ASp
la revue du GERAS

31-33 | 2001
Varia

Mediating between lexis and texts: collocational networks in specialised corpora

Geoffrey Williams

Electronic version
URL: http://journals.openedition.org/asp/1782
DOI: 10.4000/asp.1782
ISBN: 978-2-8218-0384-8
ISSN: 2108-6354

Publisher
Groupe d'étude et de recherche en anglais de spécialité

Printed version
Date of publication: 1 October 2001
Number of pages: 63-76
ISSN: 1246-8185

Electronic reference
Geoffrey Williams, « Mediating between lexis and texts: collocational networks in specialised corpora », *ASp* [Online], 31-33 | 2001, Online since 25 October 2010, connection on 19 April 2019. URL : http://journals.openedition.org/asp/1782 ; DOI : 10.4000/asp.1782

Mediating between lexis and texts: collocational networks in specialised corpora

Geoffrey Williams

1 Lexis is something that is stored in dictionaries or wordlists. Texts are pieces of paper that we study in class, explaining structure and vocabulary so as to assist with comprehension. Students, as we know, are often reluctant to buy monolingual dictionaries and tend to stock their own vocabulary in the form of decontextualised lists. As teachers, our own attempts to mediate between lexis and text are invariably subjective, even if we construct a **corpus**, we still have to wade through large volumes of data to find what we consider significant. **Collocational networks** propose a more objective methodology for the extraction of the lexis that typifies a field by making use of the lexical relationships into which words enter, that is through collocation.

2 Since Firth first put forward the notion of collocation in the nineteen thirties,[1] this particular linguistic phenomenon has gradually reached acceptance, especially with the advent of easily accessible machine corpora. Being idiomatic by nature, collocation did not get the favour of dominant structuralist and formal schools of linguistics on the European continent and in the United States, it has however gradually come to be seen as playing a central role in language learning and translation. The rise of the field of phraseology is symptomatic of this change; words are no longer seen simply as elements in isolation that can be slotted into syntactic frameworks, but as forming larger units. In language teaching this has been explored as lexical phrases by such researchers as Nattinger and DeCarrico (1992). Collocation on the other hand has largely been the field of lexicographers, the monumental work being the BBI dictionary (Benson et al. 1986), a boon to language learners at intermediate level and beyond and to translators alike.

3 Whereas their nebulous nature makes lexical phrases difficult to formalise linguistically, we seem to be on clearer ground with collocations as being a binary relationship between two words. Firth does not greatly help in formalising the concept in that he essentially defined collocation through example, for instance his canonical *silly ass* (1957). This

example would imply that collocations are binomial syntagmatic relationships realised through a wide variety of syntactic pairings as adjective-noun or verb-noun. Indeed, in the BBI we do find a classification into lexical and grammatical pairings. At the semantic level these are interpreted through a distributional approach to meaning; one word giving semantic weight to another. However, the lexicographical approach to this phenomenon is not the only one, the textual approach, epitomised by the Birmingham school of Sinclair see collocation as essentially associative revealing the thematic coherence of a text. Consequently, it is perhaps easier if we recognise two main traditions in collocational studies (Williams 2000a) in which we have a lexicographical school, which seeks to formalise collocation in functional and syntactic terms, and a contextualist school, which views the concept as a textual phenomenon. Both may be treating similar subject matter, but from a different standpoint.

4 The research described here takes as its ultimate aim a lexicographic description of collocation, but is based on a textual theory. The basis of this research is the notion of collocational networks (Williams 1998, 1999a). These are networks of collocationally-related lexical items in which equal weight is given to each item in a collocational pair as a potential base node. The networks are then built by exploring the lexical relationships into which each item enters. In this text, I shall first discuss the nature of collocation before introducing the network theory. I shall then illustrate collocational networks through the example of a corpus in parasitic plant biology and conclude with some practical applications.

Collocation: the nature of the beast

5 As seen above, the notion of collocation was illustrated rather than defined by Firth (1957: 179) who claimed that "you shall know a word by the company it keeps". This then is not formal linguistics based on intuition and the study of isolated, made-up sentences, but language in context. Heavily influenced by Malinowski's (1921) work on the social and cultural factors inherent in language use and evolution, Firth placed word meaning as being a function of context:

> The basic assumption of the theory of analysis by levels is that any text can be regarded as a constituent of a *context of situation*... (1957: 175)

6 Collocations are not simply fixed elements, but the product of the dynamic process of language production, becoming gradually institutionalised through usage in a given language environment. This entails that different situational contexts, that is national or regional varieties, special languages or specific genres would develop collocations unique to that environment. Institutionalisation takes us away from what Sinclair (1991) has termed the open choice principle to the idiom principle where we make use of reusable chunks of language within specific contexts. In this approach, language use is seen as being basically idiomatic in nature.

7 Within the wider field of fixed expression and idioms (Moon 1998), collocation is generally illustrated by examples such as Firth's *silly ass* or pairs as *powerful*/strong tea*. However, whilst we can easily point to central examples of this phenomenon, there will be some disagreement as we move away from these canonical examples. Generally, in order to define collocation, four characteristics are brought forward, namely that collocations are:

- Habitual

- Lexically transparent
- Arbitrary
- Grammatically well-formed

8 Unfortunately, different linguists give more or less weight to individual criteria, depending on the school of thought in which they are working. It is then useful to rapidly take each in turn to discuss its validity; a more complete discussion may be found in Williams (2001).

Habitual

9 The first condition is that proposed by Firth who claimed that "collocations of a given word are statements of the habitual or customary places of that word in collocational order" (1957: 181). In textual collocation, following the Birmingham school, this is interpreted as signifying that collocations may be measured statistically. Indeed, according to Sinclair:

> SIGNIFICANT COLLOCATION is regular collocation between two items, such that they co-occur more often than their respective frequencies and the length of the text in which they appear would predict. (Sinclair 1970: 150)

10 This essentially statistical view allows for the automatic extraction of potential collocational pairs, a possibility exploited by researchers such as Smadja (1993). Although the use of statistics makes a valuable contribution to the extraction of collocations, they pose a problem as to their status as a defining criterion in that different measures will, potentially, extract different pairings. A number of tools have been used such as the z-score, the t-score and mutual information. Each has been shown to be efficient for the purpose for which it was chosen, but that is hardly a defining characteristic; it would be better to say that such tools provide potential or candidate collocations.

Lexically transparent

11 This is a linguistic criterion much discussed within lexical semantics (Cruse 1986). It must be said that here we are discussing only lexical collocation, that is a relationship between two semantically full words, to the exclusion of functional words. In this context, according to Cruse, the essential difference between a collocation and an idiom is that in the former each word remains fully transparent whereas in the latter the meaning can no longer be decomposed. Thus *kick the bucket* would be discarded as opaque and *heavy drinker* accepted as transparent. However, in real language we cannot make hard and fast rules, *make*, for instance, can be transparent as in *make a cake*, where it does have the meaning of creating something. However, in other cases, the verb is less transparent, thus *make the bed* is more problematic and *make love* even more so. The same problem can be seen with *public*, which is transparent in *public baths*, but less so in *public school*. Thus, a word may form a collocation in one case, but an idiom in another. Collocation is obviously a question of degree.

Arbitrary

12 According to the lexicographer Morton Benson (1989 : 3) "collocation should be defined as not just 'recurrent word combinations' but as 'arbitrary word combinations'". What

this points to is that much of what we take for granted is simply not translatable literally from one language to another. This is apparent with adjective-noun forms as *heavy traffic/ circulation intense* as well as verb-noun collocations as *break the lawenfreindre la loi.* The problem is to decide as to the degree of arbitrariness on a free-fixed continuum (Haussman 1976), something which can only really be appreciated by comparing across languages. However, if we take examples from French and English, we can find freely translatable phrases as "he had a heart attack" which also have arbitrary translations as in "*il a fait un infarctus*". There can be no hard and fast dividing line for, as Haussman (1997) has pointed out, "*tout est idiomatique dans la langue*".

Grammatically well-formed

13 The condition of well-formedness has been added by lexicographers as Kjellmer who need to describe syntactically coherent units. The use of such linguistic filters coupled with statistical tools has allowed a better extraction of terms and collocations using Natural Language Processing (NLP) methodologies as for example in Xtract, the programme created by Smadja (1993) for the extraction of collocations and specialised phraseological units. However, the textual collocation approach does not necessarily require this restriction. Indeed, according to Firth, partially cited earlier:

> Collocations of a given word are statements of the habitual or customary places of that word in collocational order but not in any other contextual order and emphatically not in grammatical order. (1957: 181)

14 In other words, collocation is syntagmatic recurrence, which may be described in syntactical terms, but this is not a condition.

From collocation to collocational networks

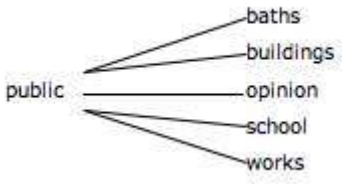
15 Quite apart from the questions of degree of collocability, linguists do not necessarily agree as to the validity of all four conditions. Whilst most lexicographers would accept all four, many would give less weight to the statistical aspect in that statistics can only indicate, but not prove a relationship. What may be more important to the lexicographer is the intuition of the linguist in revealing native speaker competence. Similarly, transparency may be only partially applied as, for instance, in some work within the Melcukian lexical function school (Kahane & Polguère 2000) where an idiomatic form as *colère noire* is treated as collocation.

16 Whilst rigid definition may not be possible, different attempts have been made to isolate what is prototypical of collocation. This has given rise to two main schools of thought: a lexicographic school looking at syntagmatic structures, often out of context, and a contextualist school heavily dominated by Sinclair's corpus linguistics approach (Williams 2000a, 2000b). The latter approach does not deny the lexicographical definition, but simply looks at the phenomenon from the viewpoint of word association. The theory of collocational networks (Williams 1998, 1999a) belongs to this latter approach.

17 In traditional lexicographical practice, collocation is seen as a binary relationship between a node or base, generally a noun or verb, and its collocates, adjective or adverb. Rather than looking at simple collocational pairs, network theory posits the possibility of moving out from a central node so as to explore the significant word associations of both base and collocate. Given that it is purely contextualist, this theory does not take into

account grammatical well-formedness and does not consider the problem of transparency in extraction. Consequently the following collocational paradigm for *public* is quite acceptable (Fig. 1).

Figure 1. Collocational paradigm for public



18 The only limiting factor is the corpus: the more specialised the corpus, the more restricted the paradigm.

19 Whereas the above combinations would be quite sufficient in standard collocational theory, in collocational networks the collocations of each of the collocates listed would then be explored. This means that whereas Melcukian **lexicography** would see *heavy* only an intensifier for a given noun, network theory would treat it as significant and explore its collocates. The combinations arising from adjectives or adverbs are not trivial insofar as within a given field only a restricted number of nouns or verbs could be thus modified. Whilst it is inevitable that such an uncontrolled system would allow ambiguity to slip in, this is not considered a problem as disambiguation would come in later as part of the semantic analysis of the networks.

20 The lexicographical interpretation of collocation sees only binomial pairings within a certain context. In Hallidayan grammar, collocation plays a more active role in textual cohesion. According to Halliday:

> [...] even where there is a relation of synonymy between lexical items, their cohesive effect tends to depend more on collocation, a simple tendency to co-occur. (1995 : 333)

21 In exploring the patterns of relationships from networks, we find that collocation can go beyond this role in intra-textual cohesion to reveal patterns that are significant for texts emanating from a discourse community. These patterns may then be used to demonstrate the essential lexis of that community.

22 Such a possibility was foreseen by Berry-Roghe (1973) who was working on literary texts, although the capacity of computing power at the time limited further exploration. Later, Phillips (1985) worked on scientific texts from the COBUILD corpus building restricted networks to define what he termed the "aboutness" of the text. Collocational networks as described here go much further in that they link essential concepts of a community by isolating the essential lexis from a corpus.

Collocational networks: methodology

23 Collocational networks adopt a statistical approach in which the cohesive nature of the phenomenon is exploited. The notion of cohesive collocation as used here may be defined as the habitual and statistically significant relationship between word forms within a predefined window and for a defined discourse community, expressed through an electronic corpus of texts. The corpus used here, BIVEG1[2] was carefully constructed

following external (Williams 1998b, 1999a) and internal (Williams 1999b) selection criteria so as to have a coherent, justifiable source. It currently consists of 155 research articles from the field of parasitic plant biology amounting to about 500,000 words. The corpus is entirely marked up in Standard Generalized Markup Language following the Text Encoding Initiative recommendations.

24 The statistical measure used is that of mutual information (Church & Hanks 1990). This tends to privilege rare and specialised usage which makes it ideal for the exploration of special language corpora. Other than a stoplist to eliminate function words, no linguistic filters are applied. The measures were prepared using Excel and wordlists produced by WordSmith Tools. A more rough-and-ready network can be built quickly using the mutual information facility of WordSmith.

25 Networks may be produced from high frequency keywords or from a semantically-related group of words. The start word is referred to as the central node. The left and right collocates are calculated for the central node and noted on a graph. The graph serves only to illustrate the evolving network; a stage is quickly reached where the interrelationships are too complex to be displayed in this way. The collocative links and resulting lexicon are stored on an Excel spreadsheet for later semantic analysis of corpus content. In every case the collocates of a given node are considered as nodes in their own right allowing the network to grow. Once a word is repeated in the word lists, it is marked and not re-explored as this would introduce circularity.

26 The approach is best illustrated through an example drawn from the BIVEG1 corpus of plant biology research articles.

Building collocational networks

27 In this particular theme-specific discourse community, the topic under study is parasitic plants, either as botanical species or as highly invasive weeds. The cohesive nature of collocation in this community can be illustrated through the introduction to a single research article in this field:

> Witchweed (Striga asiatica [L.] Kuntze) is an important **parasitic** weed on several poaceous **crops**, including sorghum. **Crop** yields may be reduced by as much as 90% in infested land. The **parasite** produces large numbers of seeds with prolonged viability and special germination requirements. (Babiker *et al.* 1993: 89)

28 The first two sentences use complex lexical repetition (Hoey 1991: 55) in that use is made of repetition of lexical morphemes, in this case "parasite-parasitic" and "crop-crops". This is in itself a form of lexical cohesion. However, these words also perform a role of collocational cohesion through their regular co-occurrence with other lexical items. If we accept that collocation can be cohesive, not just for individual texts, but also for a discourse community, we can look at the collocative patterns for these words in the corpus. Using mutual information, the following collocations for "crop" can be observed:

attractive - attractive crop
induces - induces crop losses
sown - directly sown crop
vigor - crop vigor
rotation - crop rotation

29 These are also lexicographic collocations of the adjective-noun, verb-noun or noun-noun form. Their variations may not necessarily be described in terms of grammatical relations, but they remain collocative in the textual cohesive sense, for instance:

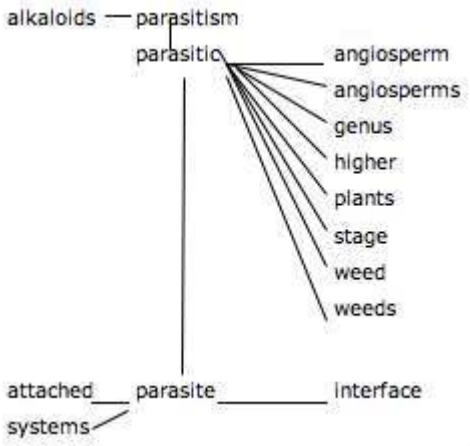
> crop vigor
> The vigor of the crop
> Many factors affect the crop; its vigor...

30 The same may be done for the adjective "parasitic" which forms the following collocational groupings:

> angiosperms - parasitic angiosperms
> plants - parasitic plants
> weeds - parasitic weeds
> non-parasitic - parasitic & non-parasitic plants
> life - parasitic way of life

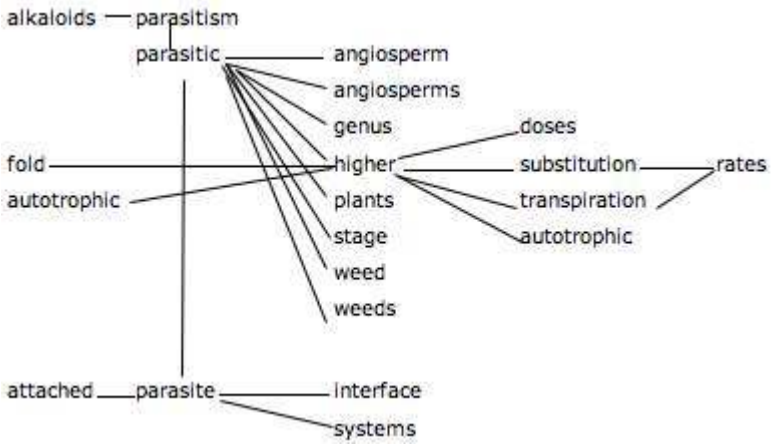
31 These collocations, and those of their morphemic paradigm may be illustrated in a local network (see Fig. 2).

Figure 2. Local network



32 The collocates are clearly central to the field in that they concern angiosperms, that is flowering higher plants, which can be seen as both parasitic weeds, invasive and therefore a nuisance in agriculture, and as parasitic plants, objects of biological study. A number of collocational pairings and multi-word units may be formed with these words. Insofar as the collocate is a full lexical item, they can now be treated as nodes and their collocates explored. The result is a more complex display (fig. 3):

Figure 3. Nodes and collocates

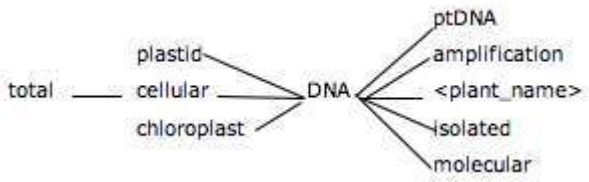


33 The networks are clearly selecting key words of the field. In addition, this example shows the use of "unimportant" adjectives as being useful collocational mediators within a given textual environment.

General to specific networks

34 Networks may illustrate a general theme in a corpus or a more specific one. The theme-specific discourse community of parasitic plant biology is multidisciplinary and calls upon a number of disciplines in biology. Among the disciplines concerned by this theme, molecular biology is essential for studying both the intra-cellular structure of the plants and for grouping them into phylogenetic families and clades. Consequently, as in any molecular study, DNA becomes a key word forming the specific local network shown in Figure 4:

Figure 4. Specific local network



35 Looking at theme specific networks entails subdividing the corpus, moving from the general theme to specific aspects of that theme. Such subdivision is carried out through internal selection criteria (Williams 1999b) and helps illustrate potential bias in a corpus through the weighting given to individual approaches. This is all the more important in biology, where the molecular approach has greatly increased in importance in recent years and has introduced a vast amount of jargon based on a computational usage rather than derived from a Latin base. Its approach has also highlighted a different way of looking at cellular and intra-cellular phenomena. Thus, local networks can show very different viewpoints on a single concept. From a physiological point of view (Fig. 6.) "gene" is seen as a unit performing certain functions as resistance and defence; it is the character, recessive or dominant, that is important. The same word used by molecular biologists deals with internal structure of the gene, whether the gene is presence or not

and what it may code for. Exploring these networks allows an understanding of domain-specific usage in relation to theme-specific concerns and can assist in cross communication within the field by highlighting the concerns of the different disciplines involved. It is all too easy to lose the specific in a general overview of a field.

Ever-expanding circles

36 Networks do not grow indefinitely, but do tend to grow rapidly at three removes from the start node only to tail off at five removes. The most frequent lexical words form obvious starting points, although in reality they rapidly join into one network that will take in central themes covered in the corpus. Another starting point would be group elements that are linked in one way or another to form a complex central node. This has been done to look at the vocabulary related to the parasitic plants themselves. In this case, instead of taking one plant, Striga, for instance, we take all the identified parasitic plants and explore the collocates of this mass node. The result is a large number of initial collocates forming a network that is so extensive that it can only be fully explored by automating the extraction. This automation process, and its applications in lexicography and artificial intelligence, is currently underway and should lead to the easier mapping of complex phenomena.

37 Building networks reveals the thematic content of a corpus, a topic of study in itself. Using networks to locate internal selection criteria can then lead to subcategorisation of corpora and the exploration of the lexical contribution of individual disciplines in multidisciplinary studies. In exploring the lexis we gradually reveal terms, lexical phrases and lexicographic collocations that typify the lexis of the domain. The essentially objective nature of the methodology allows the extraction of lexis that can then be studied and taught in relation to a precise field.

Conclusion

38 It is obvious that no system is perfect, especially one that seeks to automatically process that most fundamental aspect of human nature, language. As noted earlier all statistical measures have their advantages and inconveniences; mutual information does seem well adapted to special languages, but will inevitably introduce both noise and silence.

39 Noise, that is the presence of non-pertinent material, is often partially overcome with linguistic filters, these, however, are not without their inconveniences. Stoplists are far from infallible and potentially exclude interesting material, pattern grammar (Hunston & Francis 1999), for example, requires the presence of function words. For some languages other than English their presence is vital in all collocational studies (van der Wouden 2000). For collocational networks, however, function words can be ignored, and would almost certainly be eliminated by the mutual information process anyway. Other linguistic filters have been used for collocation extraction (Smadja 1993), but these require part of speech tagging in an attempt to eliminate all that is not well-formed. In a textual approach to collocation, such an approach would not be appropriate. Another difficulty in the elimination of noise is to decide what this phenomenon actually entails. It must be borne in mind that the ultimate aim here is the building of a dictionary, not a terminology. Terminologies do not look at “non-scientific” words, but for a dictionary these may need to be present in that they provide the context in which terminology

operates. It is upon contextual wordings that the non-native writer generally stumbles rather than on domain-specific terms. Elimination of noise in terminology can lead to silence in lexicography.

40 Silence is another problem, how to know how much is left out and, if possible, measure the degree of loss. This can really only be calculated with reference to the community of users. Although statistics have been used, we should not adopt a purely quantitative view of language, but also adopt a qualitative one in which the community of users is consulted. Only the users can legitimately say what is missing, but these are also often unaware of what should be there from a language point of view. Linguistic analysis must be a cooperation in which the expertise of both researcher and linguist is respected. It must also be recognised that what is not in the corpus will not be seen. Networks work within a finite corpus, but that corpus, and the networks, need to be maintained if the concerns of the domain are to be adequately covered.

41 The role of these networks is to mediate between lexis and text, allowing the building of a lexical picture of a domain in a way similar to brainstorming for word association, with the exception that the methodology is objective and text-based. Working within a carefully constructed corpus based on a discourse community, the networks gradually reveal the significant lexis of the field and the candidate combinations for more complex structures. They are not the only way, but one way of classifying words. As Wittgenstein has said:

> [...] how we group words into kinds will depend on the aim of the classification, — and on our own inclination. (1957: 17)

Figure 5. "Gene" in the molecular biology sub corpus

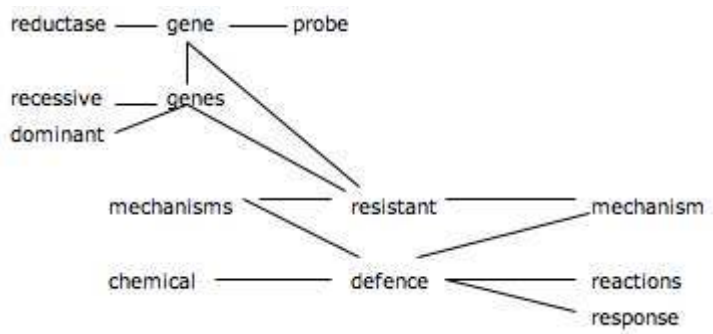
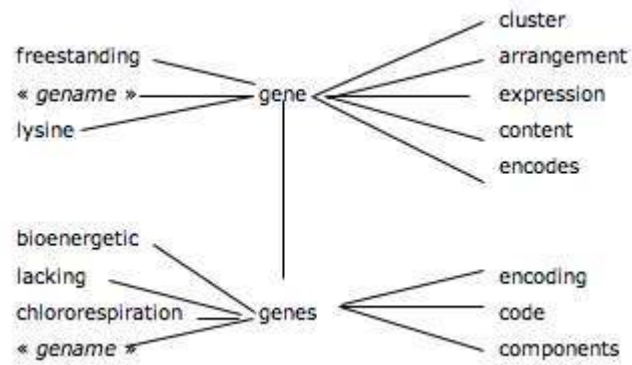


Figure 6. "Gene" in the plant physiology sub corpus



BIBLIOGRAPHY

Babiker, A.G.T., L.R. Butler, G. Ejeta & W.R. Woodson. 1993. "Enhancement of ethylene biosynthesis and germination by cytokinins and 1-aminocyclopropane-1- carboxylic acid in Striga asiatica seeds". *Physiologia Plantarum* 89.

Benson, M., E. Benson, R. Ilson. 1986. *The BBI Dictionary of English Word Combinations*. Amsterdam: John Benjamins.

Berry-Roghe, G.L.M. 1973. "The computation of collocations and their relevance in lexical studies". In Aitken A.J. *et al.* (eds.), *The Computer and Literary Studies*. Edinburgh: Edinburgh University Press.

Church, K. & P. Hanks. 1990. "Word association norms, mutual information and lexicography". *Computational Linguistics* 16/1, 22-29.

Cruse, D.A. 1986. *Lexical Semantics*. Cambridge: Cambridge University Press.

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

Halliday, M.A.K. 1994. *Introduction to Functional Grammar* (2nd edition). London: Edward Arnold.

Haussman, F. J. 1976. "Linguistik des Wortschatzlernens im Französischstudium". *Grazer linguistische Studien* 4, 49-60.

Haussman, F.J. 1977. "Tout est idiomatique dans les langues". In Martins-Baltar, M. (ed.), *La locution entre langue et usages*. Fontenay St Cloud: CNRS Editions, 277-290.

Hoey, M. 1991. *Patterns of Lexis in Text*. Oxford: Oxford University Press.

Hunston, S. & G. Francis. 1999. *Pattern Grammar*. Amsterdam: John Benjamins.

Kahane, S. & A. Polguère. 2000. "Un langage formel d'encodage des fonctions lexicales et son application à la modélisation des collocations". In Daille, B. & G. Williams (eds.), *La Collocation – Journée d'études de l'ATALA, Paris 13 janvier 2001* IRIN rapport de recherche 00.13 décembre 2000, 15-18.

Kjellmer, G. 1984. "Some thoughts on collocational distinctiveness". In Aarts, J. & W. Meijs (eds.), *Corpus Linguistics: Recent advances in the use of computer corpora in English language research*. Amsterdam: Rodopi, 163-171.

Nattinger, J. & J. DeCarrico. 1992. *Lexical Phrases and Language Teaching*. Oxford: Oxford University Press.

Palmer, F.R. (ed.). 1968. *Selected Papers of J.R. Firth 1952-59*. London/Harlow: Longmans.

Phillips, M. 1985. *Aspects of Text Structure: An investigation of the lexical organisation of text*. Amsterdam: North-Holland.

Sinclair, J. McH. *et al.* 1970. *English Lexical Studies: Report to OSTI on Project C/LP/08*. Department of English, University of Birmingham.

Smadja, F. 1993. "Retrieving collocations from text: Xtract". *Computational Linguistics* 19/1, 399-413.

Van der Wouden, T. 2000. *Collocational Behaviour in the Modal Realm*. DGfS, Philipps-Universität, Marburg, March 2000.

Williams, G. 1998. "Collocational networks: Interlocking patterns of lexis in a corpus of plant biology". *International Journal of Corpus Linguistics* 3/1, 151-171.

Williams, G. 1999a. "Les réseaux collocationnels dans la construction et l'exploitation d'un corpus dans le cadre d'une communauté de discours scientifique". Thèse en anglais – linguistique de corpus, Université de Nantes http://geoffrey.williams.pagesperso-orange.fr/thesis/index.htm/.

Williams, G. 1999b. "Looking in before looking out: Internal selection criteria in a corpus of plant biology". *Proceedings of the 4th International Conference on Computational Lexicography*, Complex '99. Budapest.

Williams, G. 2000a. "Collocational Networks as the realisation of a specialised textual environment". DGfS, Philipps-Universität, Marburg, March 2000.

Williams, G. 2000b. "Collocational networks for prototypical thematic grouping in a French single source corpus". *Proceedings 5th TELRI Seminar,* Ljubljana, Slovenia, September 2000.

Williams, G. 2001. "Sur les caractéristiques de la collocation". *Actes de TALN* 2001, Tours vol. 2, 9-16.

Wittgenstein, L. 1953. *Philosophical Investigations*. Oxford: Basil Blackwell.

NOTES

1. The great British linguist, John Rupert Firth (1890-1960) was the founder of what came to be known as the British Contextualist school wherein meaning is seen as inextricably linked to context of use. He published numerous articles during his lifetime; the most significant of these have been grouped in two volumes*, Papers in Linguistics 1934-1951* (Firth 1957) and *Selected Papers of J.R. Firth* (Palmer 1968).

2. BIVEG1 is version 1 of a corpus of research articles in parasitic plant biology. It consists of articles from specialised journals and conference proceedings. BIVEG is a monitor corpus in that new material is constantly being added so as take into account diachronic aspects of language use. The corpus is part of speech tagged and marked up in SGML following the TEI guidelines. It is currently being converted to XML.

ABSTRACTS

The theory of collocational networks adopts a textual definition of the phenomenon of collocation in an electronic corpus. The networks move out from a central node exploring the significant word associations of both base and collocate. This is done by measuring statistical significance using mutual information. Working within a carefully-constructed corpus based on a discourse community, the networks gradually reveal the significant lexis of the field and collocational combinations for more complex structures. These networks are currently being used for the extraction of headwords in a specialised pedagogical dictionary.

La théorie des réseaux collocationnels est basée sur une définition textuelle du phénomène de collocation dans un corpus électronique. Les réseaux se construisent à partir d'un nœud central en explorant les associations significatives entre la base et ses collocats. Ceci repose sur une mesure de signifiance statistique à l'aide de l'information mutuelle. Les réseaux, extraits d'un corpus de communauté de discours soigneusement élaboré, révèlent le lexique significatif et les combinaisons collocationnelles participant à des structures plus complexes. Ces réseaux sont actuellement exploités pour l'extraction des entrées destinées à un dictionnaire pédagogique spécialisé.

INDEX

Mots-clés: biologie, collocation, corpus, lexicographie, réseau (collocationnel)
Keywords: biology, collocation, corpus, lexicography, network (collocational)

AUTHOR

GEOFFREY WILLIAMS

Geoffrey Williams est professeur des universités à l'Université de Bretagne Sud, Lorient où il enseigne la linguistique de corpus et dirige le département d'Ingénierie du document. Il a auparavant enseigné l'anglais en LEA. Il a également enseigné la linguistique de corpus en licence et en DEA des Sciences du langage à l'Université de Nantes. Il est membre de nombreuses associations internationales en linguistique de corpus et en lexicographie. Il fait partie du HCTI, Héritages et Constructions dans le Texte et l'Image, laboratoire de recherche co-habilité avec l'UBO. Geoffrey.Williams@univ-ubs.fr