Extracting Multiword Expressions with A Semantic Tagger

Scott S. L. Piao
Dept. of Linguistics and MEL
Lancaster University
s.piao@lancaster.ac.uk

Paul Rayson
Computing Department
Lancaster University
paul@comp.lancs.ac.uk

Dawn Archer
Dept. of Linguistics and MEL
Lancaster University
d.archer@lancaster.ac.uk

Andrew Wilson
Dept. of Linguistics and MEL
Lancaster University
eiaaw@exchange.lancs.ac.uk

Tony McEnery
Dept. of Linguistics and MEL
Lancaster University
amcenery@lancaster.ac.uk

Abstract

Automatic extraction of multiword expressions (MWE) presents a tough challenge for the NLP community and corpus linguistics. Although various statistically driven or knowledge-based approaches have been proposed and tested, efficient MWE extraction still remains an unsolved issue. In this paper, we present our research work in which we tested approaching the MWE issue using a semantic field annotator. We use an English semantic tagger (USAS) developed at Lancaster University to identify multiword units which depict single semantic concepts. The Meter Corpus (Gaizauskas *et al*., 2001; Clough *et al*., 2002) built in Sheffield was used to evaluate our approach. In our evaluation, this approach extracted a total of 4,195 MWE candidates, of which, after manual checking, 3,792 were accepted as valid MWEs, producing a precision of 90.39% and an estimated recall of 39.38%. Of the accepted MWEs, 68.22% or 2,587 are low frequency terms, occurring only once or twice in the corpus. These results show that our approach provides a practical solution to MWE extraction.

1 Introduction

Automatic extraction of Multiword expressions (MWE) is an important issue in the NLP community and corpus linguistics. An efficient tool for MWE extraction can be useful to numerous areas, including terminology extraction, machine translation, bilingual/multilingual MWE alignment, automatic interpretation and generation of language. A number of approaches have been suggested and tested to address this problem. However, efficient extraction of MWEs still remains an unsolved issue, to the extent that Sag *et al*. (2001b) call it "a pain in the neck of NLP".

In this paper, we present our work in which we approach the issue of MWE extraction by using a semantic field annotator. Specifically, we use the UCREL Semantic Analysis System (henceforth USAS), developed at Lancaster University to identify multiword units that depict single semantic concepts, i.e. multiword expressions. We have drawn from the Meter Corpus (Gaizauskas *et al*., 2001; Clough *et al*., 2002) a collection of British newspaper reports on court stories to evaluate our approach. Our experiment shows that it is efficient in identifying MWEs, in particular MWEs of low frequencies. In the following sections, we describe this approach to MWE extraction and its evaluation.

2 Related Works

Generally speaking, approaches to MWE extraction proposed so far can be divided into three categories: a) statistical approaches based on frequency and co-occurrence affinity, b) knowledge–based or symbolic ap-

proaches using parsers, lexicons and language filters, and c) hybrid approaches combining different methods (Smadja 1993; Dagan and Church 1994; Daille 1995; McEnery *et al*. 1997; Wu 1997; Wermter *et al*. 1997; Michiels and Dufour 1998; Merkel and Andersson 2000; Piao and McEnery 2001; Sag *et al*. 2001a, 2001b; Biber *et al*. 2003).

In practice, most statistical approaches use linguistic filters to collect candidate MWEs. Such approaches include Dagan and Church's (1994) *Termight* Tool. In this tool, they first collect candidate nominal terms with a POS syntactic pattern filter, then use concordances to identify frequently co-occurring multiword units. In his Xtract system, Smadja (1993) first extracted significant pairs of words that consistently co-occur within a single syntactic structure using statistical scores called *distance*, *strength* and *spread*, and then examined concordances of the bi-grams to find longer frequent multiword units. Similarly, Merkel and Andersson (2000) compared frequency-based and entropy based algorithms, each of which was combined with a language filter. They reported that the entropy-based algorithm produced better results.

One of the main problems facing statistical approaches, however, is that they are unable to deal with low-frequency MWEs. In fact, the majority of the words in most corpora have low frequencies, occurring only once or twice. This means that a major part of true multiword expressions are left out by statistical approaches. Lexical resources and parsers are used to obtain better coverage of the lexicon in MWE extraction. For example, Wu (1997) used an English-Chinese bilingual parser based on stochastic transduction grammars to identify terms, including multiword expressions. In their DEFI Project, Michiels and Dufour (1998) used dictionaries to identify English and French multiword expressions and their translations in the other language. Wehrli (1998) employed a generative grammar framework to identify compounds and idioms in their ITS-2 MT English-French system. Sag *et al*. (2001b) introduced Head-driven Phrase Structure Grammar for analyzing MWEs. Like pure statistical approaches, purely knowledge-based symbolic approaches also face problems. They are language dependent and not flexible enough to cope with complex structures of MWEs. As Sag *et al*. (2001b) suggest, it is important to find the right balance between symbolic and statistical approaches.

In this paper, we propose a new approach to MWEs extraction using semantic field information. In this approach, multiword units depicting single semantic concepts are recognized using the Lancaster USAS semantic tagger. We describe that system and the algorithms used for identifying single and multiword units in the following section.

3 Lancaster Semantic tagger

The USAS system has been in development at Lancaster University since 1990[1]. Based on POS annotation provided by the CLAWS tagger (Garside and Smith, 1997), USAS assigns a set of semantic tags to each item in running text and then attempts to disambiguate the tags in order to choose the most likely candidate in each context. Items can be single words or multiword expressions. The semantic tags indicate semantic fields which group together word senses that are related by virtue of their being connected at some level of generality with the same mental concept. The groups include not only synonyms and antonyms but also hypernyms and hyponyms.

The initial tagset was loosely based on Tom McArthur's Longman Lexicon of Contemporary English (McArthur, 1981) as this appeared to offer the most appropriate thesaurus type classification of word senses for this kind of analysis. The tagset has since been considerably revised in the light of practical tagging problems met in the course of the research. The revised tagset is arranged in a hierarchy with 21 major discourse fields expanding into 232 category labels. The following list shows the 21 labels at the top level of the hierarchy (for the full tagset, see website: http://www.comp.lancs.ac.uk/ucrel/usas).

[1] This work is continuing to be supported by the Benedict project, EU project IST-2001-34237.

A general and abstract terms
B the body and the individual
C arts and crafts
E emotion
F food and farming
G government and the public domain
H architecture, buildings, houses and the home
I money and commerce in industry
K entertainment, sports and games
L life and living things
M movement, location, travel and transport
N numbers and measurement
O substances, materials, objects and equipment
P education
Q linguistic actions, states and processes
S social actions, states and processes
T time
W the world and our environment
X psychological actions, states and processes
Y science and technology
Z names and grammatical words

Currently, the lexicon contains just over 37,000 words and the template list contains over 16,000 multiword units. These resources were created manually by extending and expanding dictionaries from the CLAWS tagger with observations from large text corpora. Generally, only the base form of nouns and verbs are stored in the lexicon and a lemmatisation procedure is used for look-up. However, the base form is not sufficient in some cases. Stubbs (1996: 40) observes that "meaning is not constant across the inflected forms of a lemma", and Tognini-Bonelli (2001: 92) notes that lemma variants have different senses.

In the USAS lexicon, each entry consists of a word with one POS tag and one or more semantic tags assigned to it. At present, in cases where a word has more than one syntactic tag, it is duplicated (i.e. each syntactic tag is given a separate entry).

The semantic tags for each entry in the lexicon are arranged in approximate rank frequency order to assist in manual post editing, and to allow for gross automatic selection of the common tag, subject to weighting by domain of discourse.

In the multi-word-unit list, each template consists of a pattern of words and part-of-speech tags. The semantic tags for each template are arranged in rank frequency order in the same way as the lexicon. Various types of multiword expressions are included: phrasal verbs (e.g. *stubbed out*), noun phrases (e.g. *ski boots*), proper names (e.g. *United States*), true idioms (e.g. *life of Riley*).

Figure 1 below shows samples of the actual templates used to identify these MWUs. Each of these example templates has only one semantic tag associated with it, listed on the right-hand end of the template. However, the second example (ski boot) combines the clothing (B5) and sports (K5.1) fields into one tag. The pattern on the left of each template consists of a sequence of words joined to POS tags with the underscore character. The words and POS fields can include the asterisk wildcard character to allow for inflectional variants and to write more powerful templates with wider coverage. USAS templates can match discontinuous MWUs, and this is illustrated by the first example, which includes optional intervening POS items marked within curly brackets. Thus this template can match *stubbed out* and *stubbed the cigarette out*. 'Np' is used to match simple noun phrases identified with a noun-phrase chunker.

```
stub*_* {Np/P*/R*} out_RP     O4.6-
ski_NN1 boot*_NN*           B5/K5.1
United_* States_N*              Z2
life_NN1 of_IO Riley_NP1        K1
```

Figure 1 Sample of USAS multiword templates

As in the case of grammatical tagging, the task of semantic tagging subdivides broadly into two phases: Phase I (Tag assignment): attaching a set of potential semantic tags to each lexical unit and Phase II (Tag disambiguation): selecting the contextually appropriate semantic tag from the set provided by Phase I. USAS makes use of seven major techniques or sources of information in phase II. We will list these only briefly here, since

they are described in more detail elsewhere (Garside and Rayson, 1997).

1. *POS tag*. Some senses can be eliminated by prior POS tagging. The CLAWS part-of-speech tagger is run prior to semantic tagging.
2. *General likelihood ranking for single-word and MWU tags*. In the lexicon and MWU list senses are ranked in terms of frequency, even though at present such ranking is derived from limited or unverified sources such as frequency-based dictionaries, past tagging experience and intuition.
3. *Overlapping MWU resolution*. Normally, semantic multi-word units take priority over single word tagging, but in some cases a set of templates will produce overlapping candidate taggings for the same set of words. A set of heuristics is applied to enable the most likely template to be treated as the preferred one for tag assignment.
4. *Domain of discourse*. Knowledge of the current domain or topic of discourse is used to alter rank ordering of semantic tags in the lexicon and template list for a particular domain.
5. *Text-based disambiguation*. It has been claimed (by Gale *et al*, 1992) on the basis of corpus analysis that to a very large extent a word keeps the same meaning throughout a text.
6. *Contextual rules*. The template mechanism is also used in identifying regular contexts in which a word is constrained to occur in a particular sense.
7. *Local probabilistic disambiguation*. It is generally supposed that the correct semantic tag for a given word is substantially determined by the local surrounding context.

After automatic tag assignment has been carried out, manual post-editing can take place, if desired, to ensure that each word and idiom carries the correct semantic classification.

From these seven disambiguation methods, our main interest in this paper is the third technique of overlapping MWU resolution. When more than one template match overlaps in a sentence, the following heuristics are applied in sequence:

1. Prefer longer templates over shorter templates
2. For templates of the same length, prefer shorter span matches over longer span matches (a longer span indicates more intervening items for discontinuous templates)
3. If the templates do not apply to the same sequence of words, prefer the one that begins earlier in the sentence
4. For templates matching the same sequence of words, prefer the one which contains the more fully defined template pattern (with fewer wildcards in the word fields)
5. Prefer templates with a more fully defined first word in the template
6. Prefer templates with fewer wildcards in the POS tags

These six rules were found to differentiate in all cases of overlapping MWU templates. Cases which failed to be differentiated indicated that two (or more) templates in our MWU list were in fact identical, apart from the semantic tags and required merging together.

4 Experiment of MWE extraction

In order to test our approach of extracting MWEs using semantic information, we first tagged the newspaper part of the METER Corpus with the USAS tagger. We then collected the multiword units assigned as a single semantic unit. Finally, we manually checked the results.

The Meter Corpus chosen as the test data is a collection of court reports from the British Press Association (PA) and some leading British newspapers (Gaizauskas 2001; Clough *et al*., 2002). In our experiment, we used the newspaper part of the corpus containing 774 articles with more than 250,000 words. It provides a homogeneous corpus (in the sense that the reports come from a restricted domain of court events) and is thus a good source from which to extract domain-specific MWEs.

Another reason for choosing this corpus is that it has not been used in training the USAS system. As an open test, we assume the results of the experiment should reflect true capability of our approach for real-life applications.

The current USAS tagger may assign multiple possible semantic tags for a term when it fails to disambiguate between them. As mentioned previously, the first one denotes the most likely semantic field of the term. Therefore, in our experiment we chose the first tag when such situations arose.

A major problem we faced in our experiment is the definition of a MWE. Although it has been several years since people started to work on MWE extraction, we found that there is, as yet, no available "clear-cut" definition for MWEs. We noticed various possible definitions have been suggested for MWE/MWU. For example, Smadja (1993) suggests a basic characteristic of collocations and multiword units is recurrent, domain-dependent and cohesive lexical clusters. Sag *et el*. (2001b) suggest that MWEs can roughly be defined as "idiosyncratic interpretations that cross word boundaries (or spaces)". Biber *et al*. (2003) describe MWEs as lexical bundles, which they go on to define as combinations of words that can be repeated frequently and tend to be used frequently by many different speakers/writers within a register.

Although it is not difficult to interpret these deifications in theory, things became much more complicated when we undertook our practical checking of the MWE candidates. Quite often, we experienced disagreement between us about whether or not to accept a MWE candidate as a good one. In practice, we generally followed Biber *et al*.'s definition, i.e. accept a candidate MWE as a good one if it can repeatedly co-occur in the corpus.

Another difficulty we experienced relates to estimating recall. Because the MWEs in the METER Corpus are not marked-up, we could not automatically calculate the number of MWEs contained in the corpus. Consequently, we had to manually estimate this figure. Obviously it is not practical to manually check though the whole corpus within the limited time allowed. Therefore, we had to estimate the recall on a sample of the corpus, as will be described in the following section.

5 Evaluation

In this section, we analyze the results of the MWE extraction in detail for a full evaluation of our approach to MWE extraction.

Overall, after we processed the test corpus, the USAS tagger extracted 4,195 MWE candidates from the test corpus. After manually checking through the candidates, we selected 3,792 as good MWEs, resulting in overall precision of 90.39%.

As we explained earlier, due to the difficulty of obtaining the total number of true MWEs in the entire test corpus, we had to estimate recall of the MWE extraction on a sample corpus. In detail, we first randomly selected fifty texts containing 14,711 words from the test corpus, then manually marked-up good MWEs in the sample texts, finally counted the number of the marked-up MWUs. As a result, 1,511 good MWEs were found in the sample. Since the number of automatically extracted good MWEs in the sample is 595, the recall on the sample is calculated as follows:

$$\text{Recall}=(595\div1511)\times100\%=39.38\%.$$

Considering the homogenous feature of the test data, we assume this local recall is roughly approximate to the global recall of the test corpus.

To analyze the performance of USAS in respect to the different semantic field categories, we divided candidates according to the assigned semantic tag, and calculated the precision for each of them. Table 1 lists these precisions, sorting the semantic fields by the number of MWE candidates (refer to section 3 for definitions of the twenty-one main semantic field categories). As shown in this table, the USAS semantic tagger obtained precisions between 91.23% to 100.00% for each semantic field except for the field of "names and grammatical words" denoted by Z. As Z was the biggest field (containing 45.39% of the total MWEs and 43.12% of the accepted MWEs), we examined these MWEs

more closely. We discovered that numerous pairs of words are tagged as person names (Z1) and geographical names (Z2) by mistake, e.g. *Blackfriars crown* (tagged as Z1), *stabbed Constance* (tagged as Z2) etc.

Semantic field	Total MWEs	Accepted MWEs	Precision
Z	1,904	1,635	85.87%
T	497	459	92.35%
A	351	328	93.44%
M	254	241	94.88%
N	227	211	92.95%
S	180	177	98.33%
B	131	128	97.71%
G	118	110	93.22%
X	114	104	91.23%
I	74	72	97.30%
Q	67	63	94.03%
E	58	53	91.38%
H	53	52	98.11%
K	48	45	93.75%
P	39	37	94.87%
O	32	29	90.63%
F	24	24	100.00%
L	11	11	100.00%
Y	6	6	100.00%
C	5	5	100.00%
W	2	2	100.00%
Total	4,195	3,792	90.39%

Table 1: Precisions for different semantic categories

Another possible factor that affects the performance of the USAS tagger is the length of the MWEs. To observe the performance of our approach from this perspective, we grouped the MWEs by their lengths, and then checked precision for each of the categories. Table 2 shows the results (once again, they are sorted in descending order by MWE lengths). As we might expect, the number of MWEs decreases as the length increases. In fact, bi-grams alone constitute 80.52% and 81.88% of the candidate and accepted MWEs respectively. The precision also showed a generally increasing trend as the MWE length increases, but with a major divergence of tri-grams. One main type of error occurred on tri-grams is that those with the structure of *CIW*(capital-initial word) + *conjunction* + *CIW* tend to be tagged as Z2 (geographical name). The table shows relatively high precision for longer MWEs, reaching 100% for 6-grams. Because the longest MWEs extracted have six words, no longer MWEs could be examined.

MWE length	Total MWEs	Accepted MWEs	Precision
2	3,378	3,105	91.92%
3	700	575	82.14%
4	95	91	95.44%
5	18	17	94.44%
6	4	4	100.00%
Total	4,195	3,792	90.39%

Table 2: Precisions for MWEs of different lengths

As discussed earlier, purely statistical algorithms of MWE extraction generally filter out candidates of low frequencies. However, such low-frequency terms in fact form major part of MWEs in most corpora. In our study, we attempted to investigate the possibility of extracting low frequency MWEs by using semantic field annotation. We divided MWEs into different frequency groups, then checked precision for each of the categories. Table 3 shows the results, which are sorted by the candidate MWE frequencies. As we expected, 69.46% of the candidate MWEs and 68.22% of the accepted MWEs occur in the corpus only once or twice. This means that, with a frequency filter of *Min*(f)=3, a purely statistical algorithm would exclude more than half of the candidates from the process.

Freq. of MWE	Total number	Accepted MWEs	Precision
1	2,164	1,892	87.43%
2	750	695	92.67%
3 - 4	616	570	92.53%
5 - 7	357	345	96.64%
8 - 20	253	238	94.07%
21 - 117	55	52	94.55%
Total	4,195	3,792	90.39%

Table 3: Precisions for MWEs with different frequencies

Table 3 also displays an interesting relationship between the precisions and the frequencies. Generally, we would expect better precisions for MWEs of higher frequencies,

as higher co-occurrence frequencies are expected to reflect stronger affinity between the words within the MWEs. By and large, slightly higher precisions were obtained for the latter groups of higher frequencies (5–7, 8-20 and 21-117) than those for the preceding lower frequency groups, i.e. 94.07%-96.64% versus 87.43%-92.67%. Nevertheless, for the latter three groups of the higher frequencies (5-7, 8-20 and 21–117) the precision did not increase as the frequency increases, as we initially expected.

When we made a closer examination of the error MWEs in this frequency range, we found that some frequent domain-specific terms are misclassified by the USAS tagger. For example, since the texts in the test corpus are newspaper reports of court stories, many law courts (e.g. *Manchester crown court*, *Norwich crown court*) are frequently mentioned throughout the corpus, causing high frequencies of such terms (f=20 and f=31 respectively). Unfortunately, the templates used in the USAS tagger did not capture them as complete terms. Rather, fragments were assigned a Z1 person name tag (e.g. *Manchester crown*). A solution to this type of problem is to improve the multiword unit templates used in the USAS tagger. Other possible solutions may include incorporating a statistical algorithm to help detect boundaries of complete MWEs.

When we examined the error distribution within the semantic fields more closely, we found that most errors occurred within the Z and T categories (refer to Table 1). The errors occurring in these semantic field categories and their sub-divisions make up 76.18% of the total errors (403). Table 4 shows the error distribution across 14 sub-divisions (for definitions of these subdivisions, see: website: http://www.comp.lancs.ac.uk/ucrel/usas). Notice that the majority of errors are from four semantic sub-categories: Z1, Z2, Z3 and T1.3. Notice, also, that the first two of these account for 60.55% of the total errors. This shows that the main cause of the errors in the USAS tool is the algorithm and lexical entries used for identifying names - personal and geographical and, to a lesser extent, the algorithm and lexical entries for identifying periods of time. If these components of the USAS can be improved, a much higher precision can be expected.

In sum, our evaluation shows that our semantic approach to MWE extraction is efficient in identifying MWEs, in particular those of lower frequencies. In addition, a reasonably wide lexical coverage is obtained, as indicated by the recall of 39.38%, which is important for terminology building. Our approach provides a practical way for extracting MWEs on a large scale, which we envisage can be useful for both linguistic research and practical NLP applications.

Stag	Err.	Stag	Err.
Z1:person names	119	T1.1.1:time-past	1
Z2:geog. names	125	T1.1.2:time-present	1
Z3:other names	16	T1.2:time-momentary	8
Z4:discourse bin	3	T1.3:time-period	23
Z5:gram. bin	2	T2:time-begin/end	2
Z8:pronouns etc.	2	T3:time-age	1
Z99:unmatched	2	T4:time-early/late	2

Table 4: Errors for some semantic sub-divisions

6 Conclusion

In this paper, we have shown that it is a practical way to extract MWEs using semantic field information. Since MWEs are lexical units carrying single semantic concepts, it is reasonable to consider the issue of MWE extraction as an issue of identifying word bundles depicting single semantic units. The main difficulty facing such an approach is that very few reliable automatic tools available for identifying lexical semantic units. However, a semantic field annotator, USAS, has been built in Lancaster University. Although it was not built aiming to the MWE extraction, we thought it might be very well suited for this purpose. Our experiment shows that the USAS tagger is indeed an efficient tool for MWE extraction.

Nevertheless, the current semantic tagger does not provide a complete solution to the problem. During our experiment, we found that not all of the multiword units it collects are valid MWEs. An efficient algorithm is needed for distinguishing between free word

combinations and relatively fixed, closely affiliated word bundles.

References

Douglas Biber, Susan Conrad and Viviana Cortes. 2003. Lexical bundles in speech and writing: an initial taxonomy. In A. Wilson, P. Rayson and T. McEnery (eds.) *Corpus Linguistics by the Lune: a festschrift for Geoffrey Leech*, pp. 71-92. Peter Lang, Frankfurt.

Paul Clough, Robert Gaizauskas and S. L. Piao. 2002. Building and annotating a corpus for the study of journalistic text reuse. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*, pp. 1678-1691. Los Palmas de Gran Canaria, Spain.

Ido Dagan, and Ken Church. 1994. Termight: identifying and translating technical terminology. In *Proceedings of the 4th Conference on Applied Natural Language Processing*, pp. 34-40. Stuttgard, German.

Béatrice Daille. 1995. Combined approach for terminology extraction: lexical statistics and linguistic filtering. Technical paper. UCREL, Lancaster University.

Robert Gaizauskas, Jonathan Foster, Yorick Wilks, John Arundel, Paul Clough and Scott Piao. 2001. The METER corpus: a corpus for analysing journalistic text reuse. In *the Proceedings of the Corpus Linguistics 2001,* pp: 214-223. Lancaster, UK.

William Gale, Kenneth Church, and David Yarowsky. 1992. One sense per discourse. In *Proceedings of the 4th DARPA Speech and Natural Language Workshop*, pp 233-7.

Roger Garside and Nick Smith. 1997. A hybrid grammatical tagger: CLAWS4. In R. Garside, G. Leech and A. McEnery (eds.), *Corpus Annotation: Linguistic Information from Computer Text Corpora*, pp. 102-121. Longman, London.

Roger Garside and Paul Rayson. 1997. Higher-level annotation tools. In. Roger Garside, Geoffrey Leech, and Tony McEnery (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*, pp. 179 - 193. Longman, London.

Tom McArthur. 1981. *Longman Lexicon of Contemporary English*. Longman, London.

Tony McEnery, Langé Jean-Marc, Oakes Michael and Véronis Jean. 1997. The exploitation of multilingual annotated corpora for term extraction. In Garside Roger, Leech Geoffrey and McEnery Anthony (eds), *Corpus annotation --- linguistic information from computer text corpora*, pp 220-230. London & New York, Longman.

Magnus Merkel and Mikael Andersson. 2000. *Knowledge-lite extraction of multi-word units with language filters and entropy thresholds*. In *Proceedings of 2000 Conference User-Oriented Content-Based Text and Image Handling (RIAO'00)*, pages 737--746, Paris, France.

Archibald Michiels and Nicolas Dufour.1998. DEFI, a tool for automatic multi-word unit recognition, meaning assignment and translation selection. In *Proceedings of the First International Conference on Language Resources & Evaluation*, pp. 1179-1186. Granada, Spain.

Scott Songlin Piao and Tony McEnery. 2001. Multi-word unit alignment in English-Chinese parallel corpora. In *the Proceedings of the Corpus Linguistics 2001*, pp. 466-475. Lancaster, UK.

Ivan A. Sag, Francis Bond, Ann Copestake and Dan Flickinger. 2001a. Multiword Expressions. *LinGO Working Paper No. 2001-01*. Stanford University, CA.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake and Dan Flickinger. 2001b. Multiword Expressions: A Pain in the Neck for NLP. *LinGO Working Paper No. 2001-03*. Stanford University, CA.

Frank Smadja. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics* 19(1):143-177.

Michael Stubbs. 1996. *Text and corpus analysis: computer-assisted studies of language and culture*. Blackwell, Oxford.

Elena Tognini-Bonelli. 2001. *Corpus linguistics at work*. Benjamins, The Netherlands.

Eric Wehrli. 1998. Translating idioms. In *Proceedings of COLING-ACL '98*, Montreal, Canada, Vol. 2, pp. 1388-1392.

Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics* 23(3): 377-401.