Sylviane Granger, Magali Paquot, Paul Rayson

Extraction of multiword units from EFL and native English corpora. The phraseology of the verb 'make'

1. Introduction

Phraseology has traditionally been on the fringes of foreign language learning and teaching, but it is now fast becoming one of its major foci of interest and activity. This development is rooted in a decade or more of feverish corpus-based research activity, which has served to establish phraseology as a major component of language, taking in both lexis and grammar and thus helping to widen its scope. While previous phraseological studies had been firmly focused on opacity and fixedness, corpus research showed that the most frequent multiword units, and hence arguably the most useful to teach, were in fact at the more transparent and variable end of the phraseological continuum. This has led to a shift of interest from pure or quasi-idioms to units such as collocations and frequently recurring sequences with compositional meaning which allow for a great deal of substitution and grammatical change.

The importance of phraseology in foreign language learning is now widely recognized (Nation 2001) and a range of methods for integrating it into the classroom has been designed (Lewis 1997 & 2000). There is still a major hurdle to overcome before phraseology can truly come into its own, however, viz. the lack of reliable descriptions of multiword units. When it comes to assessing the habitual company words keep, intuition is useful but not sufficient. What is needed are large electronic collections of native texts from which multiword units can be extracted and analysed. In addition, if the investigation has an applied aim, i.e. to improve pedagogical tools and methods, information from learner corpora, i.e. electronic collections of texts produced by foreign or second language learners, is also valuable. Native corpora provide crucial descriptions of multiword units, but learner corpora are a window into learners' difficulties with multiword units and help uncover the patterns of overuse, underuse and misuse of these units that characterize learner language (Nesselhauf 2004).

In this article we report on the preliminary stage of a 5-year research project at the University of Louvain, which aims to clarify the role of phraseology in foreign language learning and teaching using a broad range of native and learner corpora. As we are dealing with large amounts of data, our preliminary objective has been to find ways of automating the extraction of multiword units. Here we report the results of one particular extraction method, that of automatic filtering with a pre-established list. We highlight its advantages and disadvantages and suggest complementary approaches. To test the method, we focused on the verb *make*, which, like most high-frequency verbs, displays a high rate of phraseological use (Howarth 1996: 118-119) and always proves a major stumbling-block for learners (Nesselhauf 2004).

2. A pre-established look-up list

For twenty years or more, multiword units have been attracting wide attention from researchers in subject areas as diverse as lexicography (Sinclair 1987; Heid 1994), terminology (Bourigault 1993; Daille 1994), language teaching (Lewis 1997; 2000), psycholinguistics (Wray 2002; Schmitt 2004), and natural language processing (Sag & al. 2002; Bond & al. 2003; Dias & al. 2004). Their perspectives on multiword units may have differed, but these researchers have all come up against two recurrent and thorny issues in their multiword unit research: how to define multiword units and how to extract them from textual data. In this article we focus on the issue of extraction, which is a major challenge for automated analysis as "in so many cases [multiword units] are not predictable, not common, not fixed formally, and not fixed temporally" (Moon 1998: 51).

Most previous approaches to multiword unit extraction have been either purely statistical (Dunning 1993; Pedersen 1996) or have combined statistical and linguistic information. In studies of the latter type, statistical measures are applied on pre-selected candidates that generally correspond to a combination of part-of-speech tags, e.g. adjective + noun, verb + noun (Smadja 1993; Daille 1994) or to a syntactic pattern, e.g. verb + object (Church & al. 1991). Although the success of combined approaches largely depends on the accuracy of the annotation tools used, the results yielded by such techniques are quite promising. Kilgarriff & al. (2004), for example, developed the *Sketch Engine*, a corpus tool which generates word sketches, i.e. corpus-based summaries of a word's grammatical and collocational behaviour. For a verb, recurrent subjects, objects, modifying adverbs, prepositions, prepositional objects and typical grammatical structures all feature in separate collocate lists. A sample of the Word Sketch[1] for *make* is given in Table 1.

Table 1: A sample of the Word Sketch for *make* (v), BNC frequency = 213469

	Freq	Stat	**Object**	Freq	Stat	**Modifier**	Freq	Stat
~ by	**2912**	15.8	Mistake	1738	31.1	sure	2322	33.6
~ in	**4307**	15.0	Attempt	1695	28.8	easier	661	25.3
~ from	**1620**	13.2	Sense	2466	28.5	clear	280	21.4
material	53	12.2	Decision	3356	27.8	public	309	20.0
plastic	17	11.8	Debut	621	27.4	good	217	18.9
grape	24	11.7	Progress	1332	26.6	easy	232	17.5
steel	19	11.2	Contribution	1331	26.4	plain	167	14.4
~ of	**3312**	11.0	Use	2336	26.4	worse	69	14.2
wood	93	10.8	Effort	1703	26.2	harder	108	13.8
plastic	49	10.6	Difference	1845	25.5	abundantly	53	12.8

Statistical methods require large amounts of data as it is only in large corpora that "you get repetitions of multiword choices in combination" (Sinclair 2001: x). Kilgarriff & al. (2004) used the 100-million word British National Corpus[2] to generate their word sketches. This is a problem when it comes to using learner corpora for phraseology research, as although large in comparison to the types of data regularly used in second language acquisition studies, most learner corpora are quite limited in comparison to current native corpora. In this study, we have therefore opted for a method which does not rely on fully automatic multiword unit extraction but is instead based on a manually compiled list, converted into a machine-readable format for subsequent automatic extraction, a method inspired by Moon's (1998) study of variability in fixed expressions and idioms (FEIs). After compiling a look-up list of FEIs mainly based on the *Collins Cobuild English Language Dictionary*, Moon used sophisticated search patterns to retrieve FEIs and their variants from part-of-speech (POS) tagged and semantically annotated corpora. For example, a search for the lemma *spill* in collocation with a word denoting food or drink within a span of 5 words allowed her to retrieve occurrences of the idiom *spill the beans* as well as possible variations of this particular multiword unit like in the following sentence: *Alarmed by this turn of events the Grand Master of the Freemasons (Nosher Powell) employs two inept hit men, Mig (Tim McInnerny) and Mog (Alexei Sayle), to kill Bertoli before he can* ***spill the pasta****.* (quoted in Moon 1998: 51).

As the efficiency of this approach largely hinges on the completeness of the look-up list, our first task consisted in compiling a comprehensive list of multiword units with the verb *make* taken from corpora, current monolingual learners' dictionaries and dictionaries of collocations and idioms. The list includes 338 word combinations of various types, especially collocations and, to a lesser extent, idioms. The collocations are essentially *make* + noun sequences, where *make* is used as a delexical verb (*make a statement, make a concession, make an offer, make a threat*).

The term idiom is used to cover the more fixed and/or opaque combinations such as *make it, make oneself at home* or *make something of*. Phrasal verbs are excluded from the analysis.

Although compiling such a list is very time-consuming, it has the considerable advantage of being re-usable. For example, De Cock & Granger (2004) used this list to compare the phraseological coverage of five monolingual learners' dictionaries of English. Similar studies could be carried out to assess the phraseological coverage of English-French bilingual dictionaries or ELT textbooks.

Moon used a concordancer to extract fixed expressions and idioms. As concordancers can only search for occurrences of one multiword unit at a time (cf. Colson, this volume), she needed to build a new query for each unit (cf. the example of *spill the beans* above). She did however highlight several benefits of more automatic methods such as preprocessing routines for recognizing FEIs in text, which would make it "possible to investigate more robustly the distribution of FEIs, in specific genres, varieties, or idiolects" (ibid: 56). We decided to follow this route and used the *Wmatrix* corpus processing tool (Rayson 2003) to automate the identification of multiword units with *make* in native and learner corpora.

3. Extracting multiword units

Wmatrix is a web-based corpus processing environment which allows users to annotate corpora both grammatically and semantically[3]. It makes use of two lexicons: (a) a single-word lexicon and (b) a large multiword expression lexicon, which currently contains 18,710 patterns. Multiword expressions can be phrasal verbs (*stubbed out*), noun phrases (*riding boots*), proper names (*United States of America*), collocations (*make a decision*) or idioms (*living the life of Riley*). They are described as regular expressions or templates, i.e. sequences of words, parts of words, grammatical and/or semantic categories used to match similar patterns of text and extract them. Thus, the template *"ahead_II21 of_II22 *_APPGE time_NN1*" will identify all occurrences of the complex preposition (II)[4] *ahead of* directly followed by a possessive pronoun (APPGE) and the singular noun (NN1) *time* and will consequently retrieve all instances of the multiword unit *ahead of one's time.* As users can compile their own lexicons and use them instead of or in addition to Wmatrix's inbuilt lexicons, we turned all the phraseological uses of *make* in our look-up list into templates conforming to Wmatrix's pattern-matching syntax. We used the following three types of templates:

1) Highly constrained templates for fully fixed multiword units, e.g. *ma[dk]*_V* ends_NN2 meet_VVI,* will identify all forms of the verb *make*[5] directly followed by the plural noun (NN2) *ends* and the infinitive (VVI) *meet*.
2) Slightly less constrained templates for multiword units which are quite fixed but allow for some well-defined variation. In *ma[kd]*_V* {JJ, D*, AT*} sense_NN1*, braces are used to enclose alternatives in the search pattern. The template allows for the extraction of the multiword unit *make sense* and its variants *make no sense, little sense, more sense, common sense,* etc., as *make* can be separated from *sense* by an adjective (JJ), all kinds of determiners (D*) or articles (A*).
3) Highly flexible templates for multiword units such as collocations with a potentially high degree of variability. For example, the template *ma[dk]*_V* {A*, D*, JJ*, M*, NN*, IO, CC*, R*, UH} mistake*_N**[6] allows for the extraction of a wide range of patterns, e.g. *a person has* ***made*** *a* ***mistake****, they will* ***make*** *that* ***mistake****, they* ***made*** *a huge* ***mistake****, they are* ***making*** *the same stupid* ***mistake***, etc. For each multiword unit in this category, another template is necessary to describe patterns where the collocate precedes the verb *make* as in *a lot of procedural* ***mistakes*** *were* ***made*** or ***mistakes*** *that were* ***made*** *by the Public Prosecutor*.

As part of the data analysed come from learners of English and therefore contain errors, it is essential to allow for more flexibility in the templates than would be necessary if only native data were analysed. Learners make errors when they use multiword units: they use incorrect prepositions (*make fun *on* instead of *make fun of*), they insert or fail to insert determiners in multiword units (*make *plenty use of*), they use the plural when they should use the singular (*make *progresses* instead of *make progress*), etc. Narrowly defined templates such as *ma[kd]*_V* {JJ} progress_NN1* or *ma[kd]*_V* fun_NN1 of_IO* would not retrieve learners' erroneous uses of the multiword units described above. As we are equally interested in the fact that learners actually use multiword units as in the fact that they sometimes make erroneous use of them, it was essential to relax some of the templates. For example, our template for *make fun of* does not include the preposition *of* so that instances of the multiword unit containing an incorrect preposition can be retrieved. Similarly, our template for the multiword unit *make progress* allows both for singular and plural forms of the noun *progress* and article inclusion.

4. Results and evaluation

We tested this approach on six comparable corpora of about 200,000 words of argumentative writing: one native corpus, the LOCNESS (*Louvain Corpus of Native Speaker Essays*) corpus and five learner corpora, all subparts of the *International Corpus of Learner English* (ICLE)[7] (cf. Granger 1998a; 2003). The learner corpora consist of essay writing by higher intermediate to advanced EFL learners of French

(ICLE-FR), Spanish (ICLE-SP), Italian (ICLE-IT), German (ICLE-GE) and Dutch (ICLE-DU) mother tongue backgrounds. The results show that automatic filtering of multiword units with the help of a pre-established list has three major advantages: (1) the method has good precision and recall rates, (2) it is quick and robust and (3) it provides interesting insights into learners' use of phraseological units.

(1) The method has good precision and recall rates

The reliability of automatic extraction tools is usually assessed on the basis of their recall and precision rates (Salton 1989: 248). Recall is "the proportion of relevant materials retrieved" and precision is "the proportion of retrieved materials that are relevant." Applied to multiword units, the recall rate is the proportion of manually retrieved multiword units retrieved automatically. The precision rate, on the other hand, is the proportion of automatically retrieved sequences that qualify as bona fide multiword units. To calculate these formulae, we needed to manually analyse all the concordance lines of the verb *make* in the six corpora and classify them as free combinations or multiword units.

The recall rate for the six corpora described above is quite satisfactory. For each corpus, approximately 80% of the multiword units manually identified were also extracted automatically. A close examination of the 20% of multiword units not retrieved by the system shows that it would be possible to improve the recall rate in two ways. First, the look-up list is still not sufficiently exhaustive and missing multiword units such as *make mincemeat of* could be added. Secondly, some templates are still too restrictive and could be relaxed in a number of ways. For example, the multiword unit *make a statement* is not retrieved in '*this statement, made hundreds of years ago'* as the template does not allow for the insertion of a comma between *statement* and *made*. Similarly, the multiword unit *make an attempt* is not identified in '*an active attempt to remedy the situation must be made*' as the template used to extract passives does not allow for the infinitive marker *to* (tagged TO) between the two elements of the multiword unit.

The precision rates in the six corpora are also quite satisfactory: 80-90% of the word combinations extracted by *Wmatrix* are bona fide multiword units (cf. Table 2). Overgeneration, which was relatively infrequent, can be illustrated by the following two sentences, which were retrieved by the system although they do not contain the multiword units *make good* and *make a reduction*:

Hoederer believes anarchists ***make good*** *friends.*
This ***reduction made*** *things accessible.*

Table 2: Precision rate of multiword unit extraction by Wmatrix in 6 corpora

	ICLE-DU	ICLE-GE	LOCNESS	ICLE-FR	ICLE-SP	ICLE-IT
Precision rate	93%	89%	87%	84%	80%	78.4%

(2) The method is quick and robust

Once the list has been compiled, it is very easy to compare the use of multiword units in various corpora and assess the statistical significance of the frequency differences. One particularly interesting measure which can be calculated and compared across corpora on the basis of *Wmatrix*'s results is the phraseological rate, i.e. the proportion of phraseological uses of a verb. The phraseological rate is computed by dividing the number of phraseological uses of a verb by the total number of instances of this particular verb in a corpus.

Table 3: Phraseological rate of *make* in LOCNESS and 5 ICLE subcorpora

Corpus	Nr of automatically extracted multiword units with 'make' (based on post-edited Wmatrix data)	Nr of occurrences of 'make'	Phraseological rate of 'make'
ICLE-DU	242	631	38%
ICLE-GE	137	422	32.4%
LOCNESS	202	639	31.6%
ICLE-FR	137	483	28.4%
ICLE-SP	123	556	22.1%
ICLE-IT	116	575	20%

As shown in Table 3, the phraseological rate of *make* differs quite markedly across the six corpora. The Dutch and German corpora display quite high phraseological rates, while the three Romance corpora – especially the Spanish and Italian ones - display much lower rates. The native speaker corpus, LOCNESS, is situated in the middle. These quantitative results need to be thoroughly investigated in the light of "a very complex interplay of factors: developmental, teaching-induced and transfer-related, some shared by several learner populations, others more specific" (Granger 2004: 135), which lead learners to overuse some multiword units and underuse others (see Granger 1998b for studies illustrating this interplay of factors). Differences in proficiency level are certainly also partly responsible for variations in phraseological rate. It is striking to note that the phraseological cline perfectly matches the proficiency cline brought to light by an independent assessment of the five subcorpora in terms of the European Framework of Reference.[8]

(3) The method provides interesting insights into learners' use of phraseological units

The third advantage of this approach is that it provides very good insights into learners' use of multiword units, both correct and incorrect. The data show that the Spanish and Italian learners and to a lesser extent, the French learners do not only display a lower phraseological rate of the verb *make*, they also display many more errors than the German and Dutch learners. Here are a few examples:

*The political class makes ****a large*** *use of words. They are very good in speaking and convincing people.* [ICLE-IT]
... we are showing that women are making ***[…]** *way in this men's society.* [ICLE-SP]
... the students make an appeal ****towards*** *the Ministry of Education.* [ICLE-SP]

These examples show that advanced learners are aware of a large number of phraseological uses of *make* but their knowledge is incomplete. They know that variations of *make use of* may include the adjectives *good, full* or *heavy* but seem unaware that *large* is not a possibility. Similarly, *make way* requires the presence of a possessive determiner and *make an appeal* requires the preposition *to*, not *towards*. These examples support Nesselhauf's (2003: 238) claim that "[i]t is not sufficient to merely teach the lexical elements that go together, but it is necessary to teach entire combinations including prepositions, articles, etc. (e.g. knowing that *pass* can combine with *judgement* is less useful, (…), than knowing that it is *pass judgement on* and not **pass one's judgement* or **pass judgments* or **pass judgement about* or indeed anything else)."

5. Conclusion

As rightly pointed out by Nation (2001: 56), research on phraseology "can only be done to a certain point by computer". The method presented here is a good trade-off between manual and automatic analysis. After an initial and necessarily manual stage of compiling a list of multiword units and converting it into the appropriate format, all occurrences of the multiword units can be extracted automatically from a range of corpora. Although the reliability of the technique is already satisfactory, we have shown that its recall can still be improved.

It is important to bear in mind, however, that any approach based on a pre-established list has the inherent limitation of being deterministic, i.e. searches will "only report what has been sought, not what should or could have been looked for" (Moon 1998: 49). This means that the method will extract native-like multiword units used by learners but will not be capable of retrieving learner idiosyncratic units such as the following ones, all extracted from ICLE-FR: *make abstraction of,* *make influence on, make a step, make a balance, make opposition to, make a critic (to), make usage of, make one's mind, make a problem (to), make a reflexion, make an end to, make part of.* These multiword units are not used by native speakers and are therefore not part of the look-up list. Quite a few are undoubtedly L1-related as they have direct translation equivalents in French, e.g. *make abstraction of* = *faire abstraction de*, *make a step* = *faire un pas*, *make part of* = *faire partie de, make usage of* = *faire usage de*. Others seem to result from a confusion within the target language, e.g. *make an end to* instead of *put an end to* or *make one's mind* instead of *make up one's mind*.

In the future, we intend to compare learner-specific multiword units across a wide range of learner corpora to assess the role of transfer in phraseological errors. Different extraction methods will be needed to achieve this and we are currently testing a method based on the extraction of lexical dependencies[9] by means of the shallow parser Syntex (Fabre & Bourigault 2001). We are also investigating the possibility of combining grammatical and syntactic information with semantic annotation. Ideally, such an approach should allow us to retrieve, for example, all abstract nouns used as the object of the verb *make*.

Our study shows that there is not just one but many methods of retrieving multiword units and that these can be used independently or in combination. Each method has its advantages and limitations and should always be carefully selected according to the size of the corpus analysed, the type of multiword unit examined and the objective of the study.

Acknowledgements

We gratefully acknowledge the support of the Communauté française de Belgique, which funded this research within the framework of the 'Action de recherche concertée' project entitled 'Foreign Language Learning: Phraseology and Discourse' (No. 03/08-301).

Notes

[1] Word Sketches were first used in the production of the *Macmillan English Dictionary for Advanced Learners* as a means of helping lexicographers cope with the overflow of data as they "present dictionary-writers with a pre-digested outline of the most important and relevant facts about a word" (Kilgarriff & Rundell 2002: 808).

[2] http://www.natcorp.ox.ac.uk/

[3] See http://www.comp.lancs.ac.uk/ucrel for a description of the CLAWS7 part-of-speech and UCREL Semantic Analysis System (USAS) tagsets.

[4] *Ahead of* is treated as a single preposition and therefore receives the following tags: *ahead_II21 of_II22*. II stands for a general preposition. The first of the two digits indicates the number of graphemic words in the sequence, and the second digit the position of each graphemic word within that sequence.

[5] The wildcard * stands for any character.

[6] A* = articles and possessive pronouns, D* = determiners, JJ* = adjectives, M* = numbers, NN* = nouns, IO = preposition 'of', CC* = coordinating conjunctions, R* = adverb, UH = interjection

[7] A part of the ICLE corpus amounting to 2.5 million words has recently been released in CD-Rom format (cf. Granger et al. 2002).

[8] http://www.coe.int/T/E/Cultural_Co-operation/education/Languages/Language_Policy/Common_Framework_of_Reference/default.asp

[9] This method is quite similar to Kilgarriff & al. (2004) but we do not rely on statistical measures, as our corpora are too small.

Bibliography

Altenberg, B. (1998) On the Phraseology of Spoken English: The Evidence of Recurrent Word-Combinations. In Cowie, A.P. (ed.) Phraseology: Theory, Analysis and Applications. Oxford: Oxford University Press, 101-122.

Bond F., Korhonen A., McCarthy D. & Villavicencio A. (2003) (eds) Proceedings of the Workshop on Multiword Expressions: Analysis, Acquisition and Treatment at ACL 2003, 41st Annual Meeting of the Association for Computational Linguistics, Sapporo, Japan, July 12, 2003, http://acl.ldc.upenn.edu/acl2003/mwexp/cover.pdf

Bourigault D. (1993) Analyse syntaxique locale pour le repérage de termes complexes dans un texte. Traitement Automatique des Langues, vol. 34/2, 105-117.

Church K., Gale W., Hanks P. & Hindle D. (1991) Using Statistics in Lexical Analysis. In Zernik U. (ed.) Lexical Acquisition: Exploiting online resources to build a lexicon. Laurence Erlbaum Associates, Hillsdale, NJ, 115-164.

Daille, B. (1994) Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques. Thèse de Doctorat en Informatique Fondamentale. Université Paris 7.

De Cock, S. & Granger, S. (2004) High Frequency Words: the Bête Noire of Lexicographers and Learners Alike. A close look at the verb 'make' in five monolingual learners dictionaries of English. In Williams G. and S. Vesssier (eds) Proceedings of the Eleventh EURALEX International Congress. Lorient: Université de Bretagne-Sud: 233-243.

Dias, G.H., Pereira Lopes, J.G. & Vintar, Š. (2004) (eds) Proceedings of the 2004 Workshop on Methodologies and Evaluation of Multiword Units in Real-World Applications, 25 May 2004, Lisbon, Portugal, http://memura2004.di.ubi.pt/main-memura-proceedings-vInternet.pdf

Dunning, T. (1993) Accurate methods for the statistics of surprise and coincidence. Computational Linguistics 19(1):61-74.

Fabre C. & Bourigault D. (2001) Linguistic clues for corpus-based acquisition of lexical dependencies. Proceedings of the 2001 Corpus Linguistics Conference, UCREL Technical Papers 13, Lancaster University, 176-184.

Granger, S. (1998a) The computerized learner corpus: a versatile new source of data for SLA research. In Granger S. (ed.), 3-18.

Granger, S. (1998b) (ed.) Learner English on Computer. London & New York: Addison Wesley Longman.

Granger, S. (2003) The International Corpus of Learner English: A New Resource for Foreign Language Learning and Teaching and Second Language Acquisition Research. TESOL Quarterly 37/3, 538-546.

Granger, S. (2004) Computer learner corpus research: current status and future prospects. In Connor U. and T.A. Upton (eds) Applied Corpus Linguistics: A Multidimensional Perspective. Amsterdam & Atlanta: Rodopi, 123-145.

Granger, S., Dagneaux, E. & Meunier, F. (eds) (2002) The International Corpus of Learner English. CD-ROM and Handbook. Louvain-la-Neuve : Presses universitaires de Louvain. Available from http://www.i6doc.com

Heid, U. (1994) On ways words work together – research topics in lexical combinatorics. In Martin W., Meijs W., Moerland M., ten Pas E., van Sterkenburg P. & Vossen P. (eds) Euralex 1994 Proceedings, Amsterdam, 226-258.

Howarth, P. (1996) Phraseology in English Academic Writing. Tübingen: Niemeyer.

Kilgarriff A. & Rundell M. (2002) Lexical Profiling Software and its Lexicographic Applications: a case study. In Braasch A. & Povlsen C. (eds) Proceedings of the Tenth EURALEX International Congress, EURALEX 2002, Copenhagen. Copenhagen: Center for Sprogteknologi, Københavns Universitet, 807-818.

Kilgarriff A., Rychly P., Smrz P. & Tugwell D. (2004) The Sketch Engine. In Williams G. and S. Vessier (eds) Proceedings of the Eleventh EURALEX International Congress. Université de Bretagne-Sud: Lorient, vol. I, 105-116.

Lewis, M. (1997). Implementing the Lexical Approach. Putting Theory into Practice. Boston: Heinle.

Lewis, M. (ed.) (2000) Teaching Collocation. Further Developments in the Lexical Approach. Boston: Heinle.

Moon, R. (1998) Fixed Expressions and Idioms in English. A Corpus-Based Approach. Oxford: Clarendon Press.

Nation, I.S.P. (2001) Learning Vocabulary in Another Language. Cambridge: Cambridge University Press.

Nesselhauf, N. (2003) The use of collocations by advanced learners of English. Applied Linguistics 24/2: 223-242.

Nesselhauf, N. (2004) Collocations in a Learner Corpus. Amsterdam & Atlanta: Benjamins.

Pedersen, T. (1996) Fishing for Exactness. Proceedings of the South-Central SAS Users Group Conference, Austin, TX, Oct. 27-29, 1996, 188-200.

Rayson, P. (2003) Matrix: A statistical method and software tool for linguistic analysis through corpus comparison. PhD dissertation, Lancaster University. http://www.comp.lancs.ac.uk/ucrel/wmatrix/

Sag I.A., Baldwin T., Bond, F., Copestake A. & Flickinger, D. (2002) Multiword expressions: A pain in the neck for NLP. In Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002), Mexico City, 1-15.

Salton, Gerard (1989) Automatic Text Processing. Reading: Addison-Wesley.

Schmitt, N. (2004) (ed.) Formulaic sequences: Acquisition, processing and use. Amsterdam: John Benjamins.

Sinclair, J. (ed.) (1987) Looking up: an account of the COBUILD project in lexical computing and the development of the Collins COBUILD English Language Dictionary. London: Collins ELT.

Sinclair, J. (2001) Preface. In Ghadessy M., Henry A. & Roseberry R. (eds) Small corpus studies and ELT. Amsterdam/Philadephia: John Benjamins, vii-xv.

Smadja, F. (1993) Retrieving Collocations from Text: Xtract. Computational Linguistics, 19(1):143-177.

Wray, A. (2002) Formulaic language and the lexicon. Cambridge UK: Cambridge University Press