Going organic: Building an experimental bottom-up dictionary of verbs in science

Geoffrey Williams and Chrystel Millon
HCTI-LiCoRN, Université Européenne de Bretagne

Choosing what headwords to enter in a dictionary has always been a major question in lexicographical practice. Corpora have greatly helped ease both the choice of words to add, and those to remove, by resorting to frequency counts so as to monitor usage over time. This has been particular valuable in the building of learners dictionaries as, however good earlier word lists may have been, they were built largely in intuition whereas, corpora allow the consultation of large reference corpora for a better picture of current realities. In specialised dictionaries dealing with terminological issues, pure frequency is not a feasible solution for headword extraction. However, linked with extraction patterns and statistical tools, corpora still play a major role in supplying information on terms in use. In this research we aim to tackle a situation that lies in between the needs of an advanced learners dictionary and those of a terminological dictionary in attempting to build a pattern dictionary for verbs used in scientific research papers. In order to select verbs for this dictionary and put them into classes, we propose to use collocational relationships as a tool for both selection and analysis of patterns. The principle here is that a series of high frequency verbs can provide the seeds from which prototypical patterns can be extracted. By moving backwards and forwards from verb to argument and back pattern are revealed that use the statistical selectionning to highlight verbs lower in the frequency list that would otherwise be overlooked. Thus patterns will naturally enlarge the word list by selecting what is statistically significant with a textual environment. These patterns not only illustrate typical usage in a specialised environment, but will also group verbs according to textual functions as authorial positioning and description of processes.

1. Verbs and science in learners dictionaries

This project has grown from research into the dictionary as a tool for ongoing learning for practising scientists. The starting point has been research into existing advanced learners dictionaries as a tool for non-native speaker scientists, both beginner or confirmed, who need to publish in English. Research (Williams 2006, 2008a) has found that although some learner's dictionaries do highlight words that are deemed to be significant in the sciences, the target audience is not scientists but more students studying English. In addition to this, analysis revealed a great deal of inconsistency in coverage. Consequently, a first study endeavoured to show how lexicographical prototypes (Hanks 1994, 2000) could be adopted so as to expand existing definitions and show their relevance to more specialised usage. Such an approach does not get over the need to demonstrate the specialised usage patterns, which is why a complementary dictionary, in this case on-line, could fill the gap showing not only patterns commonly associated by certain verbs, but also where and why they are used.

This research has been developed as part of the SCIENTEXT initiative, a project funded by the French national research agency, which seeks to build corpora for the investigation of authorial positioning and evaluation. Three main strands have been developed based on the specificities of the research team involved, with work being carried out on a corpus of French texts in Grenoble, on learner corpora in Chambéry and on English corpora in Lorient. In the definition adopted by Grenoble, 'scientific' can be taken to mean academic publications in general, whereas the Lorient team has been essentially concentrating on the experimental sciences. Particularly specialised in lexicography for specialised languages, the LiCoRN research group (Lorient) has been primarily concerned with seeing how stance and reasoning can be captured and demonstrated in a dictionary. The English corpus developed in Lorient covers different aspects of academic usage. For the dictionary project described here we concentrate on the BMC corpus of scientific English.

2. The BioMed Central (BMC) Corpus

The English BioMed Corpus (BMC) is composed of 8945 scientific texts from 137 journals made freely accessible on-line by BioMed Central, an independent publishing house. The period covered by the corpus is from 1997 to 2005. These texts, and the corpus derived from them, are available under a creative commons licence.

The current corpus contains 33 million words that have been automatically part-of-speech tagged and lemmatised with Treetagger. The whole has been structured in conformity with the XML-TEI P5 with topic and genre-related information. Both the topic(s) (*Biochemistry*, *Blood disorders*, *Cancer*, *Cell biology*, ...) and genre (*survey*, *debate*, *database*, *report*, *short paper*, …) are encoded in the TEI header. Data concerning topic has been added automatically to the corpus, since they are absent in the original XML tagging. These topics have been extracted using the topical classification of the journals on the BMC Central Website. Data relating to genre was already included in the original XML tagging. In terms of frequency, the main topics in the corpus are: *Medical Genomics*, *Genomics*, *Bioinformatics*, *Genetics*, and *Women's Health*. Of these the first three dominate meaning that the topic distribution of the BMC Corpus is skewed in favour of a small number of topics that are particularly concerned with biomolecular analyses.

This distribution is a serious drawback for genre studies, but is not necessarily a problem for research that aims essentially at an encoding dictionary for NNS endeavouring to produce research papers in English. The bias towards biomedical language will later be measured against broader corpora of scientific data.

3. Networks in headword selection

Our working methodology is based on the use of collocational networks (Williams 1998), a methodology used to demonstrate thematic patterns in text, as well as means for selecting the lexis for a specialised language dictionary (Williams & Millon 2009). In addition to standard collocational networks we are also using selection of candidate collocates by patterns (MILLON) and collocational resonance to analyse meaning shifts (Williams 2008b). This view is not that of phraseologically related collocation that is to be found in dictionaries, but a neo-Firthian contextualist view whereby meaning is created through relations, both grammatical and lexical, within a textual environment. Sinclair (1991) amply demonstrated that the corpus, if built following carefully selected criteria, provides the environment which both typifies word meaning and provides a resource from which relevant lexis can be extracted. Collocational networks provide a means to extract that lexis.

The principle behind collocational networks is simple, words cluster. All language teachers have tested this by brainstorming so as to built a textual environment of related words. Simple statistics can perform this task by extracting collocationally related words from a corpus. Networks are statistically based chains of collocations that are built by using statistical tools to measure significant co-occurrence within a KWIC window. From a single high frequency lexical unit, lemma or group of lemmas the collocates are calculated using a statistical measure. The collocates of each collocate is then calculated in the same way leading to widening networks with each new element being entered in the lexis and subsequent dictionary. Previous work on specialised vocabulary used mutual information applied to raw data, but in this case we are using z-score applied to a POS marked-up corpus so as to extract only patterns involving particular parts of speech, notably nouns, verbs and adjectives in the current stage. Our studies have found that in looking at general and semi-technical usage in a specialised corpus, z-scores give a better picture of the environment of a search word without going for the rarer terminological items.

These thematic collocational networks have been adapted for the analysis of language change by studying how elements of meaning are carried over from one textual environment to another. This new approach to collocation has been termed collocational resonance (Williams 2008b). The methodology grew out of work on intertextual patterning presented at the Louvain 2005 phraseology conference and has been widened following work on resonance in metaphor by Patrick

Hanks that was also presented in Louvain. Hank’s work on prototypes in lexicography is central to the model that is currently being built.

The methodologies being developed are seen as building stones in the new outlook on language that we owe to John Sinclair’s insights into collocation and the idiom principle. The main influences have therefore been the Birmingham school, comprising both work carried out at the University of Birmingham and at Aston University, with the foundations of corpus analysis of scientific text in the work of Roe (1977) and later studies of the phraseology of scientific text by Gledhill (2000). Hence this list of main influences goes from Sinclair’s demonstration of the idiom principle and its ramifications seen through the work of, for example, Hunston & Francis on pattern grammars, Louw (1983, 2000) on semantic prosody and the ongoing work of Hoey from the Surface of Discourse (1983) to Lexical Priming (2005). By ultimately bringing in Corpus pattern Analysis, we believe a new approach to special language lexicography can be developed.

In this study, our start nodes are the first 100 verbs lemmas in the BMC corpus. Patterns and networks are first extracted for each of the verbs. The patterns for each verb are entered in the dictionary database which is built using Tshwanelex. The next stage is to further explore the network to locate other verbs associated with the patterns and to bring common patterns together to form pragmatically oriented classes. The latter, like the word list, are supposed to develop organically from the corpus data,. Existing classification as such as Levin (1983), Framenet or the Brandeis Ontology are also being consulted as a source of inspiration in naming classes. However, even here, class naming will prefer broad classes that are immediately understandable to a NNS user. The process can be illustrated through a case study of the verb *show.*

4. Case Study: Show

The size of the corpus means that the most frequent verbs tend to have very high frequencies (table 1) . In this first sweep we are not looking at either *be* or *have* or the modal verbs. Of the remainder, the lemma *show* is the most frequent with over 67000 occurrences, followed by *include* with 40000. Given their high frequency, the potential collocates of these verbs is extremely high, as a consequence the patterns demonstrated here will be restricted to only the most frequent collocates.

Verb	Nb occurrences
show	67287
include	39912
find	37481
increase	35422
compare	31200
follow	29449
suggest	29274
report	28546
identify	27990
provide	26831

Table 1. The ten most frequent verbs in the BMC corpus

As Sinclair (1991) has shown, different forms of word will show very different patterns which in turn can illustrate different meanings. In this study, the starting point is the lemma. Systemic Functional Grammar is then applied to see how a word is used in context.

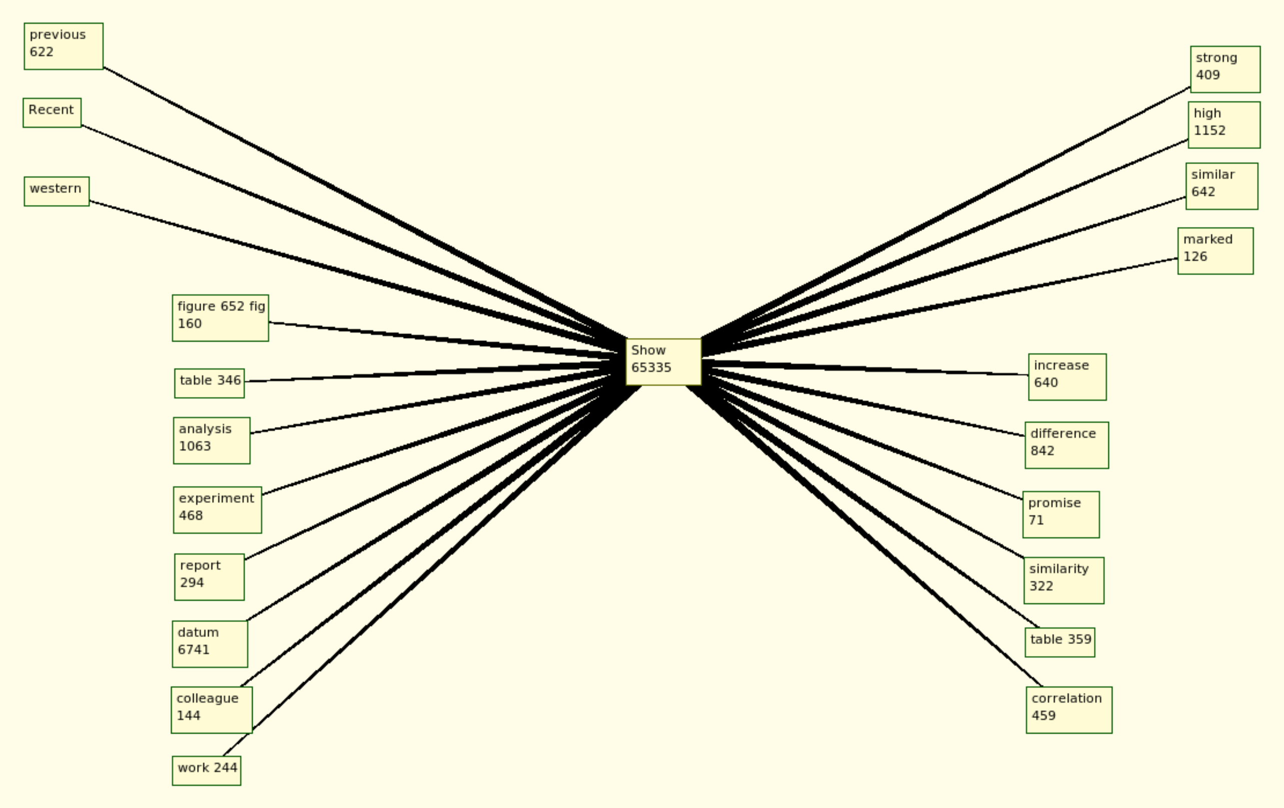

Figure 1. Noun and adjective collocates for 'show'

Figure 1. illustrates the left and right-hand noun and adjective collocates of *show*. By examining each pair in turn a number of patterns appear. As each noun is itself a lemma, content is frequently polysemic demonstrating a variety of relationships with the verb. In the case of the adjectives there is generally a link to one of the illustrated nouns, but those links also help to group the nouns into classes as can be seen in figure 2.

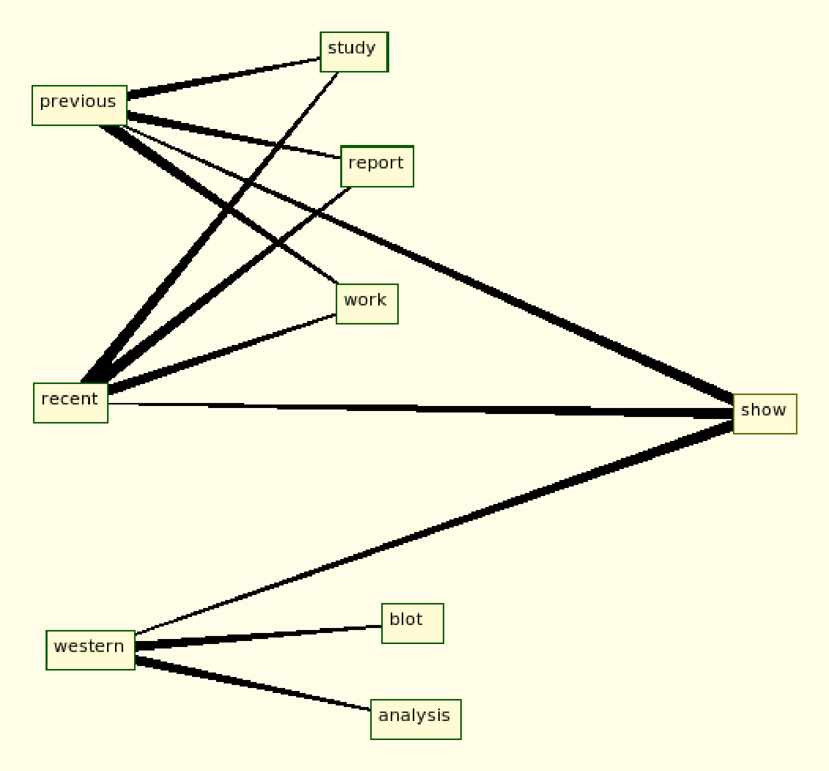

Figure 2. Interrelated adjective collocates for 'show'

Thus the adjectives *recent* and *previous* inevitably co-occur with a class we are calling *research* in that the documents subsumed under this title all report on the results from research activities. Hence the formula:

> [Recent|previous] [research[reports|studies|work] have shown that

The choice of adjective plays a connotative role in highlighting the immediate validity of the reference whilst the choice of *show* confirms factual status (Hunston and Sinclair 2003). This aspect is thus illustrated within the dictionary. The connotative aspect as well as the fact that such formula are generally found in article introductions following the IMRD model (Swales 1990) will be entered in the dictionary. The other adjective shown here, *western*, forms a terminological block with *blot*. This class is subsumed under a class of process used in carrying out research, hence:

> [Investigative process[[|Western] blot analysis]] shows that

The patterns associated with nouns also show fixed expressions as 'Table 1 shows', or 'are shown in table|figure X' as well as bracketed formula '(data not shown)' which is interesting in that it calls upon the research to trust the exactness of the authors stance. The noun 'colleagues' is found in the phrase structure "X and his colleagues' show|have shown' which is illustrated under a 'position' class in which an author, or the data, speaks. These may seem fairly straightforward formulae, but long experience in correcting research papers shows that these are not necessarily mastered by NNS scientists.

In the dictionary, all the verbs are entered in lower case and the classes in upper case. Thus 'show' comes under both EVIDENCE and POSITION classes which include the following patterns in the current dictionary:

- *EVIDENCE* patterns and collocations
  - [X] show an increase in [Y]
  - show (some/great/considerable) promise (71)
- *POSITION* patterns and collocations
  - As shown in figure X,
  - [X] is shown (schematically) in figure/table [N]
  - [X =tests] showed (no) significant differences in/between [comparison X & Y]
  - [X] shows similarity to|in [Y]

Each class gives a brief definition of the role of its members and cross links to verbs within the class.

5. Organic development

So far we have only had the individual entries and the classes to which they are linked, but building the classes and developing the lexicon is done using networks so that the total verb list will gradually increase. To do this, the collocates in figure 1. are dealt with in individual entries through their relationship with the verb node, but in network building the nominal collocates are nodes in their own right for which we shall wish to know their verbal collocates. This is process can be shown in figure 3 looking at the verbs associated with 'analysis', 'data' and 'experiment'..

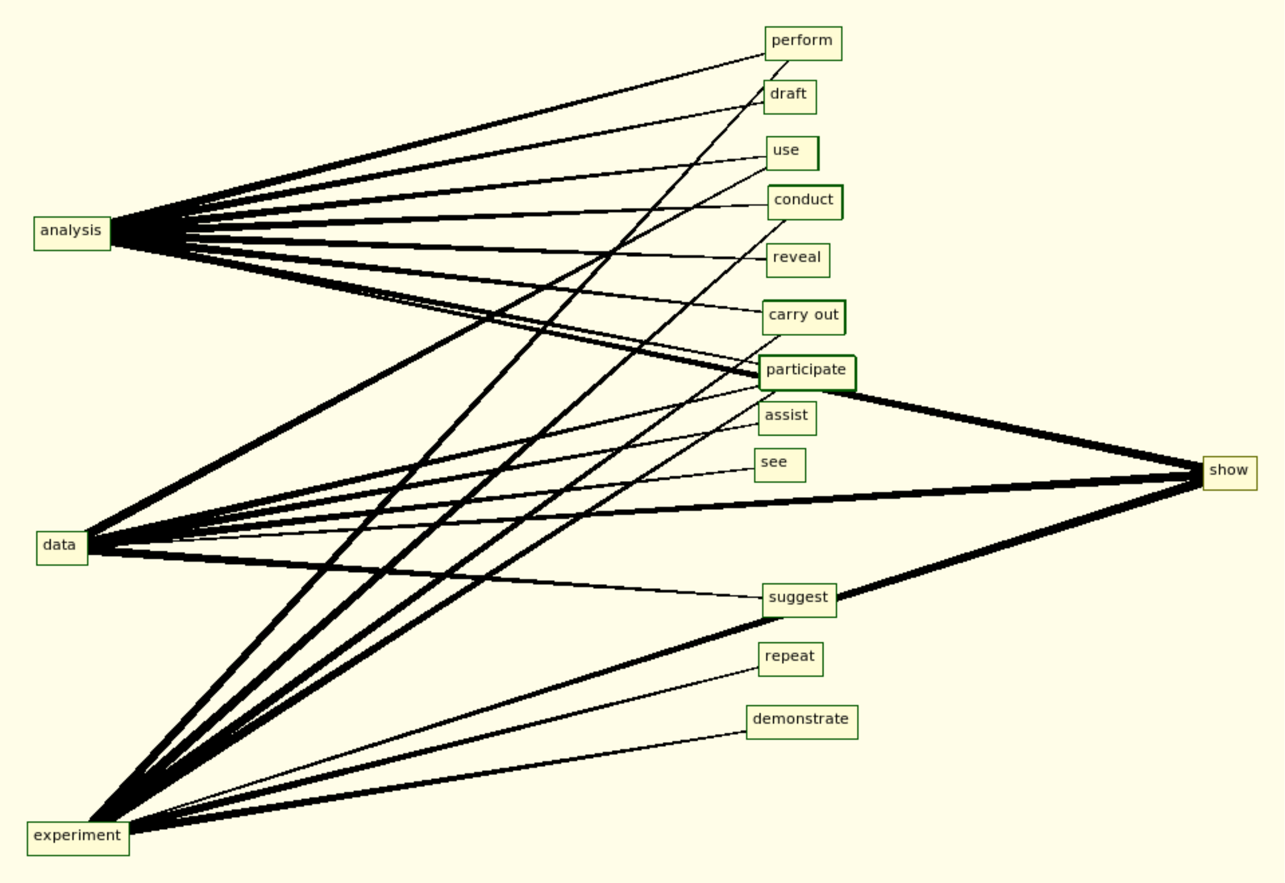

Figure 3. Noun and adjective collocates for 'show'

As can be seen from the diagram, these three nouns share a number of verbs, and will therefore join a new ANALYSIS class, but more importantly they also introduce three new verbs, 'assist', 'draft' and 'participate' that were not in the original 100. Thus the lexicon will continue to grow whilst the verb patterning and collocational environment will help classify content for ease of access.

6. Conclusion

This study does not presume to provide a fully operational dictionary in the short term. Neither does it seek to provide an exhaustive analysis of the potential patterns of each verb. The obvious methodological choice is carrying out a fuller analysis would be Corpus Pattern Analysis (Hanks), but this will require a considerable investment in time that is not currently possible. Instead, its immediate aim is to both experiment with the building process whilst rapidly making available an on-line resource that can gradually be expanded and improved upon.

The current dictionary aims at verbs, but as we have shown noun classes are also formed in the dictionary building process thereby allowing future work in a more comprehensive usage dictionary for scientists. Organic dictionaries are designed to develop naturally through statistically significant collocations. The process being iterative ,we cannot predict the number of headwords in the complete dictionary until the first sweep is completed, and even then the new words will have to be explored with the ultimate cut-off being based on frequency and statistical data. As the corpus grows, the networks can be applied to see what verbs will need to be entered, and what removed, but like with any tree, we first let it grow before contemplating pruning.

References

Hanks, P. (1994). ‘Linguistic Norms and Pragmatic Exploitations or, Why Lexicographers Need Prototype Theory, and Vice Verse’. In *Papers in Computational Lexicography: Complex 94*. 89-113.

Hanks, P. (2000). ‘Do word meanings exist?’ In A. Kilgarriff & Palmer (eds). *Sensival: Evaluating Word Sense Disambiguation Programmes. Computers and the Humanities*, 34/1-2, 205-215.

Hanks, P. (forthcoming). *Lexical Analysis: Norms and Exploitation.*

Hunston, S., Sinclair, J. (2003). ‘A local grammar of evaluation’. In Hunston, S. & Thompson, G. (2003). *Evaluation in Text: Authorial stance and the construction of discourse.* Oxford: OUP. pp 74-101.

Levin, B. (1993). *English verb classes and alternations: a preliminary investigation*. Chicago: University of Chicago Press.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Swales, J.M. (1990). *Genre Analysis. Cambridge.*: Cambridge University Press.

Williams, G. (1998). ‘Collocational Networks: Interlocking Patterns of Lexis in a Corpus of Plant Biology Research Articles’. In *International Journal of Corpus Linguistics*. Vol. 3/1. pp. 151-171.

Williams, G. (2006). ‘Advanced ESP and the Learner’s Dictionary: Tools for the non-language specialist’. In Marello, C. (éd) 2006. *Proceedings of the 12th EURALEX International Congress*. Turin, Université de Turin.

Williams G. (2008) ‘Verbs of Science and the Learner’s Dictionary’. In Bernal, E.; DeCesaris, J. (éd). 2008. *Proceedings of the 13th EURALEX International Congress*. Barcelona, Universitat Pompeu Fabra.

Williams G. (2008b). ‘The Good Lord and his works: A corpus-based study of collocational resonance’. In Granger, S. & Meunier, F. (eds). *Phraseology: an interdisciplinary perspective* Amsterdam: Benjamins. pp 159-173.

Williams, G. & Millon, C. (2009). ‘The general and the specific: collocational resonance of scientific language’. In *Proceedings of Corpus Linguistics 2009*. University of Liverpool.