Corpus-Driven Lexicography and the Specialised Dictionary: Headword Extraction for the Parasitic Plant Research Dictionary.

Geoffrey Clive Williams
U.F.R Lettres et Sciences Humaines
Université de Bretagne Sud
4 rue Jean Zay
B.P. 92116
56321 LORIENT CEDEX
France
Geoffrey.Williams@univ-ubs.fr

Abstract

Specialised dictionaries tend to address the needs of language-aware users such as translators, or to be terminological seeking only to standardise usage. This paper looks at some of the problems encountered in using a corpus-directed approach for the design of a specialised pedagogical dictionary. Collocational networks are demonstrated as a objective means of headword extraction. In network building, collocations are seen as statistically significant pairings extracted from within a pre-defined window. The resulting networks represent the thematic environment of the corpus and include both technical and semi-technical words. The advantages and drawbacks of these networks in compiling a specialised pedagogical dictionary are discussed.

Introduction

In recent years there has been a great deal of activity in the world of pedagogical dictionaries, especially for English [Rundell, 1998]. With an increasing emphasis being placed on encoding, these dictionaries address the needs of learners, particularly those studying languages. Unfortunately, students following non-language courses with a Languages for Specific Purposes component tend to prefer bilingual dictionaries, if they use one at all, whilst failing to recognise the pitfalls in encoding. However, science students entering the world of academic research are rapidly faced with the need to publish. Although they generally acquire field-specific terminology quickly, the problem of expressing this in coherent English remains a problem, both for the experienced and inexperienced researcher. When specialised dictionaries are to be found they do not necessarily help in that these generally tend to be terminological in nature and mainly address translators who have little problem with the co-text. Academics can rarely afford the services of professional translators and are left to fend for themselves.

This paper looks at some of the problems encountered in designing an encoding/decoding dictionary for a multidisciplinary discourse community, that of parasitic plant biology research.. In building a specialised dictionary a number of problems arise; defining the community to be served, choosing headwords and defining both subject or field specific and essential general words. This article looks at the research community under study and discusses the problems of extracting a headwords. Collocational networks are demonstrated

as a corpus-driven approach to the extraction of headwords. The advantages and drawbacks of these networks are discussed.

Defining the community

The first factor, defining the community, is essential in a corpus driven approach if a suitably representative corpus is to be achieved. One of the disadvantages of adopting disciplines as the basis of a dictionary is that these are merely convenient categorisations of human knowledge; the reality tends to be multidisciplinary. For a small lexicographical project it would seem easier to begin with a user community centred round a definable research topic. The discourse community (DC) as the basis of English for Academic Research (EAP) has been defined by Swales [1990] as a form of special interest group. The advantage of this approach in special language lexicography is the sense of belonging; there is a known, definable user community with whom the lexicographer can interact in adapting the dictionary to their needs. In the present case Swales' criteria have been refined so as to take into account the particularities of scientific research through the Scientific Discourse Community, or SDC [Williams, 2001a]

In this research the SDC under study is that of parasitic plant research. Up to recently this has been an *ad hoc* community, but has now taken the form of a learned society, the International Parasitic Plant Society (IPPS). Members of the community come from a wide variety of both research and linguistic backgrounds. This domain calls upon a number of biological disciplines ranging from field-based research in agronomy to laboratory-based work in cell biology and biochemistry. The *lingua franca* of the community is English, which entails that both full members and apprentices are faced with two non-scientific problems: producing correct English within the confines of the research article and the correct use of terminology in a multidisciplinary environment. In addition members of the community have to reconcile their field-specific usage with the need to address experts from other disciplines within a topic-based community, a potentially complicating factor in reconciling usage and terminological norms.

The BIVEG Corpus

The corpus used to investigate this community and as the basis of the Parasitic Plant Research Dictionary (PPRD) is the BIVEG, *Biologie Végétale*, corpus, a 560,000-word corpus built from 279 published research papers. This corpus was originally started as part of a teaching project in a French university, hence the name. In the initial project other corpora dealing with other aspects of biology were envisaged, these were abandoned in favour of carrying out a deeper analysis in the area of plant biology. The corpus was further refined so as to concentrate on parasitic plant biology, but with papers from wider aspects of plant biology being retained a part of the global cohesion needed for corpus analysis [Williams, 1999]. The papers included in the corpus take two forms; articles destined to the wider scientific community, generally field-specific DCs such as molecular biology or plant physiology and which have been published in peer-reviewed journals, and theme-specific conference proceedings, addressed primarily to other members of the topic-based DC. Potential headwords for the encoding dictionary will be exclusively taken from this corpus, which raises the difficulty as to which words are the most salient..

Headword extraction and collocational networks

The recent corpus-based dictionaries have been forced to go back to the sources so as to reflect only what can be attested in the corpus and to eliminate entries that are simply the result of years of accumulation. However, going back to basics raises the question as to which words to include. For the COBUILD [Sinclair, 1987] frequency has been the principle criteria, but whilst this is feasible in a general language pedagogical dictionary, it is not necessarily so in a specialised one.

Collocation networks [Williams, 1998; 2001a] were developed within a corpus driven perspective [Tognini Bonelli, 2001] as a means for objective extraction of a lexis by calling upon statistically significant co-occurrence patterns in text. The idea is simple. If high frequency lexical items can be seen as significant within a corpus, or subset of that corpus, then they can serve as nodes for the extraction of statistical collocations.

Collocational networks are based on a textual, rather than a lexicographic, approach to collocation [Williams, 2001b]. Collocation is seen as the regular co-occurrence of two or more lexical items [Sinclair, 1991], which here means that other linguistic parameters are temporarily laid to one side. The degree of collocation is measured statistically with clusters of collocates being seen as demonstrating the central themes in a text. This approach is not new, local networks were exploited by Berry-Roghe [1973] on literary texts and by Phillips [1985], who, working on scientific texts, developed the concept of "aboutness". Furthermore this contextualist approach to collocation has been discussed by Clear [1994] as a means of disambiguating between polysemic forms and is also now a central characteristic of word sense disambiguation programmes as Word Sketch (Kilgarriff & Tugwell, 2001].

Rather than viewing collocation as a dominance relationship between base and collocate, collocational networks are formed by considering each collocate as a node in its own right. The collocates of a node are extracted using the mutual information score [Church & Hanks, 1990; Church et al. 1994]. The drawbacks of mutual information as opposed to other statistical techniques are well known [Clear, 1993; Kilgarriff, 2001], however, the principle defect, that of emphasising rare or more technical words has been found to be ideal for the extraction of salient specialised items as potential headwords [Williams, 2001a]. Unlike previous lexical networks, collocational networks are not purely local, instead each unit in a collocational pair as a node a network is allowed to spread out naturally from a central start node, often a high frequency lexical item. This process usually comes to a natural end at about 5 removes from the start node. The networks can best be illustrated with an example.

In any corpus subset concerning molecular biology, *DNA* inevitably appears as a high frequency unit. Figure 1 shows the immediate collocates of *DNA* in the molecular biology subset of the BIVEG corpus. As can be seen, all the collocates clearly relate to the discipline, two belonging specifically to plant biology. All ten collocates could be useful headwords in their own right, six nouns, three adjectives and one verb. The left collocates can be seen as forming terms relating to different forms of *DNA*, to the right *sequences* (the product), *sequencing* (the process) and *methylation* are also terms relating to *DNA* and are worthy of a full entry. *Genome* is a contextual collocate, that is one occurring within a text window, but not forming a contiguous collocational pair. *DNA* and *digested* form a classic noun-verb

collocation. However, rather than stopping here we can treat each of these collocates as a node thereby seeking their collocates and creating a network of related words. This gives a much more complex network as can be seen in figure 2. Items that appear more than once indicate circularity and would not be explored further. It does not require an extensive knowledge of plant molecular biology to see that all the forms represented clearly relate to the topic under study.

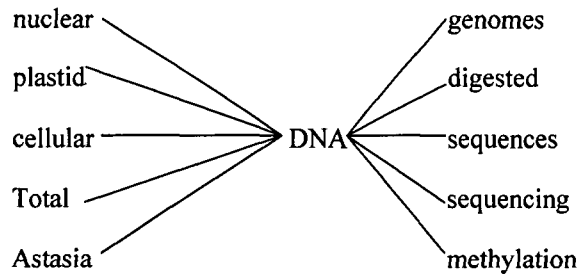


Fig. 1. Immediate statistical collocates of *DNA*

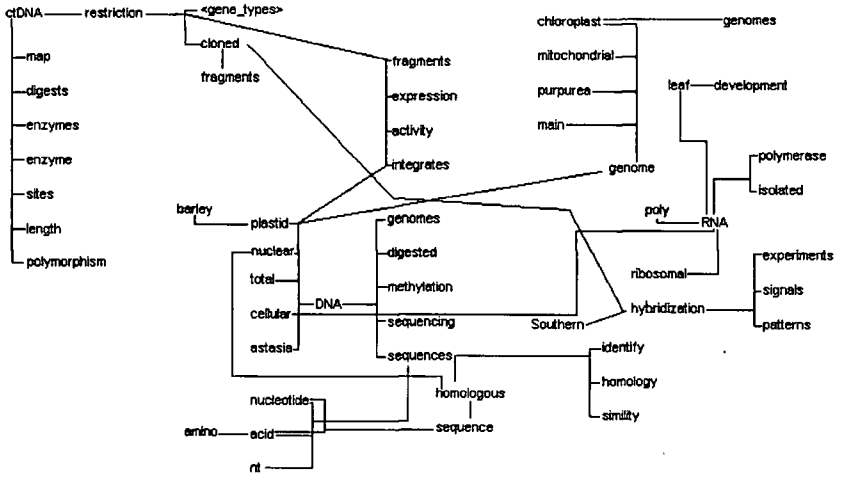
The extraction process is not lemmatised and is carried out on raw text. This is deliberate as the patterns from individual forms are in themselves significant in a data-driven perspective and because the annotation of texts does not necessarily increase the analytical yield, but does distance the corpus linguist from the text itself (Tognini Bonelli, 2001). Obviously, in the dictionary construction stage, word paradigms and the classification of entries into lexemes are taken into account, but this is a later process at the network stage the forms are seen only as entry points, lexical hooks on which entries may be hung.

Building networks in this way does not produce a terminology, the results mix both scientific and non-scientific words; only function words are eliminated by means of a stop list although MI tends to eliminate these anyway. What must be borne in mind is that collocational networks are textual in nature and demonstrate surface relationships between words in context. The aim here is to isolate significant items used within a community as potential headwords in a specialised pedagogical dictionary, which means going beyond the purely technical usage to show wider contextual words that are just as important, if not more so, to the potential dictionary user considered here.

As with any automated system, the two major dangers are silence and noise. Both are being tackled by working closely with the user community, cross checking proposed headwords and word lists compiled from the corpus. To date silence has largely been the result of unattested terms, inevitable in any fast moving discipline. Noise is a question of perspective. In its encoding function, the dictionary primarily addresses non-native members of the community who need to see examples of terms in context and require an increased semi-technical vocabulary. On the other hand, native speakers and experienced scientists are less interested in "non-scientific" words, but are more concerned by a prescriptive standardisation of terminology. This inevitably leads to a conflict of interest, and a problem of methodology.

Conclusion

This paper is an overview of a small project that opens big questions. It must be borne in mind that both the dictionary and the solutions are experimental. Collocational networks do allow the extraction of relevant headwords for a dictionary that aims to include significant non-scientific words. For a specialised dictionary, or terminology, some form of filter would be necessary. Whilst a corpus-driven approach may work well for extraction from a specialised corpus, it works less well for defining purposes, but as Teubert [2001] has pointed out, in a production-oriented dictionary definitions may not even be necessary. For more technical usage, the corpus will serve as a source of examples and as a basis for negotiating 'correct' usage. A discourse community is a living thing, studied over time it will be possible to see whether the definitions resist or will need adapting, the diachronic aspect can only increase in significance.



Figur 2:. A collocational network for *DNA*

References

[Berry-Roghe 1973] Berry-Roghe G.L.M. 1973. The Computation of Collocations and their Relevance in Lexical Studies, in: Aitken A.J., Bailey R., Hamilton-Smith N. (eds) *The Computer and Literary Studies*. Edinburgh : Edinburgh University Press.

[Church & Hanks 1990] Church, K., & P. HANKS, 1990. Word Association Norms, Mutual Information, and Lexicography, in: *Computational Linguistics* 16 (1). pp 22-29. MIT Press.

[Church et al. 1994] Church, K., W. Gale, P. Hanks, D. Hindle & R. Moon. 1994. Lexical Substitutability, in: Atkins, B.T.S. & A. Zampolli, A. (eds.) *Computational Approaches to the Lexicon.* pp. 153-177. Clarendon Press, Oxford.
[Clear 1994] Clear, J. 1994. I can't see the sense in a large corpus, in: *Papers in Computational Lexicography. Complex '94.* pp 33-48. Hungary: Budapest.
[Kilgarrif 2001] Kilgarriff, A. 2001. in *International Journal of Corpus Linguistics.* Vol3 (1). pp 151-171. John Benjamin's, Amsterdam.
[Kilgarriff & Tugwell 2001] Kilgarriff, A. & D. Tugwell. 2001. Word Sketch: Extraction, Combination and Display of Significant Collocation for Lexicography, in: *Proceedings of ACL Workshop on Collocation: Computational Extraction, Analysis and Exploitation.* Toulouse July 7th 2001: 32-38. Université de Toulouse, France.
[Phillips 1985] Phillips, M. 1985. *Aspects of Text Structure: An investigation of the lexical organisation of text.* Amsterdam, North Holland.
[Rundell 1998] Rundell, M. 1998. Recent Trends in English Pedagogical Lexicography in *International Journal of Lexicography* 11 (4) pp 315-342. Oxford University Press, Oxford.
[Sinclair 1987] Sinclair J. (ed) 1987. *Looking Up : An Account of the COBUILD Project in Lexical Computing.* London : Collins.
[Sinclair 1991] Sinclair, J. 1991. Corpus, Concordance, Collocation. Oxford: Oxford University Press.
[Swales 1990] Swales, J. 1990. *Genre Analysis.* Cambridge University Press, Cambridge.
[Teubert 2001] Teubert, W. 2001. Corpus Linguistics and Lexicography in *International Journal of Corpus Linguistics.* Special Issue. pp 125-153. John Benjamin's, Amsterdam.
[Tognini Bonelli 2001] Tognini Bonelli, E. 2001. *Corpus Linguistics at Work.* Amsterdam: John Benjamin.
[Williams 1998] Williams, G. 1998. Collocational Networks : Interlocking Patterns of Lexis in a Corpus of Plant Biology Research Articles, in: *International Journal of Corpus Linguistics.* Vol3 (1). pp 151-171. John Benjamin's, Amsterdam.
[Williams 1999] Williams, G. 1999. Looking in before looking out: Internal selection criteria in a corpus of plant biology, in: *Papers in Computational Lexicography. Complex '99.* pp 195-204. Hungary: Budapest.
[Williams 2001a] Williams, G. 2001. *Les réseaux collocationnels dans la construction et l'exploitation d'un corpus dans le cadre d'une communauté de discours scientifique.* Presses Universitaires de Septentrion, Lille.
[Williams 2001b] Williams, G. 2001. *Sur les caractéristiques de la collocation,*.in : Actes du 8ème Conférence sur le Traitement Automatique des Langues Naturelles. Tome 2. Université de Tours, France.