Growing naturally: The DicSci Organic E-Advanced Learner's Dictionary of Verbs in Science[1]

Geoffrey Williams, Chrystel Millon & Araceli Alonso

Keywords: *collocational networks, verbal patterns, learner's dictionary, specialised dictionary, organic dictionary, phraseology*.

Abstract

In this paper we illustrate the principles and building methodology of the E-Advanced Learner's Dictionary of Verbs in Science (DicSci), paying special attention to the methodology being developed for its compilation which is based on the application of collocational networks and the adaptation of Corpus Pattern Analysis (Hanks 2004, forthcoming) to specialised language environments. DicSci focuses on showing specialised usage patterns commonly associated with certain verbs used in specialised contexts by means of collocational networks (Williams 1998). The different steps to create the dictionary, its present state and plans for its completion and future are explained.

1. Introduction

The DicSci project is on-going research developed by members of the research group LiCoRN at the Université de Bretagne-Sud (France) with the aim of creating an on-line dictionary of verbs used in scientific research papers for non-native speakers who need to produce scientific texts in English. The project has grown from research into the dictionary as a tool for on-going learning for practising scientists. In previous studies (Williams 2006, 2008) related to existing advanced learners' dictionaries, it has been stated that current dictionaries do not assist in the particular production environment of scientific articles. Although studies on dictionary use (Atkins and Varantola 1997, Nesi and Haill 2002) have contributed to bringing dictionaries ever closer to user needs, learner's dictionaries still largely ignore the needs of the non-language specialist user and are of little help in the encoding process; they are still aim at general language rather than the language of science.

The DicSci project attempts to give account of a model for creating learner's dictionaries of science which help the non-native user to produce texts in English by reflecting the full extension of real usage of lexical units used in science. It seeks to go beyond being a static model by using collocational networks to both build and manage data so that the dictionary can grow in line with the user's needs and with changes in lexical patterning as genres evolve. The DicSci project also aims to explore the potential of e-dictionaries by developing new and innovative outlooks on their management and distribution.

2. The DicSCi E-Advanced Learner's Dictionary of Verbs in Science

2.1. *An organic dictionary*

DicSci is a corpus-driven English dictionary of verbal patterns in scientific usage which shows not only patterns of certain verbs, but also where and why they are used. To do this, the mechanism of collocational networks is used as an analytical tool — see Williams 1998, forthcoming and Williams and Millon 2010, for more detailed information. This methodology is enhanced by the adaptation of the Corpus Pattern Analysis or CPA method developed by

Hanks (Hanks 2004, forthcoming) to specialised language environments (Alonso 1999, Alonso and Renau forthcoming).

One of the main innovative characteristics of the DicSci is the way in which compilation is 'organic' (Williams and Millon 2010, Alonso et al. 2011) in that DicSci is a 'living' dictionary that will organise itself in a natural way thanks to the links between words shown by means of collocational networks of verbs for headword selection, and also for structuring and classifying verbs. The principle of growing naturally of the conceptual classes and the lexicon is exemplified in the following chapter with one of the listed verbs serving as a departure node.

The corpus on which DicSci is based is the BioMed Central Corpus (BMC). The BMC was created as part of the SCIENTEXT initiative.[2] The BMC, which is freely available online[3], contains 33 millions of occurrences from more than one hundred medical and biological journals.

2.2. *The working methodology*

As explained earlier, the methodology centres on the use of *collocational networks*. A collocation network is a chain of collocations identified statistically and built in two stages. In the first stage, collocates (nominal, verbal and adjectival) of a given verb are extracted from the corpus using proximity and statistical procedures. During the second stage, the local network is extended by including the ten most statistically significant collocates of each collocate of the given verb stated in the local network. The collocates of the target verb and the collocates of the collocates thus form a collocational network of the target verb in which new verbs will be drawn in in order to enter them in the dictionary. Figure 2 illustrates how a network is developed from a central node, using the example of the verb 'call'.

The second element is the adaptation of *Corpus Pattern Analysis* to detect in a systematic way the different patterns of use of the collocates reflected in the collocational network. Corpus Pattern Analysis is an ongoing work corpus-driven methodology developed by Hanks for the creation of a *Pattern Dictionary of English Verbs* (PDEV)[4] which allows a systematic analyse of a corpus to detect the 'norms' of a word by means of identifying prototypical patterns with which words in use are associated (Hanks 2004). As established in previous work (Alonso 1999, Alonso et al. 2011), CPA is a technique which complements the information given by the collocational networks and allows differentiate uses of an item.

3. Compiling DicSci

3.1. *The compilation process*

The compilation process of DicSci must be seen as iterative. This is a natural consequence of the working methodology being used as each time a collocational network is created, new verbs, nouns and adjectives will enter the dictionary. However, not all items present in the network will be selected as headwords, as this also depends on frequency. Entries that are already included may be affected by the information extracted from the new collocational network. In this sense, the dictionary is dynamic and grows naturally; it is continuously updated. Even though the process is not linear, there are different steps which can be clearly distinguished:

Step 1: Building collocational networks. Our starting point is the one hundred most frequent verbs in the BioMed Central corpus. The choice of lexical entries is not arbitrary, but is the result of a choice guided by statistically significant collocations found in the collocational networks. The most frequent verbs are considered as nodes of the collocational networks. For each node the collocates are calculated using a statistical measure, mainly MI or Z-score. The network is extended until a point is reached where either no more significant collocates are found or where words that have occurred earlier in the network are repeated.[5] In this case a lemmatised network is desirable for having a complete panorama of the total environment of the node.
Step 2: Comparing general uses. The corresponding collocational network of the same node in a general corpus is built in order to check the similarities and differences. The PDEV is also consulted to determine the differences between the 'general' use and 'specialised' use of a verb. This stage may be developed in parallel to stage 1 as it will serve as a guide to analyse data in the specialised corpus.
Step 3: Establishing verbal patterns. Once the collocational networks are created, a thorough analysis of each main node and its collocates must be made. CPA is then applied in order to analyse every concordance and determine the most prototypical semantic patterns of usage. The CPA ontology, which is used to populate the patterns with semantic information about each collocate, is a bottom-up ontology driven by the general language corpus data and cannot be totally used in our case, as it is not adapted to our specific corpus; however it has been taken into account for getting a way to group verbs into classes.
Step 4: Linking patterns to corpus data. The linking of patterns to corpus data is done semi-automatically by using XML mark-up links.
Step 5: Classifying verbs. The lexicographical analysis of the collocational networks allows to group verbs according to some conceptual function. The functional grouping of verbs into conceptual classes is a task driven by the data and can be made in an automatic way, as has been tested by Millon (2011). However, the naming of classes is the task of the lexicographer and classes must be named so as to clarify usage to the dictionary user.
Step 6: Writing entries. For the writing of entries, the dictionary editor TshwaneLex[6] is being used. This editor will let us extract the entries in XML so that entries will be linked to each of the networks nodes. At this moment, the microstructure of the dictionary is work-in-progress and has not as yet been completed, as it depends mainly on the data extracted from the collocational networks.
Step 7: Developing online interface. The interface is based on web and mind mapping technology and different levels of information will be available according to the different users' profiles.

We are currently working on the first five steps but have also started to develop insights into step 6, specifically on the delimitation of the microstructure of the dictionary entries. Some research has also been done on the interface design, even though it is at the early stages.

3.2. *A first insight into DicSci: the verb 'to call'*

In order to illustrate our proposal we have chosen one of the most frequent verbs in our corpus, 'call'. If we check this verb in the PDEV, 'to call' is quite a complex verb with 34 different patterns, here we give only a brief insight (Fig. 1):

Sample size: **1000 / 983**

No.	%	Pattern / Implicature
1	31%	**[[Anything]] be called [NO OBJ] {[N]}** The name of [[Anything]] is [N] [[Anything]] may be an individual, a set of things, a person, a human group, an idea, or anything
2	25%	**[[Human | Institution]] call [[Anything]] {[N | ADJ]}** [[Human | Institution]] invents or uses the term [N | ADJ] to refer to [[Anything]]
3	4%	**[[Human | Institution]] call [[Event = Meeting | Action]]** [[Human | Institution]] instructs people to cause [[Event = Meeting | Action]] to happen immediately and (normally) [[Event = Meeting | Action]] does happen immediately
4	16%	**[[Human | Institution | Document]] call [NO OBJ] {for [[Action]] | for [[State]]}** [[Human | Institution | Document]] says that other people should do [[Action]] or create [[State]]

Figure 1. First four patterns of 'to call' in the PDEV.

In the BioMed corpus, 'to call' has 3355 occurrences as a lemma. The networks were obtained using the methods described above. Once the data were obtained, the collocational network was built by using the tool Gephi.[7] In figure 2, the nodes in red are the nouns and the ones in green are the verbs. The verbs can be shared by two or more nodes in the network.

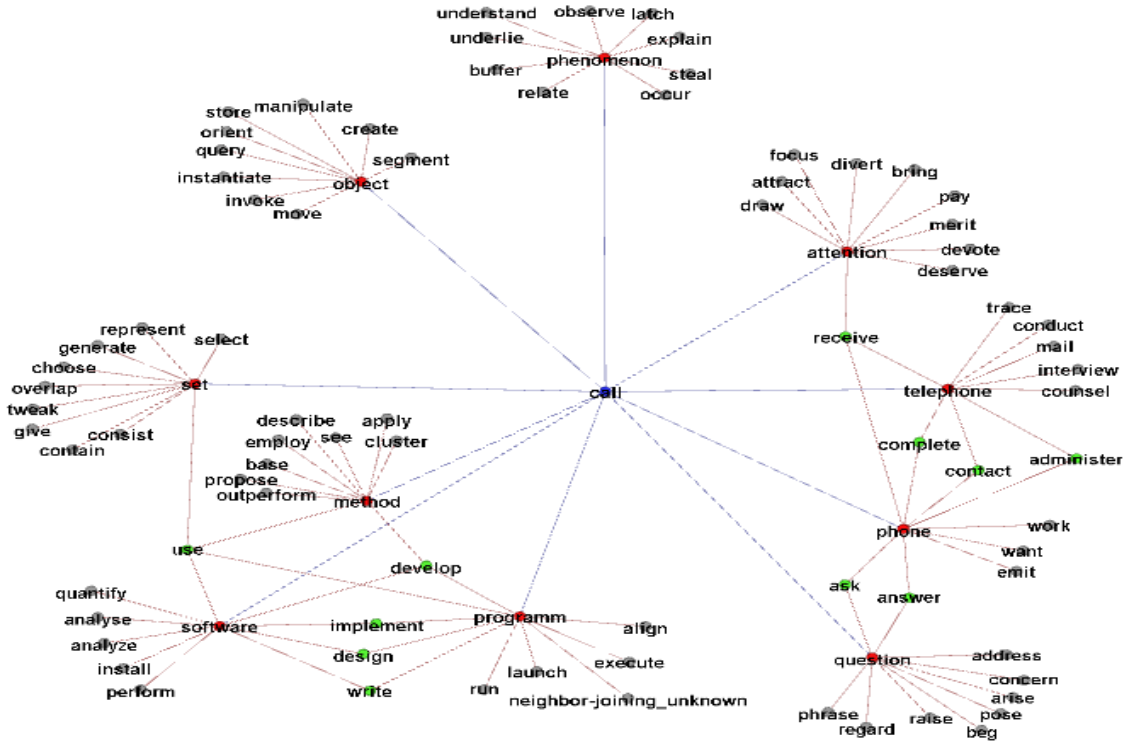

Figure 2. First levels of the collocational network of 'to call' in the BMC.

By looking at the collocational network in the BMC and observe the most frequent collocates two groups can be made: *programme* (51 occurrences), *software* (27), *phenomenon* (23) and *object* (18) in the first group and *attention* (26) in the second group. The first group is fairly straightforward, often using the passive with 'X is called' as the phraseological pattern, although 'X call Y Z', can be also found. The patterns are similar to the first two patterns in the PDEV. It can be observed that in this case the verb fulfils the task of NAMING. By looking at the concordances, it can be observed that even though the patterns are similar to that used in general language, in the BMC the difference lying in the arguments of the verb; in terms of Hanks, on the number of collocates that normally populate an

argument slot in relation to the verb. While in general language '[[Anything]] is called', in our corpus 'X' is a more limited set of lexical items.

The second group represented by the collocate 'attention' is a richer class. It corresponds to the pattern 'X call attention to Y'. In the PDEV this pattern is also described but not as one of the most frequent ones:

[[Human]] call {attention} {to [[Event]]}
[[Human]] asks people to notice [[Event]].

In this case, the class is not that of NAMING, but can be defined as UNDERLINING. It is an essential function as in the BMC by using this pattern, the author seeks to draw attention to a key point or issue.

By widening the network to a new level, other verbs used with an specific collocate are shown. For instance, the top five verbal collocates of the collocate 'attention' are *receive* (239 occurrences), *focus* (187), *draw* (109), *attract* (71), *deserve* (51). The network allows differentiating between potential synonyms.

As can be seen, collocational networks bring about a rich analysis of how language is used in scientific texts giving the non-native speaker of English a valuable source of information to improve communication skills and produce scientific texts in English.

4. Conclusions and perspectives

Very much like CPA, DicSci is unfunded ongoing work, which means that advances depend heavily on the availability of the researchers involved. DicSci is both experimental and practical, that is to say seeks to create a usable dictionary using techniques that are in themselves evolving. Collocational networks are not new, first put forward by Williams in 1995, the methodology has been revised ever since and has been adopted in other projects. Their use in experimental dictionaries is not new either as this was proposed early on and has been largely experiment. What is new is the integral approach that is currently being developed that includes CPA. By combining three approaches, we hope to quickly get a usable dictionary available as it is only in this way that we can move on to experiment with new ways to display the data.

Notes

[1] Research for this article was funded by the Equipe LiCoRN of the HCTI research group and also by the Spanish Ministry of Education as part of the *National Mobility Programme of Human Resources of the R+D National Programme 2008-2011* which has made possible the post-doctoral work of one of the authors. It has also been supported by the Spanish National Project HUM2009-07588/FILO.

[2] An explanation of the SCIENTEXT initiative is available at http://scientext.msh-alpes.fr/scientext-site/spip.php?article1

[3] http://scientext.msh-alpes.fr/scientext-site/spip.php?article30

[4] The PDEV is available at http://deb.fi.muni.cz/pdev/

[5] A more detailed description of the procedure is shown in Williams (1998).

[6] See http://tshwanedje.com/tshwanelex/

[7] The open-source graph visualization platform Gephi is available at http://gephi.org/

References

A. Dictionaries

Hanks, P. (ed.). In progress. *Pattern Dictionary of English Verbs* (PDEV). http://deb.fi.muni.cz/pdev/.

B. Other literature

Alonso, A. 2009. *Características del léxico del medio ambiente y pautas de representación en el diccionario general*. PhD Thesis, Institut Universitari de Lingüística Aplicada – University of Pompeu Fabra.

Alonso, A. and I. Renau Forthcoming. 'Corpus Pattern Analysis for the Description of Verbs Used in Science.' *Terminàlia*.

Alonso, A., C. Millon and G. Williams 2011. 'Collocational Networks and their Application to an E-Advanced Learner's Dictionary of Verbs in Science (DicSci).' In I. Kosem and K. Kosem (eds.), *Electronic Lexicography in the 21st Century: New Applications for New Users. Proceedings of eLex 2011*, Bled, 10-12 November 2011. Ljubljana: Trojina, 12–22. http://www.trojina.si/elex2011/Vsebine/proceedings/eLex2011-2.pdf

Atkins, B. T. S. and K. Varantola 1997. 'Monitoring Dictionary Use'. *International Journal of Lexicography* 10.1: 1–45.

Fillmore, C. J. 1982. 'Frame Semantics.' In Linguistic Society of Korea (ed.), *Linguistics in the Morning Calm*. Seoul: Hanshin Publishing Co., 111–137.

Hanks, P. 2004. 'The Syntagmatics of Metaphor and Idiom.' *International Journal of Lexicography*, 17.3: 245–274.

Hanks, P. Forthcoming. *Lexical Analysis: Norms and Exploitations*. Massachusetts: The MIT Press.

Ježek, E. and Hanks, P. 2010. 'What Lexical Sets Tell us about Conceptual Categories.' *Corpus Linguistics and the Lexicon, Special Issue of Lexis, E-Journal in English Lexicology* 4: 7–22.

Millon, C. 2011. *Acquisition automatique de relations lexicales désambiguïsées à partir du Web*. PhD Thesis, University of Bretagne-Sud.

Nesi, H. and R. Haill 2002. 'A Study of Dictionary Use by International Students at a British University.' *International Journal of Lexicography* 15.4: 277–306.

Williams, G. 1998. 'Collocational Networks: Interlocking Patterns of Lexis in a Corpus of Plant Biology Research Articles.' *International Journal of Corpus Linguistics* 3.1: 151-171.

Williams, G. 2006. 'Advanced ESP and the Learner's Dictionary.' In E. Corino, C. Marello and C. Onesti. (eds.), *Proceedings of the XII EURALEX International Congress*. Torin: University of Torin, 795–801.

Williams, G. 2008. 'Verbs of Science and the Learner's Dictionary.' In E. Bernal and J. DeCesaris (eds.), *Proceedings of the XIII EURALEX International Congress.* Barcelona: Institut Universitari de Lingüística Aplicada – University of Pompeu Fabra, Documenta Universitaria, 797–806.

Williams, G. Forthcoming. 'Bringing Data and Dictionary Together: Real Science in Real Dictionaries.' In A. Bolton, S. Thomas and E. Rowley-Jolivet (eds.), *Corpus-Informed Research and Learning in ESP: Issues and Applications*. Amsterdam: John Benjamins, 219–240.

Williams, G. and C. Millon 2010. 'Going Organic: Building an Experimental Bottom-up Dictionary of Verbs in Science.' In A. Dykstra and T. Schoonheim (eds.), *Proceedings of the XIV EURALEX International Congress*. Leeuwarden: Fryske Akademy, 1251–1257.