SciE-Lex: A Lexical Database of Collocations in Scientific English for Spanish Scientists

Isabel Verdaguer Clavera
Anna Poch Higueras
Natalia Judith Laso Martín
Eva Giménez Domínguez
Universitat de Barcelona

As a result of the widespread use of English in science and scholarship, there is an increasing need of reference tools which provide accurate information to non-native-especially junior-researchers on the correct use of lexico-grammatical patterns of non-technical words when writing their scientific papers in English and on the conventionalized phraseological characteristics of the genre. Our aim is to present SciE-Lex, a lexical database which provides information to help Spanish researchers to write research papers in English accurately. Whereas there are specialized monolingual and bilingual dictionaries with specific terminological information, there is a shortage of reference tools supplying information on the correct use of syntactic and collocational patterns of non-technical words in the scientific register and on the conventionalized phraseological characteristics of the genre. Based on the analysis of a 3+ million word corpus of scientific English, in its first stage, SciE-Lex displays information on: word class, morphological variants, equivalent(s) in Spanish, patterns of occurrence, list of collocations, examples of real use, and notes to clarify usage. In a second stage we plan to include lexical bundles*, compositional recurrent sequences of words, since several studies have confirmed the difficulties that learners have with them. Further research will provide SciE-Lex with information about the distribution of lexical bundles across the different sections and/or moves of the academic research article as well as their function in discourse.*

Preliminary considerations

The role of English as the lingua franca of science and scholarship is undeniable. Scientists who wish international diffusion and recognition write their research articles in English (*Swales* 2004) and there is an increasing need of reference tools which provide accurate information to non-native—especially junior—researchers on the correct use of syntactic and collocational patterns of non-technical words when writing their scientific papers in English and on the conventionalized phraseological characteristics of the genre. The number of studies based on specialized corpora (Howarth 1998, Hyland 1998, Gledhill 2000, Luzón Marco 2000, Flowerdew 2002, Gavioli, 2005, Noguchi, Orr & Tono 2006, Verdaguer & Laso 2006, Biber & Barbieri 2007, among others) reflects the attention that they are currently receiving. Furthermore, the fact that the second edition of the Macmillan English dictionary for advanced learners (2007) includes a section to improve the writing skills of learners in academic or professional contexts confirms the existence of such a need.

Whereas there are specialized monolingual and bilingual dictionaries providing specific terminological information—either encyclopaedic or the equivalent in other languages—there is a shortage of reference tools providing information about the use of non-technical nouns and verbs in scientific register.

Aim

Our aim is to present *SciE-Lex*, a lexical database which provides information to help Spanish researchers to write research papers in English accurately. Based on the analysis of a 3+ million word corpus of scientific English, in its first stage *SciE-Lex* displays information on: *Word*

class, *Morphological variants*, *Equivalent(s) in Spanish*, *Patterns of occurrence*, *List of collocates*, *Examples of real use* and *Notes to clarify usage.*

Entry

The following lexical entry of the node word *evidence*, which is a frequent noun in research articles, is aimed at illustrating the contents of the database (see Figures 1, 2 and 3):

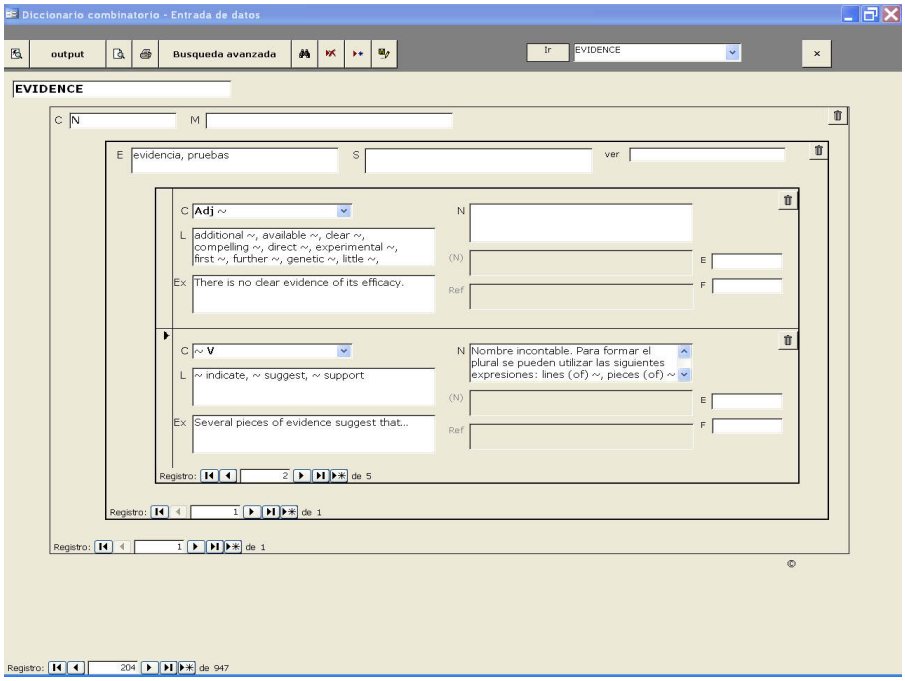

Figure 1. Lexical entry of the node word *evidence* (I)

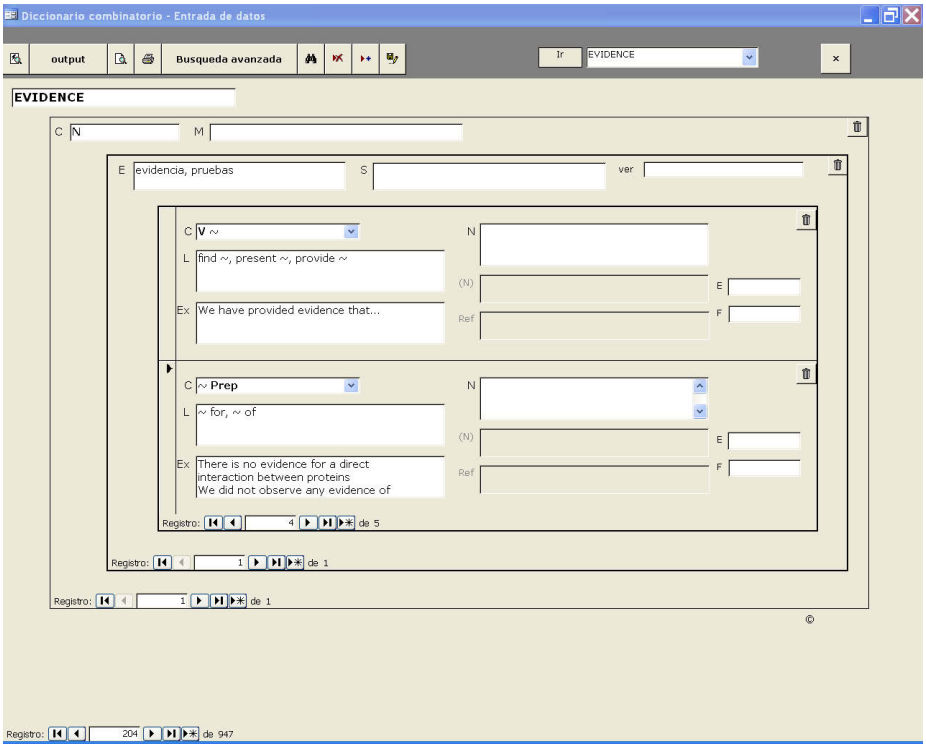

Figure 2. Lexical entry of the node word *evidence* (II)

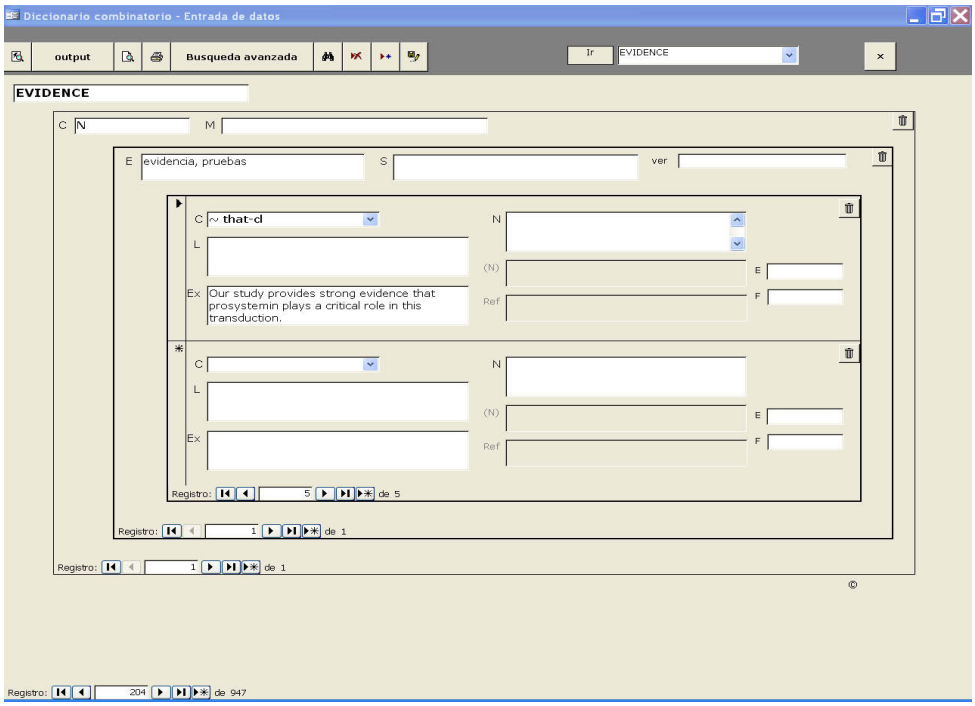

Figure 3. Lexical entry of the node word *evidence* (III)

Word Class (C): it is a noun

Morphological variants (M): it is uncountable (so, it is invariable)

Equivalent in Spanish (E): (*evidencia*, *pruebas*), which can be further specified if necessary, especially in cases of homonymy or polysemy.

Next, there are the *patterns (C)* in which the lexical item can appear, followed by a list of the most frequent *collocates (L)* and *examples (Ex)* illustrating them:

evidence can be preceded by an adjective (*additional*, *available*…) (see Figure 1)

evidence can be the Subject and, consequently, can precede the verbs *indicate*, *suggest*, *support*… (see Figure 2)

evidence can be the Object and, consequently, can follow the verbs *find*, *present*, *provide*… (see Figure 2)

evidence can be followed by a *for*-PP or *of*-PP (see Figure 3)

evidence can be followed by a *that*-clause (see Figure 3)

Finally, there may be *notes* that help users or highlight special usages (see Figure 1)

Future work: Lexical bundles

As Gledhill (2000: 130-131) points out, phraseology is a central issue in ESP studies: "the direct correspondence between lexis and grammar is now so pervasive that it is difficult to conceive of a general characterization of science writing or the design of teaching materials for the benefit of science writers which can afford to ignore phraseology as a central level of analysis". In addition, scientists are not consciously aware that there are phraseological conventions specific to each genre.

In a second stage, then, in line with the current recognition of the importance of phraseology (Oakey 2002, Charles 2006), we plan to include the multiword units that Biber et al. (1999 & 2004) refer to as *lexical bundles*. Several studies on learner corpora (Cortes 2004, Granger 2005, Nesselhauf 2005, Paquot 2005) have confirmed the difficulties that learners have with these compositional recurrent sequences of words. After subdividing the corpus into several subcorpora according to the different sections in some research papers compiled in the *SciE-Lex* corpus, we will provide the following information:

Most statistically frequent lexical bundles

Patterns and its variants:

1) THERE + BE + evidence + that-clause
   THERE + BE + evidence + for-PP
   THERE + BE + evidence + of-PP

There + is/was/isn't/wasn't + (some/no // clear/direct/further/strong) + evidence + that-clause/for-PP/of-PP

2) evidence + FOR + NP [the absence/presence of, the/a role of]
3) lines of evidence + VP[indicate/suggest/support]
4) a (large) body of + evidence
5) in the light of + evidence

Further research will provide *SciE-Lex* with information as regards the distribution of the abovementioned lexical bundles across the different sections and/or moves characteristic of the academic research article as well as their function in the discourse.

In conclusion

SciE-Lex supplies not only lexico-grammatical information of the language used in the scientific research article but also aims at helping scientists in the phraseological conventions of the genre.

Acknowledgements

This is part of a project financed by the Spanish Ministry of Science and Education and FEDER (HUM2007-64332/FILO).

References

Biber, D. et al. (1999). *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Biber, D. (2004). "Lexical bundles in academic speech and writing". In Lewandowska-Tomaszcyk, B. (ed.). *Practical Applications in Languages and Computers*. Frankfurt: Peter Lang, 165-178.

Charles, W. (2006). "Phraseological patterns in reporting clauses used in citation: A corpus-based study of theses in two disciplines". *English for Specific Purposes* 25 (3). 310-331.

Cortes, V. (2004). "Lexical bundles in published and student disciplinary writing: Examples from history and biology". *English for Specific Purposes* 23 (4). 397-423.

Flowerdew, L. (2002). "Corpus-based analyses in EAP". In Flowerdew, J. (ed.). *Academic discourse*. London: Longman. 95-114.

Gavioli, L. (2005). *Exploring Corpora for ESP Learning*. Amsterdam: John Benjamins.

Gledhill, C. J. (2000). "The discourse function of collocation in research article introductions". *English for Specific Purposes* 19. 115-135.

Granger, S. (2005). "Pushing back the limits of phraseology: How far can we go?". In Cosme, C.; Gouverneur, C.; Meunier, F.; Paquot, M. (eds.). *Proceedings of the Phraseology 2005 Conference*. Louvain-la-Neuve, 13-15 October 2005. 165-168.

Howarth, P. (1998). "The phraseology of learners' academic writing". In Cowie, A. P. (ed.). *Phraseology: Theory, Analysis, and Applications*. Oxford: Oxford University Press. 161-186.

Hyland, K. (1998). *Hedging in scientific research articles*. Amsterdam: John Benjamins.

Luzón Marco, M. J. (2000). "Collocational frameworks in medical research papers: a genre-based study". *English for Specific Purposes* 19. 63-86.

Macmillan English Dictionary for Advanced Learners. Oxford: Macmillan, 2007.

Nesselhauf, N. (2005). *Collocations in a Learner Corpus*. Amsterdam: John Benjamins.

Noguchi, J.; Orr, T.; Tono, Y. (2006). "Using a dedicated corpus to identify features of professional English usage: What do *we* do in science journal articles?". In Wilson, A.; Archer, D.; Rayson, P. (eds.). *Corpus Linguistics Around the World*. Amsterdam: Rodopi. 155-166.

Oakey, D. (2002). "Formulaic language in English academic writing". In Reppen, R. et al. (eds.). *Using corpora to explore linguistic variation*. Amsterdam: John Benjamins. 111-129.

Paquot, M. (2005). "EAP vocabulary in learner corpora: a cross-linguistic perspective". In Cosme, C.; Gouverneur, C.; Meunier, F.; Paquot, M. (eds.). *Proceedings of the Phraseology 2005 Conference*. Louvain-la-Neuve, 13-15 October 2005. 323-326.

Swales, J. (2004). *Research Genres*. Cambridge: Cambridge University Press.

Verdaguer, I.; Laso, N. J. (2006). "Delexicalisation in a Corpus of Scientific English". In Hornero, A.; Luzón, M. J.; Murillo, S. (eds.). *Corpus Linguistics*. Bern: Peter Lang. 417-428.

Wray, A. (2002). *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.

Wray, A. (2005). "Looking at the WHY in Phraseology: a psycholinguistic perspective on patterns in text". In Cosme, C.; Gouverneur, C.; Meunier, F.; Paquot, M. (eds.). *Proceedings of the Phraseology 2005 Conference*. Louvain-la-Neuve, 13-15 October 2005. 23-26.