximatel
25	umenal ER marker, BiP, wasdetected by immunoblotting using an antibody raised against the T.brucei BiP (Bangs et al., 1993). This antibody recogni
26	nst the rat GM130 from N.Nakamura and M.Lowe; rabbit polyclonal FBA19 raised against the peptide YLHRIPTQPSSQYK; mouse monoclonal 7E10
27	References Antibodies Crude serum and affinity-purified rabbit antibodies raised against a truncated Paired protein containing amino acids 355-613
28	ressed in bacteria, or synthetic peptides. One of hese antibodies, FBA19, raised against the sequence YLHRIPTQPSSQYK (underlined in Figure 1),
29	ical characterization by Western blotting showed eactivity with antibodies raised against human perlecan (Figure 4, lane 5). Electron micrographs cou
30	r has been demonstrated in human FF [23], the reactivity using antibodies raised against ITI was investigated. Western blot analysis showed that this
31	. Gottschling, Fred Hutchinson Cancer Center). Monoclonal antibodies were raised against glutaraldehyde cross-linked ubiquitin according to the proced
32	86 (which includes the part remaining in the Cul-3 knockout), and antibodies raised against the carboxy terminal end of Cul-3 (see Materials and Method
33	e columns used in the purification of B-type activity (p97/lunr) with antibodies raised against the recombinant proteins. For the coimmunoprecipitation t
34	not shown). In immunodiffusion experiments the polyclonal rabbit antiserum raised against wildtype amidase revealed structural differences between th

Figure 1: Concordance list

4. Case study: RAISE

In scientific register, words, especially polysemous verbs, have more restricted senses than in general language and characteristic grammatical and collocational patterns. However, the terms included in scientific dictionaries are predominantly specialized nouns with well-defined meanings; on the whole, information about the use of general open-class words in scientific texts is not included (see Norman 2002; L'Homme 2003).

Polysemous verbs used in specific registers usually have fewer senses and subsenses than in the general language (Pearson 1998). Also, they are determined by their linguistic environment and the collocates they appear with are specific too. This paper presents the results of the analysis of the different meanings and the grammatical and collocational patterns of the verb **raise** in a corpus of scientific English to justify the need to include context-specific information in specialized dictionaries.

Raise is a highly polysemous verb in the general language: The New Oxford Dictionary of English, which presents information based on the analysis of the British National Corpus and other corpora and citation databases, includes a total of 9 senses (with their respective subsenses) for this verb. The first sense is "lift or move to a higher position or level", which is also the first sense that other monolingual English dictionaries (Collins Cobuild English Dictionary, Cambridge International Dictionary of English) provide.

On the other hand, the analysis of the concordance lists of **raise** in our corpus of biochemical and biological sciences reveals a different behaviour of this word in this register, as the following table illustrates:

i	ii	iii	iv	v
raise: produce; cause to grow	*raise*: put forward for consideration	*raise*: cause to move to higher level	*raise*: grow	*raise*: rear
V NP (PP *against/in*) V NP (PP *in*) (PP *against*)	V NP	V NP	V NP (PP *at*)	V NP (PP *on*)
NP antibody antiserum	NP concern consideration doubt hypothesis implication question	NP components concentration level ratio proportion temperature	NP animal embryo	NP animal hyponyms
An antibody was raised in sheep against the X protein.	**Recent results raise the possibility that...**	*The concentration was raised to 1%.*	*All embryos were raised at 16°.*	*Flies were raised on cornmeal.*
50.7%	32.7%	10.5%	4.5%	1.5%

Table 1: Semantic, syntactic and collocational analysis of raise

As the table above shows, in our corpus of biological and biomedical sciences, there can be found five meanings of the verb **raise**. Cases i, iii, iv, and v correspond to its concrete uses, while case ii reflects its figurative use.

The most frequent meaning is "to produce, cause to grow" (case i), which shows a consistent collocation with the words antibody and antiserum and appears in the passive voice or as a past participle postmodifying these two nouns. Optionally, it collocates with prepositional phrases introduced by in and/or against in the right co-text.

Cases iv and v are closely connected both formally and from a semantic point of view. Similar collocational patterns indicate a close semantic relationship: in the sentences in which both are used, conditions that favour the growth and development of the participants (embryos and animals) are specified (temperature in case iv, and food in case v).

It is interesting to note that in this register **raise** as "cause to move to a higher level" (case iii) is not used as frequently as it is used in the general language. Its contextual patterning is more restricted also: in this context, **raise** does not connote physical movement upwards but it indicates an increase in measurable parameters or units (see its strong collocation with words such as level, proportion or temperature).

Case ii leads us to consider the issue of figurative meaning. Although here the meaning of **raise** derives from its primary, literal sense, it has lost all its physical connotations in a strong co-selection pattern with abstract nouns in the right co-text. Unlike the previous cases, most of the instances of this use are in the active voice and the noun phrases which realize the subject function have scientific research and its results as referents (eg. studies, results, findings, data). Tense is also a distinctive feature of case ii, as the

present tense is predominant while the use of the past tense is more widespread in the other four cases.

The figure below shows how the information for case i has been codified in our lexical database.

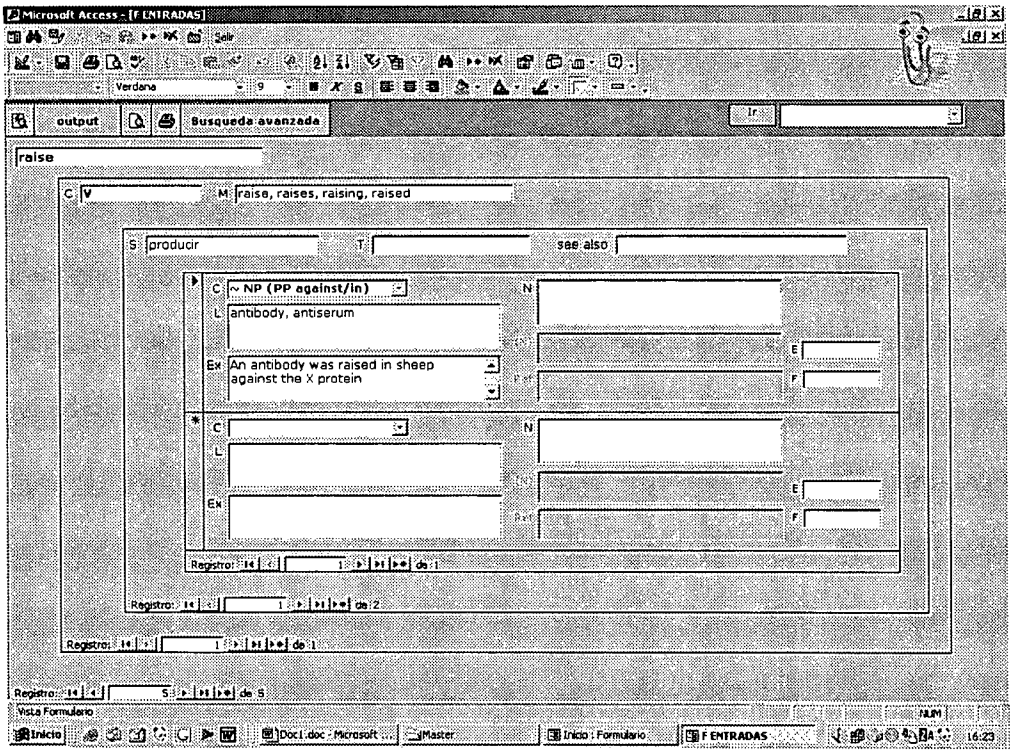

Figure 2: Database information for **raise** (case i)

5. Conclusions

Our analysis of the meanings and contextual patterning of the verb **raise** in scientific texts corroborates the assumption that form and meaning are systematically interrelated: variations in the contextual patterning of a word lead to changes in meaning.

When used in scientific texts, general words show specific collocational patterns and a more restricted set of senses. Thus, it proves relevant to include information about the combinatorial and semantic profiles of these lexical items in specialized reference tools aimed at non-native writers of research papers.

6. Acknowledgements

This project, reference BFF2001-2988, is financed by the Spanish Ministry of Science and Technology and FEDER.

7. Endnotes

1. See Williams' *Parasitic Plant Dictionary* at http://perso.wanadoo.fr/geoffrey.williams/

2. We use this term in a broad sense, meaning frequently co-occurring terms.

3. *Biochemical Journal, Genes and Development, British Medical Journal and The Journal of Cell Biology*

8. References

8.1 Dictionaries

Cambridge International Dictionary of English. 1995. Cambridge: Cambridge University Press.

Collins Cobuild English Dictionary. *1995. London: HarperCollins Publishers.*

The New Oxford Dictionary of English. 1998. Oxford: Oxford University Press.

8.2 Other references

Altenberg, B. and Granger, S. 2002. *Lexis in Contrast.* Amsterdam/Philadelphia: John Benjamins Publishing Company.

Firth, J. R. 1957. *Papers in Linguistics 1934-1951.* London: Oxford University Press.

Flowerdew, J. 2002. *Academic Discourse.* Edinburgh: Pearson Education Limited.

Gledhill, C. J. 2000. *Collocations in Science Writing.* Tübingen: Gunter Narr Verlag.

Howarth, P.H. 1996. *Phraseology in English Academic Writing.* Tübingen: Max Niemeyer Verlag.

L'Homme, M. C. 2003. 'Verbs and Verbal Derivatives. A Model for Specialized Lexicography'. *International Journal of Lexicography* 16.4: 403-22.

Lee, D. 2001. *Cognitive Linguistics: An Introduction.* Oxford: Oxford University Press.

Levin, B. 1993. *English Verb Classes and Alternations.* Chicago/London: The University of Chicago Press.

Norman, G. 2002. 'Description and Prescription in Dictionaries of Scientific Terms.' *International Journal of Lexicography* 15.4: 259-76.

Oakey, D. 2002. 'Formulaic Language in English Academic Writing' in Reppen, R. et al. (eds.), *Using Corpora to Explore Linguistic Variation.* Amsterdam/Philadelphia: John Benjamins Publishing Company. 111-29.

Pearson, J. 1998. *Terms in Context.* Amsterdam/Philadelphia: John Benjamins Publishing Company.

Sinclair, J. 1991. *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Stubbs, M. 2002. *Words and Phrases.* Oxford: Blackwell Publishing.

Swales, J.M. 1990. *Genre Analysis: English in Academic and Research Settings.* Cambridge: Cambridge University Press.

Williams, G. 1999. *Les Réseux Collocationnels dans la Construction et l'Exploitation d'un Corpus dans le Cadre d'une Communauté de Discours Scientifique.* Doctoral Thesis. http://perso.wanadoo.fr/geoffrey.williams.