Evaluation of alternative association measures for extraction of terminology based on a large Norwegian corpus

Gisle Andersen

NHH Norwegian School of Economics

Summary
Multiword expressions are words that co-occur so often that they are perceived as a linguistic unit (Stubbs 2007). Identifying them correctly is important for a variety of tasks within terminology, lexicography and language technology. This paper presents a methodology for the systematic and corpus-driven study of multiword expressions in Norwegian. It reports on a series of experiments using a variety of different association measures in order to identify multiword expressions that occur in a large corpus consisting of Norwegian newspapers (Andersen & Hofland forthcoming). The output of each association measure is a ranked list of bigrams and trigrams in the corpus. The value of different association measures for terminology purposes is assessed by considering the relevance and salience of ranked candidates among the bigrams and trigrams in the data. It is shown that the association measures differ greatly in their ability to pick out relevant term candidates. The paper also briefly evaluates the corpus itself and its relevance for terminology work (Kristiansen Forthcoming).

1 Introduction

Multiword expressions (MWEs) are pairs or longer combinations of word which co-occur more often than would be predicted by chance. The creation of databases for terminology and lexicography require systematic approaches to the correct identification of MWEs. The category of MWEs incorporate an array of structurally and conceptually different items, and the correct identification and segmentation of MWEs is necessary for a variety of purposes in natural language processing, terminology, lexicography and related disciplines. For instance, technical terminology is very often realised as phrasal units. Examples of recent date occurring in Norwegian are *acute respiratory distress syndrome*, *predatory pricing*, *tabula gratulatoria* and *spinal muskelatrofi*, some of which also illustrate terms realised as anglicisms or other foreign items. For more general lexicographical purposes it is also necessary to develop procedures for identifying various types of phrasal idioms such as *sakens kjerne* 'the core of the matter' and various types of idiomatic formulae. Moreover, the performance of language processing tools such as word class taggers and grammatical parsers (treebanks) can be enhanced with knowledge of which forms constitute different types of grammatical MWEs, for instance phrasal prepositions such as *på grunn av* 'because of', adverbs such as *i tide* 'on time', etc. These should be segmented as phrasal units and not processed further by the tagger.

The present paper is concerned with testing alternative methods for identifying MWEs in a large corpus of Norwegian text. This is a task which in principle amounts to sorting the wheat from the chaff, that is to say, distinguishing between, on the one hand, "frequent collocations [...] which are of no real interest to lexicography" (Atkins & Rundell 2008: 166) and on the other hand "phrases which have some degree of idiomatic meaning or behaviour (ibid.). Although the proposed distinction is drawn from a lexicography setting, the same holds true for terminological purposes, in that for multiword terms, a phrasal unit, commonly an adj+N

or N+N combination, will be taken to represent a single concept. Similarly, Biber argues for "the existence of two underlying linguistic constructs: 'multi-word lexical collocations' versus 'multiword formulaic sequences'" (Biber 2009: 227). Endorsing this fundamental distinction, I shall be concerned only with the former type, which includes multiword technical terms generally consisting of a sequence of lexical words, while the latter are high-frequency sequences that may well be perceptually salient but are uninteresting from a terminological/lexicographical point of view because they represent recurrent phrase fragments or grammatical patterns, such as *it is the*. These 'lexical bundles', as Biber calls them, are 'the most frequent recurring sequences of words' (Biber 2006: 133); they are 'usually not idiomatic in meaning, and they are usually not complete grammatical structures' (Biber 2006: 134)

It is clear that absolute frequencies are in no way able to capture the word associations that are important for terminological or lexicographical work. The alternative to relying on absolute frequencies is to use a statistical measure of association, like the Mutual Information (MI) score. Association measure (AM) scores reflect the collocational strength of pairs and longer combinations of words by comparing the frequency of a word combination to the overall frequencies of each of the individual words (Stubbs 1995). As the individual word frequencies become higher, it becomes more likely that the word combination would occur just by random chance, and therefore the combination has less importance. Like Biber (2009), I adopt a radical corpus-driven approach to identify the most common multiword patterns in a corpus. The paper reports on research that applies various statistical association measures to word sequences from the Norwegian Newspaper Corpus (NNC), which is a large monitor corpus consisting of Norwegian newspapers (avis.uib.no; (Andersen & Hofland Forthcoming). This paper concurs with Fontenelle, who says that "the notions of relevance and salience are also crucial issues [...] which should not be neglected when discussing the usefulness of such tests as mutual information [...] or t-scores" (Fontenelle 2002: 221). The tendency of words to co-occur in prefabricated chunks of language has attracted a lot of attention recently (Renouf & Sinclair 1991; Sinclair 1991; Renouf 1996; Sinclair 2004; Sandford 2008). "The availability of very large corpora has made it possible to shed some new light onto the concept of collocation and statistical tools are now the norm rather than the exception" (Fontenelle 2002: 219). The task of having to investigate thousands of concordances to extract the most relevant facts about the behaviour of individual lexical items is untenable for most purposes. One of the questions which arise is how to extract such collocational combinations in the most optimal way. It is becoming increasingly common to use large corpora and apply sophisticated statistical techniques on them in order to identify patterns which are "salient, relevant and typical patterns" (Tognini-Bonelli 2001: 221).

The primary purpose of the current study is therefore to present a research methodology for the identification of MWEs and to investigate which AMs are the better for extracting terminology. I investigate a set of ranked lists of two- and three-word sequences (bigrams and trigrams) in terms of the tendency of words to collocate. Preliminary results (Lyse & Andersen Forthcoming) show that some statistical measures favour relatively frequent MWEs (e.g. *i motsetning til* 'as opposed to'), whereas other measures favour relatively low-frequent units, which typically comprise loan words (*de facto*), technical terms (*notarius publicus*) and phrasal anglicisms (*practical jokes*). I evaluate the relevance of each of these measures for terminology, lexicography and language technology purposes. The extracted forms should be viewed as term candidates rather than *bona fide* terms, and ought to be the subject of subsequent inspection of a terminologist or lexicographer. I shall make no attempt to distinguish or predict the association of these terms to specific domains – this would have to

be done in a follow-up study. Some of the extracted terms are relevant for lexicographical purposes while other categories are clearly terminological. I restrict my account to the investigation of bigrams and trigrams, but longer n-grams have also been extracted and to some extent studied in the NNC project. The paper also serves a secondary purpose of assessing the relevance of a large corpus of written Norwegian general newspaper text to the field of terminology (Kristiansen forthcoming).

2 Material and method

The work on MWEs in the NNC has taken place in several steps:

1. Compile a large (1 bn word) Norwegian corpus of newspapers from the web
2. Calculate all the 1-5 gram statistics
3. Rank the bigrams and trigrams according to different association measures (9 AMs used for bigrams; 4 AMs used for trigrams)
4. Manually identify terminologically/lexicographically salient MWEs
5. Semi-automatic consistency check of manually classified items
6. Calculate statistics and evaluate

The present paper is mainly concerned with the last three steps and reports the manual classification of highly ranked bigrams and trigrams, specifically the top 500 of each ranking, and the subsequent evaluation of these rankings. The first three steps are described thoroughly in a forthcoming anthology about the NNC (Andersen & Hofland forthcoming; Lyse & Andersen forthcoming). This work is in close cooperation with Knut Hofland of Uni Digital and Gunn Inger Lyse of the University of Bergen.

The following association measures were used for the bigram analysis:

Pearson's chi square (homogeneity corrected)
Log likelihood
Logarithmic Odds Ratio
Z-score-regular
Z-score-corrected
T-score
Pointwise Mutual Information
Dice coefficient
Jaccard coefficient

The following association measures were used for the trigram analysis:

Log likelihood
Poisson-Stirling
Pointwise Mutual Information
True Mutual Information

In the classification of term candidates and other relevant items, several types had to be disregarded as irrelevant for the current purposes, although highly ranked and salient in the data. These were tokens that are the result of recurrent code switching patterns, titles, product names, common quotations, etc. relevant examples being *the twain shall never meet*, *macht frei*, *formerly known* (assumed to be part of *the artist formerly known as)*, *kleine nachtmusik* or *vida loca*, assumed to be part of a song title.

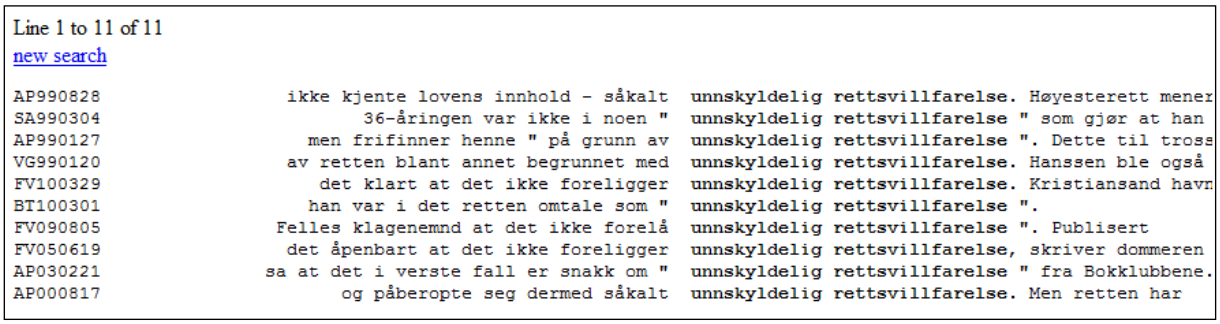
Table 1 gives a survey of the classification scheme, with examples of Norwegian term candidates and their English translations.

Table 1: survey of manual classification of relevant MWEs

anglicism MWE	*asset value*	asset value
foreign MWE	*alopecia areata*	*alopecia areata* (skin disease)
grammatical MWE	*i motsetning til*	as opposed to
idiomatic phrase	*abra kadabra*	abra cadabra
concept structure appositional phrase	*giftalgen prymnesium parvum*	the poisonous algae prymnesium parvum
term candidate	amyotrofisk lateralsklerose	amyotrophic lateral sclerosis

This classification scheme was used in the manual coding mentioned in step 4 above. Note that the classification of an item as a 'term candidate' does not necessarily mean that this item would be the preferred and standardised term in a domain-specific term base, but merely that this item appears to be a domain-specific term candidate whose termhood should be assessed further by field expert or terminologist. Moreover, note that the category 'concept structure appositional phrase' is meant to suggest that its members are not term candidates *per se*, that is, one would not consider including the salient trigram *giftalgen prymnesium parvum* as a term candidate. Such items are nevertheless relevant for term extraction purposes; they are typically composed of an appositional N+N structure whose first component, the definite NP *giftalgen* represents a superordinate concept while the last part *prymnesium parvum* is a term designating a subordinate concept to the former.

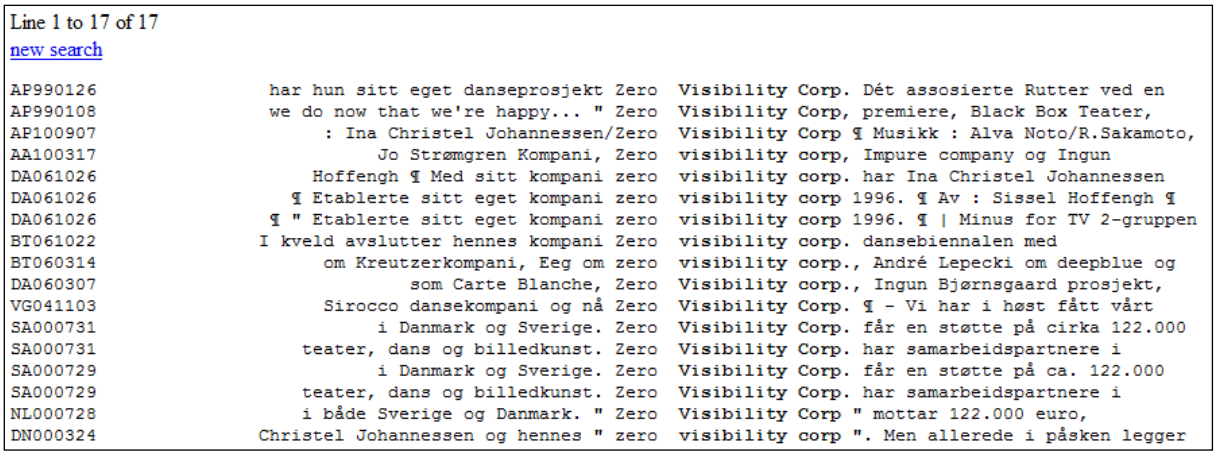
After having classified the top 500 bi- and trigrams for each ranked list, I performed a semi-automatic check of the consistency of the manual annotation of the various categories. This was done by means of a specially developed perl script which indentified inconsistencies by paired comparison of the manually annotated files. The inconsistencies were subsequently manually checked and eliminated. This manual check in some cases required lookup of unfamiliar phrases. The check was firstly done in the corpus itself, using concordance view of the phrase in question, as seen in Figure 1.


Line 1 to 11 of 11
new search

AP990828	ikke kjente lovens innhold - såkalt	**unnskyldelig rettsvillfarelse**. Høyesterett mener
SA990304	36-åringen var ikke i noen "	**unnskyldelig rettsvillfarelse** " som gjør at han
AP990127	men frifinner henne " på grunn av	**unnskyldelig rettsvillfarelse** ". Dette til tross
VG990120	av retten blant annet begrunnet med	**unnskyldelig rettsvillfarelse**. Hanssen ble også
FV100329	det klart at det ikke foreligger	**unnskyldelig rettsvillfarelse**. Kristiansand havn
BT100301	han var i det retten omtale som "	**unnskyldelig rettsvillfarelse** ".
FV090805	Felles klagenemnd at det ikke forelå	**unnskyldelig rettsvillfarelse** ". Publisert
FV050619	det åpenbart at det ikke foreligger	**unnskyldelig rettsvillfarelse**, skriver dommeren
AP030221	sa at det i verste fall er snakk om "	**unnskyldelig rettsvillfarelse** " fra Bokklubbene.
AP000817	og påberopte seg dermed såkalt	**unnskyldelig rettsvillfarelse**. Men retten har

Figure 1: Concordance view of the bigram 'unnskyldelig rettsvillfarelse'

This manual check made it possible to establish that *unnskyldelig rettsvillfarelse* is indeed a term used in legal language, and similarly to establish that *visibility corp* should be excluded because it is part of a dance company name:



Line 1 to 17 of 17
new search

AP990126	har hun sitt eget danseprosjekt Zero	Visibility Corp. Dét assosierte Rutter ved en
AP990108	we do now that we're happy... " Zero	Visibility Corp, premiere, Black Box Teater,
AP100907	: Ina Christel Johannessen/Zero	Visibility Corp ¶ Musikk : Alva Noto/R.Sakamoto,
AA100317	Jo Strømgren Kompani, Zero	visibility corp, Impure company og Ingun
DA061026	Hoffengh ¶ Med sitt kompani zero	visibility corp. har Ina Christel Johannessen
DA061026	¶ Etablerte sitt eget kompani zero	visibility corp 1996. ¶ Av : Sissel Hoffengh ¶
DA061026	¶ " Etablerte sitt eget kompani zero	visibility corp 1996. ¶ | Minus for TV 2-gruppen
BT061022	I kveld avslutter hennes kompani Zero	visibility corp. dansebiennalen med
BT060314	om Kreutzerkompani, Eeg om zero	visibility corp., André Lepecki om deepblue og
DA060307	som Carte Blanche, Zero	Visibility corp., Ingun Bjørnsgaard prosjekt,
VG041103	Sirocco dansekompani og nå Zero	Visibility Corp. ¶ - Vi har i høst fått vårt
SA000731	i Danmark og Sverige. Zero	Visibility Corp. får en støtte på cirka 122.000
SA000731	teater, dans og billedkunst. Zero	Visibility Corp. har samarbeidspartnere i
SA000729	i Danmark og Sverige. Zero	Visibility Corp. får en støtte på ca. 122.000
SA000729	teater, dans og billedkunst. Zero	Visibility Corp. har samarbeidspartnere i
NL000728	i både Sverige og Danmark. " Zero	Visibility Corp " mottar 122.000 euro,
DN000324	Christel Johannessen og hennes " zero	visibility corp ". Men allerede i påsken legger

Figure 2: Concordance view of the bigram 'visibility corp'

In other cases there was a need to supply the manual check with internet searches, so as to establish, for instance, that *lapis lazuli* was a technical term (a type of stone or jewel). All stages of the manual work were done by the same annotator (the author).

3 Bigram analysis

The overall results of the manual inspection of the top 500 bigrams in each ranked list are given in Table 2.

Table 2: Results of manual inspection of bigrams

Association measure / Cat.	anglicism MWE	foreign MWE	gram. MWE	idiomatic phrase	appos. term phr	term cand.	SUM	% rel.
Pearsons chi sq	68	8	4	49	8	127	264	52.8 %
Log likelihood	0	0	53	0	0	1	54	10.8 %
L. Odds Ratio	67	92	0	13	3	63	238	47.6 %
Z-score-reg	91	112	2	19	5	64	293	58.6 %
Z-score-corr	95	97	2	19	5	55	273	54.6 %
T-score	0	0	46	0	0	0	46	9.2 %
Pointwise MI	67	90	0	8	6	60	231	46.2 %
Dice coeff	0	0	0	0	0	10	10	2.0 %
Jaccard coeff	0	0	0	0	0	12	12	2.4 %

There are major differences between then different measures in their ability to retrieve bigrams that are considered terminologically or lexicographically relevant. Two association measures, Jaccard and Dice, are only able to retrieve a very limited number of terminologically relevant items, amounting to a mere 2 per cent of the manually inspected ranked n-grams, including *langvarig konjunkturoppgang* 'sustained cyclical expansion' and *maritime industri* 'maritime industry'. Two measures, T-score and Log Likelihood, are particularly suited for detecting grammatical multiword expressions and not any other MWE types. The retrieved items include multiword adverbials and prepositions such as *for eksempel* 'for example', *i tillegg* 'in addition', *etter hvert* 'gradually' and *blant annet* 'among others' as well as one phrasal verb, *regne(r) med* 'take into account'. Their respective 10.8 and 9.2 per cent must be considered a high proportion of grammatical MWEs, given that this category represents closed categories, which generally can be expected to have fewer members than open categories such as nouns, which is where most terms would be included. The remaining

five AMs are all relatively successful in retrieving lexically and terminologically relevant items, ranging from 46.2 (Pointwise Mutual Information) to 58.6 per cent (Z-score regular). One of these measures, Pearson's chi square, is particularly able to pick out term candidates, including *alternative energikilder* 'alternative energy sources' and *blokkerende mindretall* 'blocking minority' as well as concept structure appositional phrases of the type *tungmetallet kadmium* 'the heavy metal cadmium', which I also consider to be relevant for term extraction purposes. The other four measures are to a lesser degree able to identify domestically based term candidates but are better than Pearson's at extracting multiword expressions (including terms) of foreign or English origin, such as *consumer confidence*, *joint ventures*, *annus horribilis* and *garam masala*.

4 Trigram analysis

The overall results of the manual inspection of trigrams are as shown in Table 3.

Table 3: Results of manual inspection of trigrams

Association measure / Cat.	anglicism MWE	foreign MWE	gram. MWE	idiomatic phrase	appos. term phr	term cand.	SUM	% rel.
Log likelihood	0	0	1	0	0	0	1	0,2 %
Poisson-Stirling	0	0	0	62	0	0	62	12,4 %
Pointwise MI	59	26	0	8	17	17	127	25,4 %
True MI	0	0	0	1	0	0	1	0,2 %

There are striking differences between the four different trigam AMs in their ability to retrieve word sequences that are of terminological or lexicographical relevance. Two of the AMs, Log-Likelihood and True Mutual Information, were unable to rank highly any relevant items, with the exception of one token each, the idiomatic phrase *grøss og gru* 'shiver and horror' (True Mutial Information) and the phrasal verb *kommer til å* 'is going to' which counts as a grammatical MWE (Log Likelihood). The Poisson-Stirling measure is highly capable of picking out one specific type, namely grammatical MWEs, as 12.4 per cent of the inspected trigrams were of this category, and no other categories were represented. Finally, Pointwise Mutual Information is a more versatile measure that is capable of picking out a variety of MWEs, totalling 25.4 per cent. Note that no types representing grammatical MWEs were picked out by this AM. This shows very clearly the need for selecting the right AM depending on the specific objectives of the term extraction or lexical acquisition. However, all the other types were identified. The multiword anglicisms include multiword terms from various domains, such as *hypertext markup language*, *deficit hyperactivity disorder*, *joint stock companies*, *checks and balances*, *frequently asked questions*, *catch and release* and *stream of consciousness*, as well as other salient multiword anglicisms of a more general nature, such as *worst case scenario* and *trick or treat*. The foreign multiword trigrams are especially culinary terms, such as *gambas al ajillo*, *spaghetti alla carbonara*, *chili con carne*, *biff chop suey*, *cafe au lait* and *pain au chocolat*, but also include terms from other domains such *as homo sapiens* and *tae kwon doe*, and also more general foreign multiwords such as *quod erat demonstrandum*, *cage aux folles* and *persona non grata*. Further, this AM picks out idiomatic phrases like the formulaic *snipp snapp snute* (used at the conclusion of fairy tales) and *bitte litte granne* 'teeny weeny bit'. From a terminological point of view it is interesting to note that a number of concept structure appositional phrases are ranked highly by this measure. Nevertheless, this measure picks out fewer term candidates than the best bigram measures, limited to 17 types, presumably because multiword terms are more often realised as

bigrams than as trigrams. The term candidates are mostly from medicine and include *viral hemoragisk septikemi, amyotrofisk lateral sklerose* and *hemolytisk uremisk syndrom*.

5 Concluding remarks

The paper shows, firstly, that a large general corpus is a surprisingly rich repository for multiword terms from a variety of fields in Norwegian. Secondly, it shows the importance of selecting the right association measure depending on the specific task one is aiming for, e.g. extracting term candidates, identifying grammatical MWEs, identifying multiword anglicisms, identifying discourse markers (Andersen 2011) or the like. Moreover, with regard to multiword expressions, there are good reasons for combining work in terminology, lexicography and natural language processing, since similar methods can be used for retrieving conceptually different structures. Many aspects of the analysis require further work, however. For instance there is a need to assess whether these findings reflect language-specific features or whether they have a wider application across languages.

References

Andersen, Gisle (2011) *Corpus-driven approaches to discourse markers in spoken data*. ISLE2 Boston.

Andersen, Gisle /Hofland, Knut (forthcoming) Building a large monitor corpus based on newspapers on the web. In Andersen, Gisle (ed.) *Exploring Newspaper Language − Corpus Compilation and Research based on the Norwegian Newspaper Corpus*.

Atkins, Sue /Rundell, Michael (2008) *The Oxford Guide to Practical Lexicography.* Oxford/New York: Oxford University Press.

Biber, Douglas (2006) *University Language: A corpus-based study of spoken and written registers*. Amsterdam/Philadelphia: John Benjamins.

Biber, Douglas (2009) A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics* 14. 275-311.

Kristiansen, Marita (forthcoming) Financial jargon in a general newspaper corpus. In: Andersen, Gisle (ed.) *Exploring Newspaper Language − Corpus Compilation and Research based on the Norwegian Newspaper Corpus*: To be published by John Benjamins.

Lyse, Gunn Inger /Andersen, Gisle (forthcoming) Collocations and statistical analysis of n-grams. In: Andersen, Gisle (ed.) *Exploring Newspaper Language − Corpus Compilation and Research based on the Norwegian Newspaper Corpus*.

Renouf, Antoinette (1996) The ACRONYM Project: Discovering the textual thesaurus. In: Percy, Carol E. Meyer, Charles F. and Lancashire, Ian (eds) *Synchronic corpus linguistics* Amsterdam/Atlanta: Rodopi.

Renouf, Antoinette /Sinclair, John M. (1991) Collocational frameworks in English. In: Aijmer, Karin and Altenberg, Bengt (eds) *English Corpus Linguistics − Studies in Honour of Jan Svartvik.* London/New York: Longman.

Sandford, Daniel (2008) Discourse and metaphor: A corpus-driven inquiry. *Corpus Linguistics and Lingustic Theory* 4. 209-234.

Sinclair, John M. (2004) *Trust the Text: Language, Corpus and Discourse.* London: Routledge.

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

Stubbs, Michael (1995) Collocations and semantic profiles: On the cause of the trouble with quantitative studies. *Functions of Language* 2. 23-55.

Stubbs, Michael (2007) An example of frequent English phraseology: distributions, structures and functions. In: Facchinetti, Roberta (ed.) *Corpus Linguistics 25 Years on.* 89-105. Amsterdam/New York: Rodopi.

Tognini-Bonelli, Elena (2001) *Corpus Linguistics at Work.* Amsterdam: John Benjamins.