Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computer Speech and Language 19 (2005) 378–397
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

Comparing and combining a semantic tagger and a statistical tool for MWE extraction

Scott Songlin Piao [a,*], Paul Rayson [b], Dawn Archer [a], Tony McEnery [a]

[a] *Department of Linguistics and Modern English Language, Lancaster University, Lancaster LA1 4YT, United Kingdom*
[b] *Computing Department, Lancaster University, Lancaster LA1 4YT, United Kingdom*

Received 4 June 2004; received in revised form 5 September 2004; accepted 11 November 2004
Available online 19 March 2005

Abstract

Automatic extraction of multiword expressions (MWEs) presents a tough challenge for the NLP community and corpus linguistics. Indeed, although numerous knowledge-based symbolic approaches and statistically driven algorithms have been proposed, efficient MWE extraction still remains an unsolved issue. In this paper, we evaluate the Lancaster UCREL Semantic Analysis System (henceforth USAS (Rayson, P., Archer, D., Piao, S., McEnery, T., 2004. The UCREL semantic analysis system. In: Proceedings of the LREC-04 Workshop, Beyond Named Entity Recognition Semantic labelling for NLP tasks, Lisbon, Portugal. pp. 7–12)) for MWE extraction, and explore the possibility of improving USAS by incorporating a statistical algorithm. Developed at Lancaster University, the USAS system automatically annotates English corpora with semantic category information. Employing a large-scale semantically classified multi-word expression template database, the system is also capable of detecting many multiword expressions, as well as assigning semantic field information to the MWEs extracted. Whilst USAS therefore offers a unique tool for MWE extraction, allowing us to both extract and semantically classify MWEs, it can sometimes suffer from low recall. Consequently, we have been comparing USAS, which employs a symbolic approach, to a statistical tool, which is based on collocational information, in order to determine the pros and cons of these different tools, and more importantly, to examine the possibility of improving MWE extraction by combining them. As we report in this paper, we have found a highly complementary relation between the different tools: USAS missed many domain-specific MWEs (law/court terms in this case), and the statistical tool missed many commonly used MWEs that occur in low frequencies (lower than three in this

* Corresponding author. Tel.: +44 1524 593025; fax: +44 1524 843085.
E-mail address: s.piao@lancaster.ac.uk (S.S. Piao).

0885-2308/$ - see front matter
doi:10.1016/j.csl.2004.11.002

case). Due to their complementary relation, we are proposing that MWE coverage can be significantly increased by combining a lexicon-based symbolic approach and a collocation-based statistical approach.

1. Introduction

Automatic extraction of multiword expressions (MWEs) is an important issue in the corpus research and NLP communities. An efficient tool for MWE extraction is needed in a number of areas, including terminology extraction, machine translation, bilingual/multilingual MWE alignment, information extraction, and text summarisation, etc. A variety of approaches have been suggested and tested to address this problem. So far, however, efficient extraction of MWEs still remains an unsolved issue, to the extent that Sag et al. (2001) call it the "pain in the neck of NLP".

Generally speaking, approaches to MWE extraction proposed so far can be divided into three paradigms: (a) knowledge-based or symbolic approaches using parsers, lexicons and language filters (Wu, 1997; Michiels and Dufour, 1998; Wermter and Chen, 1997; Tanaka and Baldwin, 2003; Nivre and Nilsson, 2004), (b) statistical approaches based on frequency and co-occurrence affinity (Smadja, 1993; Dagan and Church, 1994; Daille, 1995; McEnery et al., 1997; Merkel and Andersson, 2000; Piao and McEnery, 2001; Pereira et al., 2004), and (c) hybrid approaches (Maynard and Ananiadou, 2000; Sag et al., 2001; Dias, 2003).

Symbolic systems can be efficient in identifying core MWEs, which often occur in general domains and genres, as they depend on a knowledge base drawn upon humans' perceptions of lexical usage (in our case, of expert linguists). As such systems are largely immune from frequency, they can be relatively more efficient in identifying low-frequency MWEs than statistical approaches (which tend to be inefficient in this regard). Even so, symbolic approaches have their limitations. For example, they are usually built upon large lexicons or rule bases, which tend to be expensive and difficult to produce (Dias, 2003). Moreover, their coverage is often limited to the training data. Consequently, they tend to suffer from low recall when dealing with domains/genres beyond the scope of the training data (see Section 2.3).

In contrast to symbolic approaches, statistical approaches are generally more flexible and efficient in dealing with MWEs of higher frequencies. Although they occasionally make use of small linguistic filters, they tend not to need large lexicons or rule bases. However, their statistical nature means that they are not accurate for dealing with MWEs of very low frequencies, particularly those occurring only once or twice. Not surprisingly, research on statistical approaches tends to ignore such low-frequency MWEs. Unfortunately, however, such low-frequency MWEs tend to form a significant proportion of the total MWEs in most texts (Lapata and Lascarides, 2003; Tanaka and Baldwin, 2003). Consequently, the usefulness of pure statistical approaches in practical NLP applications is limited.

In fact, many MWE extraction tools employ hybrid algorithms to various extents to obtain an optimal result. For example, Maynard and Ananiadou (2000) combined both linguistics and statistical information in their system, TRUCK, for extracting multi-word terms, reporting a precision of 64% on 250 top-ranked candidates. Dias (2003) also combined word statistics with endogenously acquired linguistic information into a hybrid system, obtaining 62% precision on

Brown Corpus. Such work suggests that it is important to find the right balance between symbolic and statistical approaches, as suggested by Sag et al. (2001).

Over the past decade, we have been developing a system for semantically tagging English corpora. Employing a large semantically classified MWE template database, USAS (as the system is known) provides a unique means of detecting and then semantically classifying *known* MWEs according to one or more of 232 pre-defined semantic field categories (see Section 2.1 for details). Potentially, USAS can also provide a means of extracting *unknown* MWEs, when combined with statistical tools. The latter is important, as pure symbolic approaches tend to be affected by the limitation of their lexicon knowledge base. In this paper, we report on an experiment in which we tested our hypothesis. Below, we begin by testing the individual efficiency of USAS and the statistical tool for MWE extraction (Sections 2 and 3).

2. Extracting MWEs using a semantic tagger

In this section, we explore the possibility of extracting MWEs using USAS. We briefly describe the semantic tagger, and evaluate the performance of this tool for MWE extraction.

2.1. The USAS semantic tagger

USAS has been in development at Lancaster University since 1990. Based on POS annotation provided by the CLAWS tagger (Garside and Smith, 1997), USAS assigns a semantic tag to each word or MWE in a running text (Rayson et al., 2004). It employs a tagset that is arranged in a hierarchical structure with 21 major semantic fields expanding into 232 categories. Table 1 lists the 21 labels at the top level of the hierarchy (for the full tagset, see website: http://www.comp.lancs.ac.uk/ucrel/usas/).

USAS is one of the few tools to offer a large machine-useable semantic lexical database (see Lewandowska-Tomaszczyk, 2003, p. 22). The lexicon, which is generally extended and expanded with observations from large corpora, contains two main parts: a single word sub-lexicon and a MWE sub-lexicon. Currently, the former contains approximately 45,000 words while the latter contains over 18,600 MWE templates. As the MWE sub-lexicon is important to our MWE extraction work, we will describe it briefly here (for a more general account of the USAS semantic tagger, see Rayson et al., 2004).

In the MWE sub-lexicon, each MWE template consists of a pattern of words and part-of-speech tags. Some template entries contain more than one candidate semantic tag. In such cases the tags are arranged in approximate frequency order. MWEs can be classified in numerous ways, of course, not least their morpho-syntactic patterns (e.g., noun phrase, phrasal verb), their orthography (e.g., via hyphenation), the fixedness of their collocational link, or their relation to distinguishable concepts in the 'real world', etc. Thus, for example, a 'yellow jersey' may operate as a MWE in some (domain-specific) contexts but not in others (cf. the prize awarded to stage leaders in the Tour de France and a jersey in another context that happens to be yellow) In the USAS database, a lexical unit is admitted as a MWE if it expresses a meaning which is distinct from the sum of its parts or has a meaning which is difficult to recover from the sum of its parts. These MWEs are then grouped into 232 semantic field categories regardless of their morpho-syntactic

Table 1
USAS semantic tagset

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
K	Entertainment, sports and games
L	Life and living things
M	Movement, location, travel and transport
N	Numbers and measurement
P	Education
O	Substances, materials, objects and equipment
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

patterns. This means that the MWEs under each semantic field category tend to represent a variety of MWE *types*, including phrasal verbs, noun phrases, proper names, and true idioms, etc. For example, the category of 'Crime, Law and Order' (denoted by code G2.1) contains the following: "*act*_NN* of_IO parliament_NN1*", "*appeal*_VV* against_II*", "*forced_JJ entry_NN1*", etc. We are not alone in categorising MWEs in terms of the semantic field(s) they represent. Maynard and Ananiadou (2000, p. 3), for example, distinguish a 'term' from a 'word' on the basis of its semantic relation to a concept or concepts and a concept's relationship, in turn, with other concepts within a particular field or fields.

Fig. 1 shows some sample template entries. Each of the example templates has only one semantic tag associated with it, listed on the right. 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 wider coverage. Many MWE templates can match discontinuous MWEs, as illustrated by the first sample entry, which includes optional intervening POS items marked within curly brackets. For example, this sample entry can match *stubbed out* and *stubbed the cigarette out*, where 'Np' is used to match simple noun phrases identified with a noun-phrase chunker.

The process 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 (for a detailed description, see Garside and Rayson, 1997; Rayson et al., 2004):

```
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
*_MC a_AT1 side_NN1                K5.1
```

Fig. 1. Sample of USAS MWE entries.

(1) POS tag.
(2) General likelihood ranking for single-word and MWE tags.
(3) Overlapping MWE resolution.
(4) Domain of discourse.
(5) Text-based disambiguation.
(6) Contextual rules.
(7) Local probabilistic disambiguation.

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.

So far, these six rules have been able to differentiate the majority of cases where overlapping MWE templates occur.

Using the MWE template sub-lexicon, the USAS tagger is capable of identifying MWEs by matching word groups, continuous or discontinuous, against the templates. Because the templates allow fuzzy matching via quasi-regular expressions such as wildcards and curly brackets, it provides a wider coverage than the number of individual entries, and thus provides a powerful tool for MWE extraction (see below).

2.2. Experiment

As explained earlier, USAS is capable of identifying MWEs. In order to evaluate this function, we (i) tagged a test corpus with the semantic tagger, (ii) collected the word groups assigned as a single semantic unit, and (iii) checked the results manually.

We chose the Meter Corpus as our test corpus, that is, a collection of court reports from the British Press Association (PA) and some leading British newspapers (Gaizauskas et al., 2001). In our experiment, we used the newspaper part of the corpus containing 774 articles with more

than 250,000 words. It provides a homogeneous data (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 therefore reflect the true capability of this tool for practical applications, in which we often expect to encounter texts from new domains/genres.

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. Consequently, we chose the first tag when such situations arose in our experiment.

As we have intimated (see Section 2.1), there are many ways of defining/categorising MWEs. Not surprisingly, there is often disagreement between the experts as to what constitutes the latter. Smadja (1993), for example, suggests a basic characteristic of collocations and multiword units is recurrent, domain-dependent and cohesive lexical clusters. In contrast, Sag et al. (2001) suggest that MWEs can roughly be defined as ''idiosyncratic interpretations that cross word boundaries (or spaces)'', and Biber et al. (2003) describe MWEs as 'lexical bundles', i.e., combinations of words that can be used and repeated frequently by many different speakers/writers within a register.[1] Although it is not difficult to interpret these definitions in theory, things became much more complicated when we undertook our practical checking of the potential MWE candidates.

As previously explained, we opted to identify as MWEs those lexical units whose meaning is distinct/not recoverable from the sum of their parts, regardless of their morpho-syntactic patterning. We also followed Biber et al.'s definition, in that we accepted a candidate MWE as a ''good'' one if it repeatedly co-occurred in the corpus, and was likely to be used by different speakers/writers. Although recurrency is often linked with domain-dependency, it is worth noting that a substantial number of MWEs tend to be general in nature. Consequently, we were able to utilize general knowledge dictionaries, thesauri and internet search engines as a means of checking whether candidate MWEs repeatedly co-occur and are used by different speakers/writers. More domain-specific terms are a little more difficult to check using the above methods, however. Consequently, we utilised the expertise of a linguist who is familiar with the law court domain to check MWEs found in the Meter Corpus (e.g., *assault and battery*, *artist's impression*, *came at*).

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 total 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 (see the following section).

2.3. Evaluation

In order to evaluate the USAS tagger for MWE extraction, we processed the test data of METER, and collected the word groups tagged as MWEs. We found that the USAS tagger extracted a total of 4195 MWE candidates. After manually checking through the candidates, we selected

[1] For a variety of linguistic and lexicographic definitions of MWE, see Tschichold (2000) and Moon (1998).

3792 as good MWEs. This gave us an overall precision of 90.39%. 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. We did so by randomly selecting 50 texts containing 14,711 words from the test corpus, then manually marking-up the MWEs in these sample texts, and finally counting the number of marked-up MWEs. As a result, 1511 MWEs were found in the sample. Given that 595 MWEs were recognised by the semantic tagger, we have calculated the recall on the sample as follows:

$$\text{Recall} = \frac{595}{1511} \times 100\% = 39.38\%.$$

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

To analyse the performance of the semantic tagger 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 2 lists these precisions, sorting the semantic fields by the number of MWE candidates in descending order (for definitions of the 21 main semantic field categories, see Section 2.1). As shown in this table, the USAS semantic tagger obtained precisions of 91.23–100.00% 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

Table 2
Precisions for different semantic categories

Semantic field	Total MWEs	Accepted MWEs	Precision (%)
Z	1904	1635	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	4195	3792	90.39

Table 3
Precisions for MWEs of different lengths

MWE length	Candidate MWEs	Accepted MWEs	Precision (%)
2	3378	3105	91.92
3	700	575	82.14
4	95	91	95.44
5	18	17	94.44
6	4	4	100.00
Total	4195	3792	90.39

as person names (Z1) and geographical names (Z2) by mistake, e.g., *Blackfriars crown* (tagged as Z1), *stabbed Constance* (tagged as Z2), etc.

Another possible factor that affects the performance of the USAS tagger is the length of the MWEs. To observe the performance of the semantic tagger from this perspective, we grouped the MWEs by their lengths, and then checked precision for each of the categories. Table 3 shows the results, this time sorting the results by MWE lengths in ascending order. As one 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 shows a generally increasing trend as the MWE length increases, but with a major divergence of tri-grams. One main type of error occurring 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, longer MWEs could not be examined.

As discussed earlier, purely statistical algorithms of MWE extraction generally filter out candidates of low frequencies. However, such low-frequency terms in fact form a major part of MWEs in most corpora. In order to investigate the tagger's capability of extracting low frequency MWEs, we divided MWEs into different frequency groups, and then checked precision for each of the categories. Table 4 shows the results, which are sorted by the candidate MWE frequencies in ascending order. As a result, 69.46% of the candidate MWEs and 68.22% of the accepted MWEs occurred 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.

Table 4 also reveals an interesting relationship between the precisions and the frequencies. Generally, one would expect better precisions for MWEs of higher frequencies, as higher

Table 4
Precisions for MWEs with different frequencies

Frequency of MWE	Candidate MWEs	Accepted MWEs	Precision (%)
1	2164	1892	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	4195	3792	90.39

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 groups of higher frequencies (5–7, 8–20 and 21–117) than for those of lower frequencies, 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 (e.g., *Manchester crown*) were assigned with Z1 (person name). A solution to this type of problem is to improve the MWE templates of the semantic tagger. Other possible solutions may include incorporating an 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 2). The errors occurring in these semantic field categories and their sub-divisions make up 76.18% of the total errors (403). Table 5 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 the errors are from four semantic sub-categories: Z1, Z2, Z3 and T1.3. Notice, also, that the first two of them 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 the semantic tagger to be an efficient tool for 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. Such a tool can provide a practical way for extracting MWEs on large scales. Nevertheless, the current

Table 5
Errors for some semantic sub-divisions

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

semantic tagger does not provide a complete solution to the problem. As shown in our experiment, not all of the candidate MWEs it collects are valid MWEs. Consequently, an efficient algorithm is needed for distinguishing between free word combinations and relatively fixed, closely affiliated word bundles. Finally, the recall may not be high enough for some practical NLP tasks.

3. Extracting MWEs using statistical algorithm

As the above experiment reveals, the Lancaster semantic tagger can accurately extract MWEs – when those MWEs are covered by the MWE template lexicon. However, as the MWE template database was basically derived from the BNC (i.e., a generic English corpora), it tends to suffer from low recall when processing texts from more specific domains like those represented by the Meter Corpus. Given MWEs such as *assault and battery*, *sent down*, etc., are likely to occur much more frequently in legal texts (i.e., particular domain-specific datasets) than in generic datasets like the BNC, statistical algorithms may help USAS improve the coverage of its MWE extraction. To test this assumption, we have evaluated a collocation-based statistical tool for MWE extraction (see the following section), and then compared it with USAS (see Section 4).

3.1. A statistical algorithm based on collocational co-occurrence association

The statistical algorithm we test in our experiment is based on collocational co-occurrence association. Given a text, it collects global co-occurrence information for all candidate word-pairs, which, in turn, is used to identify MWEs in local contexts. The following is a brief description of its algorithm:

(1) Pos-tag the input text using CLAWS POS tagger.
(2) Collect collocates using the co-occurrence association score.
(3) Using the collection of collocates as a statistical dictionary, check the affinity between closely adjacent words to create affinity distribution map.
(4) Based on the affinity distribution, collect the word clusters (not just word pairs) that are subject to relatively stronger affinity.
(5) Optionally, apply simple filters to clear highly frequent errors.

Although this algorithm can be applied to raw texts, we pre-processed the input text using the CLAWS tagger. Our reason for doing so is that word class information can influence co-occurrence association between words. For example, in our test corpus, the word ''convicted'' as adjective closely co-occurred with the word ''paedophile'' five times while ''convicted'' as past participle verb co-occurred with the word ''causing'' four times. This example shows that different POS categories of a given word can influence the selection of the word/word group with which it can associate.

Regarding the statistical metric for measuring co-occurrence affinity, quite a few statistical measures have been suggested and tested during the past decades (Manning and Schutze, 2000, pp. 151–190), including mutual information, log-likelihood, χ^2 test, etc. Among them, log-likelihood has been a popular metric for identifying collocations from sparse data (Dunning, 1993; Daille,

1995; Manning and Schutze, 2000, pp. 172–175). For example, having compared several statistical measures for text analysis, Dunning (1993) observed that log-likelihood ratio produces a better result when based on relatively smaller volumes of text. By comparing a set of statistical measures for French terminology extraction, Daille (1995) also found that the log-likelihood ratio provides the best statistical model. Because we hoped to collect collocate pairs of frequencies as low as three, and our test data had a moderate size of about 250,000 words, we adopted this metric for our algorithm. The log-likelihood is calculated with a formula adjusted for co-occurrence contingency table (Scott, 2001) as follows.

For a given pair of words X and Y and a search window W, let a be the number of windows in which X and Y co-occur, let b be the number of windows in which only X occurs, let c be the number of windows in which only Y occurs, and let d be the number of windows in which none of them occurs, then

$$G_2 = 2(a\ln a + b\ln b + c\ln c + d\ln d - (a+b)\ln(a+b) - (a+c)\ln(a+c) - (b+d)\ln(b+d) - (c+d)\ln(c+d) + (a+b+c+d)\ln(a+b+c+d)). \quad (1)$$

In order to allow for a wider range of candidate MWEs, a very low threshold of 1.00 was used for this score, i.e., collocate pairs yielding G_2-score less than 1.00 were filtered out. For practical tasks requiring high precisions, higher thresholds such as 3.84 (95% significance for degrees of freedom 1) can be used.

Besides the G_2 score described above, another t-score has also been used for filtering out insignificant co-concurrent pairs (Gale and Church, 1991; Fung and Church, 1994; Haruno and Yamazaki, 1996). Although the above two scores have similar functions, we found that the t-score can catch some false collocate pairs which are missed by the log-likelihood score. By way of illustration, because the METER Corpus contains texts that describe court cases that occurred the day prior to their respective publishing dates, the words "yesterday" and "said" occur very often in our test data. Although these two words co-occurred five times within the search window, they also occurred 809 and 2137 times, respectively, by themselves elsewhere. As a result, they produce a log-likelihood score of 6.532 but a very low t score of −2.823, which is less than the threshold. Consequently, "yesterday" and "said" were rejected from the collocate collection. In order to enhance the performance of the algorithm, the t-score is use as an additional filter, which is calculated as follows:

$$t = \frac{\text{prob}(W_a, W_b) - \text{prob}(W_a)\text{prob}(W_b)}{\sqrt{\frac{1}{M}\text{prob}(W_a, W_b)}}. \quad (2)$$

In our experiment, a t-score threshold of 1.65[2] is used; the collocation pairs producing t-score lower than this threshold were filtered out.

In order to reduce the problem caused by function words, which are omni-present in the corpus data and cause "noise" to the statistical algorithm of collocation detection, we used a stop word list of function words to exclude them from the collocation extraction process.

[2] The t-score of 1.65 indicates a confidence level greater than 95% (Fung and Church, 1994).

With regard to the searching window, based on preliminary experiments on small sample text, we set the length of the window to two words. In addition, the search proceeds in a fixed direction, i.e., from left to right. Obviously, there can be different search strategies that may produce different results. However, as our focus in this experiment was to compare a typical statistical algorithm and the semantic tagger for MWE extraction rather than developing sophisticated statistical tools, we did not explore all possible searching strategies.

Regarding the creation of the collocation dictionary, if the input text is large enough, the collocation dictionary can be directly extracted from it. Alternatively, if the input text is small, the collocation dictionary can be obtained from a relevant corpus in advance. In our case, the collocation dictionary was extracted directly from the METER Corpus test data (over 250,000 words), as it was considered to be large enough for this purpose. That is, the whole MWE extraction process was automated by (a) extracting collocates from the input text on the fly and (b) applying it as a statistical dictionary for searching for MWEs.

The next step involves testing the affinity between words. To elaborate, for a given input text – POS tagged in this case – the algorithm first examines the affinity strength between adjacent words by looking them up in the statistical dictionary. If the pair of words is not found in the dictionary, the affinity between them is set to zero. To simplify the experiment, we only considered neighboring words pairs. In this way, we obtain an affinity distribution across the given text. Fig. 2. illustrates the affinity distribution of a sentence from the *Daily Express* newspaper:

> *Deputy_NN1 principal_NN1 Alden_NN1 was_VBDZ jailed_VVN for_IF 15_MC years_NNT2 after_II being_VBG found_VVN guilty_JJ of _IO five_MC ind e cent_JJ assaults_NN2,_, one_MC1 gross_NNO indecency_NN1 and_CC four_MC serious_JJ sexual_JJ assaults_NN2._.*

As shown in Fig. 2, the curve bulges upward over the word groups "Deputy_NN1 principle_NN1", "found_VVNguilty_JJ", "indecent_JJassaults_NN2" and "serious_JJ sexual_JJ assaults_NN2". The heights of the bulges indicate the affinity strength between the words concerned. Our algorithm searches for the curve bulges and marks the word groups covered by

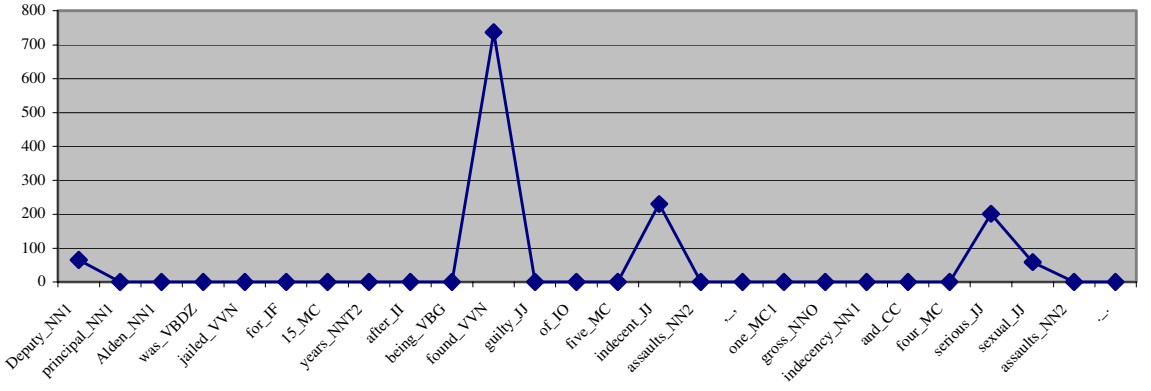

Fig. 2. Affinity distribution of a sample sentence.

them as candidate MWEs. Finally, we obtain an output text in which MWEs are marked-up using tag pairs <mwe> </mwe>, as shown below:

< s > < mwe > Deputy_NN1 principal_NN1 < /mwe > Alden_NN1 was_VBDZ jailed_VVN for_IF15_MC years_NNT2 after_II being_VBG < mwe > found_VVN guilty_JJ < /mwe > of_IO five_MC < mwe > indecent_JJ assaults_NN2 < /mwe > ,_,one_MC1 gross_NNO indecency_NN1 and_CC four_MC < mwe > serious_JJ sexual_JJ assaults_NN2 < /mwe > ._. < /s >

3.2. Evaluation of statistical tool

In order to evaluate this statistical tool, we applied it to the same test corpus used to evaluate the semantic tagger. We found that the statistical tool extracted 3306 candidate MWEs in total. After a manual check, 2705 of them were accepted as "good" MWEs, producing a precision of 81.85%.

We considered introducing lemmatisation for the MWE extraction. However, due to the POS information involved in our algorithm, the lemmatisation was found to have little impact on the result. For example, both "friend" and "friends" occur in our test corpus. When lemmatised, they were converted into a common basic form "friend", but as they carry different POS tags NN1 and NN2, respectively, they collocated with "close_JJ" separately: *close_jj friend_nn1* and *close_jj friend_nn2*. When we applied the lemmatiser to the whole test corpus, our tool produced 3159 candidates, of which 3116, or 98.64%, were duplicates of those obtained without using the lemmatiser. Although about 15 valid MWE were found in the 43 unduplicated items, we might lose many more valid MWEs by using the lemmatiser. While we acknowledge that a lemmatiser can boost MWE extraction in raw texts, we opted to exclude the lemmatiser from our experiment. Therefore, the statistical figures quoted hereafter in our evaluation are those obtained without using a lemmatiser.

For a further evaluation, we examined the above results from two aspects: frequency and MWE length. Firstly, in order to examine how MWE frequency affects the performance of the statistical tool, we divided the MWEs into six frequency groups, and then checked the precision for each of them. Table 6 shows the result. As shown in this table, by and large the precision improves from 66.34% to 91.30% as the frequency increases, with the highest precision obtained for frequency band of 5 and the lowest precision obtained for the MWEs occurring only once. The exceptional frequency band of "$\geqslant 6$" was the result of some highly frequent domain specific errors like "years_yesterday" ($f = 35$), producing a lower precision than the preceding frequency band of $f = 5$. Such a precision distribution over frequencies can be explained by the statistical nature of this algorithm, in which the co-occurrence frequency underpins the collocation extraction.

As shown in Table 6, this statistical MWE searching algorithm picked up quite a few low frequency MWEs. To be precise, 676 of the extracted MWEs, or 24.99% of them, have frequencies of one or two. In these MWES, the neighboring constituent words have strong collocational affinity in the test corpus as a whole. In other words, although the whole MWE word sequence occurs only once or twice, each pair of the neighboring words within the MWE share a strong association over the entire corpus. This result shows that a statistical algorithm, to some extent, can also deal with MWEs of low frequencies.

Table 6
Precision for MWEs of different frequencies

MWE frequency	Candidate MWEs	Accepted MWEs	Precision (%)
1	606	402	66.34
2	372	274	73.66
3	966	819	84.78
4	453	397	87.64
5	276	252	91.30
$\geqslant 6$	633	561	88.63
Total	3306	2705	81.85

Next, we further examined the precision for MWEs of different lengths. Table 7 lists the results. As shown in this table, our statistical tool seems to perform more accurately on shorter MWEs than on longer ones. Indeed, it obtained a precision of 84.79% for two-word MWEs, but only half of the candidate MWEs longer than five words were accepted. Such a result shows that our statistical tool works quite efficiently on shorter MWEs but becomes less reliable when dealing with longer MWEs.

Outputs of such a purely statistical algorithm can contain numerous errors, as indicated by the moderate precision of 81.85%. One reason for the numerous errors relates to the presence (in the test corpus) of some highly frequent domain-specific non-MWE word sequences. They formed tight collocate pairs which were collected and included into the collocation dictionary. For example, our test corpus contains some word sequences reflecting the feature of law/court news reports, such as {Name/Noun + "said"},{Name/Noun + "told"} {"told" + Name/Noun}, etc. Also, because many reports talk about events that occurred the previous day, the word "yesterday" occurs with abnormally high frequency. As a result, in the collocation list, "said" was collocated with 104 nouns/names, "told" with 67 nouns/names, and "yesterday" with 59 various words. When applied to the MWE searching algorithm, "said" created 139 errors, "told" 41 errors, and "yesterday" 89 errors. These three types of errors alone accounted for 44.76% of the total 601 errors. Linguistic filters can be used to reduce such errors. Indeed, with a simple filter blocking these three types of errors, the precision can be improved from 81.85% to 89.96%.

Due to the difficulty of exhaustively examining recall over the whole test corpus, we estimated it based on the same 50 randomly chosen sample texts used for estimating the recall of the semantic tagger (refer to Section 2.3). The MWEs in those texts were manually marked, so we could com-

Table 7
Precision for MWEs of different lengths

MWE length	Candidate MWEs	Accepted MWEs	Precision (%)
2	2413	2046	84.79
3	647	494	76.35
4	180	121	67.22
5	56	39	69.64
$\geqslant 6$	10	5	50
Total	3306	2705	81.85

pare the result of the automatic tools against human expert judgment. Of the 1511 MWEs manually marked-up, 343 were identified by the statistical tool, producing a recall of 22.70%. One of the reasons for this low recall is that our statistical tool excludes function words from collocation, and therefore it is unable to recognize MWEs containing function words. The searching algorithm needs to be improved to alleviate this problem.

In sum, our investigations suggest that the experimental statistical tool can provide an effective means of extracting recurring MWEs. To some extent, it is also capable of identifying low-frequency MWEs. Finally, although this aspect of our investigations has not been fully explored as yet, there is strong evidence to suggest that the tool's precision can be improved using simple linguistic filters.

4. Towards a hybrid MWE extraction system: comparing the semantic tagger and the statistical tool for MWE extraction

The main purpose of this experiment was to investigate the possibility of improving the USAS system for MWE extraction by incorporating a statistical algorithm into it. In particular, we hoped to estimate the extent to which the statistical tool is complementary to USAS. Because the statistical tool cannot assign semantic field information to the MWEs, we could not compare the results for individual semantic categories. Instead, we compared the outputs of these two tools under a number of other criteria.

First of all, we compared them generally in terms of the number of extracted MWEs, precision and recall. As Table 8 highlights, the semantic tagger generally performed better than the statistical algorithm in all these aspects.

The next stage involved determining if the tools were complementary to each other. An effective metric for measuring the complementary relationship between two systems is the overlap between their results. For a given pair of systems A and B, if they produce largely overlapping results, then they are not complementary; as any one of them can basically be substituted by the other. On the other hand, if they produce very different results, i.e., there is no or only a small overlap between their results, then they are highly complementary to each other, and, as such, we can expect to achieve a better result by combing them together than we would by using either of them alone.

Fig. 3 illustrates the overlap between the outputs of the statistical tool (a) and the semantic tagger (b). Notice that 655 MWEs were recognised by both tools. This means that 75.79% and 82.73% of the MWEs extracted by them, respectively, were complementary results.

In order to test the influence of MWE frequencies to the performance of these tools, we compared the MWE distributions for six different frequency bands. As Table 9 reveals, nearly half of the MWEs extracted by the semantic tagger occurred only once in the test corpus, whereas the

Table 8
General comparison between semantic tagger and statistical tool

Tools	Candidate MWEs	MWEs	Precision (%)	Recall (%)
Semantic tagger	4195	3792	90.39	39.38
Statistical tool	3306	2705	81.85	22.70

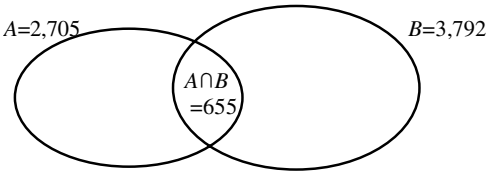

Fig. 3. Comparison of outputs of statistical and semantic tools.

Table 9
Comparison of MWE frequency distributions

MWE frequency	Semantic tagger	Percentage (%)	Statistical tool	Percentage (%)
1	1892	49.89	402	14.86
2	695	18.33	274	10.13
3–4	570	15.03	1216	44.95
5–7	345	9.10	504	18.63
8–20	238	6.28	261	9.65
$\geqslant 21$	52	1.37	48	1.77
Total	3792	100.00	2705	100.00

largest group of MWEs extracted by the statistical tool came from the frequency range of 3–4. On the whole, the distribution of the semantic tagger's output seems to follow the natural lexical distribution for frequencies, i.e., the number of MWEs decreases as the frequency increases. This shows that the performance of the semantic tagger is less affected by frequency. In contrast, the main part of the MWEs collected by the statistical tool cluster around the frequency range of 3–7, revealing its sensitiveness to frequency.

We also compared the semantic tagger and statistical tool in terms of MWE length distribution. Table 10 lists the number of MWEs and corresponding percentages for each length. Notice that the distributions for these tools have similar structures, that is, for both of the tools the majority of the extracted MWEs, 75.64% for the statistical tool and 81.88% for the semantic tagger, contain two words. Notice, also, that as the MWEs grow longer, the number of them gradually drops. Such nearly parallel distributions show that the MWE length does not affect their performance significantly.

In sum, the statistical tool and semantic tagger demonstrated contrasting but largely complementary features, e.g., a small overlap between their outputs and distinct performances on different frequency bands.

The main shortcoming identified of the semantic tagger is its inability to capture many frequently occurring domain specific MWEs unless they were already included in its MWE template base. Unidentified examples from our test corpus include core legal terms such as ''forensic scientists'', ''forensic tests'', etc. It must be noted that, very often, such domain-specific MWEs are the most significant terms needed for NLP applications, making this a serious limitation. Due to their statistical nature, statistical tools rely on a certain level of frequency to discover and predict co-occurrence patterns. Consequently, statistical tools are generally good at identifying frequently occurring MWEs, but less efficient when dealing with very low-frequent MWEs, in particular those occurring only once or twice. Although, in theory, sufficiently large corpora may provide

Table 10
Comparison of MWE length distributions

MWE length	Semantic tagger	Percentage (%)	Statistical tool	Percentage (%)
2	3105	81.88	2046	75.64
3	575	15.16	494	18.26
4	91	2.40	121	4.47
5	17	0.45	39	1.44
$\geqslant 6$	4	0.11	5	0.18
Total	3792	100.00	2705	100.00

high frequencies for every possible MWE, it is impractical, if not impossible, to obtain such an ideal corpus. This is where MWE lexicon resources like the Lancaster MWE template data can have a critical role. Indeed, out of the 3137 MWEs identified by the semantic tagger but missed by the statistical tool, 2785 or 88.78% of them occurred only once or twice in the test corpus.

The complementary features of the statistical tool in comparison with USAS make it possible to improve MWE extraction by combing them. As our comparative study shows, working together, they could extract a total of $3792 + 2705 - 655 = 5842$ MWEs. Compared to the individual statistical tool and semantic tagger, this means improvements of $((5842 - 2705)/2705) \times 100\% = 115.97\%$ and $((5842 - 3792)/3792) \times 100\% = 54.06\%$, respectively, in terms of the number of extracted MWEs. As shown in Table 8, the semantic tagger and the statistical tool extracted 4195 and 3306 candidate MWEs, respectively, of which 873 were overlapping items. This results in a total precision of $(5842/(4195 + 3306 - 873)) \times 100\% = 88.14\%$. Comparing to reports from Maynard and Ananiadou (2000) and Dias (2003) on similar systems (see Section 1), this result is rather encouraging.

In respect of recall, out of the manually identified 1511 MWEs in the 50 sample texts, the semantic tagger and the statistical tool identified 595 and 343 MWEs, respectively, with an overlap of 175. Put together, they identified $595 + 343 - 175 = 763$ MWEs, increasing the recall to 50.5% (refer to Table 8).

Due to the limited domain of our test corpus, as well as the experimental nature of the statistical tool, these results may not be conclusive. Even so, our experiment clearly demonstrates the complementary relationship between the statistical tool and the semantic tagger, and clearly points to the advantage of hybrid approaches for MWE extraction. Indeed, although there are some issues remain unsolved, such as automatic semantic classification of the MWEs extracted statistically, the incorporation of a statistical algorithm means that the USAS system markedly improves its MWE coverage and recall when dealing with texts from new domains. Such an ability to detect and extract MWEs from new domains is important for practical applications such as automatic content analysis (Sawyer et al., 2002; Onditi et al., 2004). It may also prove to be a useful means of improving machine dictionaries (Löfberg et al., 2004).

5. Conclusion

In this paper, we evaluated the Lancaster USAS semantic tagger for MWE extraction and investigated the feasibility of improving it by incorporating a statistical algorithm. We did so

by initially testing and then comparing the USAS tagger and a collocation-based statistical tool. Although our main aim was to determine the complementarity of the two systems, this work also helps us to gain a deeper insight into the distinct features and performances of the symbolic and statistical approaches to MWE extraction.

Our evaluation shows that USAS is an efficient tool for MWE extraction. Moreover, the system is such that a variety of analyses can be performed on the extracted MWEs, including classifying MWEs by semantic categories, selecting MWEs for requited semantic categories, and, potentially, helping map MWE translation equivalents across languages by matching their semantic categories.

The results of our experiment also demonstrate a marked complementary relationship between the semantic tagger and the statistical tool in terms of MWE extraction. The semantic tagger obtained a high precision in extracting core MWEs, including low-frequency ones, while the statistical tool showed a high performance in extracting domain-specific terms. Furthermore, only a small portion of overlap was observed between the outputs of these tools. By combining them together, thousands more MWEs could be extracted than by using the tools in isolation. An important implication of this finding is that, by incorporating the statistical tool with the semantic tagger, we can significantly improve the MWE coverage of the system, as well as its capability of extracting domain-specific MWEs without major expansion of the lexicon database. Such a system will provide an efficient and flexible tool for identifying and extracting MWEs in corpus-based language study and various NLP applications.

Acknowledgements

This work is continuing to be supported by the Benedict project, EU funded IST-2001-34237. We thank Janet Gough of Harper-Collins for her illuminating comments in relation to the criteria for identifying MWEs.

References

Biber, D., Conrad, S., Cortes, V., 2003. Lexical bundles in speech and writing: an initial taxonomy. In: Wilson, A., Rayson, P., McEnery, T. (Eds.), Corpus Linguistics by the Lune: A Festschrift for Geoffrey Leech. Peter Lang, Frankfurt, pp. 71–92.

Dagan, I., Church, K., 1994. Termight: identifying and translating technical terminology. In: Proceedings of the 4th Conference on Applied Natural Language Processing, Stuttgart, German, pp. 34–40.

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

Dias, G., 2003. Multiword unit hybrid extraction. In: Proceedings of the Workshop on Multiword Expressions: Analysis, Acquisition and Treatment, at ACL'03, Sapporo, Japan, pp. 41–48.

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

Fung, P., Church, K., 1994. K-vec: a new approach for aligning parallel texts. In: Proceedings of COLING'94, Kyoto, Japan, pp. 1996–2001.

Gaizauskas, R., Foster, J., Wilks, Y., Arundel, J., Clough, P., Piao, S., 2001. The METER corpus: a corpus for analysing journalistic text reuse. In: Proceedings of the Corpus Linguistics 2001, Lancaster, UK, pp. 214–223.

Gale, W., Church, K., 1991. Identifying word correspondences in parallel texts. In: The Fourth DARPA Workshop on Speech and Natural Language, Pacific Grove, CA, pp. 152–157.

Garside, R., Rayson, P., 1997. Higher-level annotation tools. In: Garside, R., Leech, G., McEnery, T. (Eds.), Corpus Annotation: Linguistic Information from Computer Text Corpora. Longman, London, pp. 179–193.

Garside, R., Smith, N., 1997. A hybrid grammatical tagger: CLAWS4. In: Garside, R., Leech, G., McEnery, A. (Eds.), Corpus Annotation: Linguistic Information from Computer Text Corpora. Longman, London, pp. 102–121.

Haruno, M., Yamazaki, T., 1996. High-performance bilingual text alignment using statistical and dictionary information. In: Proceedings of 34th Annual Meeting of the Association for Computational Linguistics, California, USA, pp. 131–138.

Lapata, M., Lascarides, A., 2003. Detecting novel compounds: the role of distributional evidence. In: Proceedings of the 11th Conference of the European Chapter for the Association of Computational Linguistics, pp. 235–242.

Lewandowska-Tomaszczyk, B., 2003. Ontologies and language corpora. In: Lewandowska-Tomaszczyk, B. (Ed.), PALC 2001: Practical Applications in Language Corpora. Peter Lang, Frankfurt.

Löfberg, L., Juntunen, J., Nykanen, A., Varantola, K., Rayson, P., Archer, D., 2004. Using a semantic tagger as dictionary search tool. In: Williams, G., Vessier, S. (Eds.), Proceedings of the EURALEX 2004. Université de Bretagne Sud, Lorient, France, pp. 127–134.

Manning, C., Schutze, H., 2000. Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge, MA.

Maynard, D., Ananiadou, S., 2000. Trucks: a model for automatic multiword term recognition. Journal of Natural Language Processing 8 (1), 101–126.

McEnery, T., Jean-Marc, L., Michael, O., Jean, V., 1997. The exploitation of multilingual annotated corpora for term extraction. In: Garside, R., Leech, G., McEnery, A. (Eds.), Corpus Annotation: Linguistic Information from Computer Text Corpora. Longman, London, pp. 220–230.

Merkel, M., Andersson, M., 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), Paris, France, pp. 737–746.

Michiels, A., Dufour, N., 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, Granada, Spain, pp. 1179–1186.

Moon, R., 1998. Fixed Expressions and Idioms in English: A Corpus-based Approach. Oxford University Press, Oxford.

Nivre, J., Nilsson, J., 2004. Multiword units in syntactic parsing. In: Proceedings of LREC-04 Workshop on Methodologies & Evaluation of Multiword Units in Real-world Applications, Lisbon, Portugal, pp. 37–46.

Onditi, V., Rayson, P., Ransom, B., Ramduny, D., Sommerville, I., Dix, A., 2004. Language resources and tools for supporting the system engineering process. In: Proceedings of 9th International Conference on Applications of Natural Language to Information Systems (NLDB 2004), Salford, UK, pp. 147–158.

Pereira, R., Crocker, P., Dias, G., 2004. A parallel multikey quicksort algorithm for mining multiword units. In: Proceedings of LREC-04 Workshop on Methodologies & Evaluation of Multiword Units in Real-world Applications, Lisbon, Portugal, pp. 17–23.

Piao, S., McEnery, T., 2001. Multi-word unit alignment in English–Chinese parallel corpora. In: Proceedings of the Corpus Linguistics 2001, Lancaster, UK, pp. 466–475.

Rayson, P., Archer, D., Piao, S., McEnery, T., 2004. The UCREL semantic analysis system. In: Proceedings of the LREC-04 Workshop, beyond Named Entity Recognition Semantic Labelling for NLP Tasks, Lisbon, Portugal, pp. 7–12.

Sag, I., Baldwin, T., Bond, F., Copestake, A., Dan, F., 2001. Multiword expressions: a pain in the neck for NLP. LinGO Working Paper No. 2001-03, Stanford University, CA.

Scott, M., 2001. Mapping key words to problem and solution. In: Scott, M., Thompson, G. (Eds.), Patterns of Text: In Honour of Michael Hoey. Benjamins, Amsterdam, pp. 109–127.

Smadja, F., 1993. Retrieving collocations from text: Xtract. Computational Linguistics 19 (1), 143–177.

Sawyer, P., Rayson, P., Garside, R., 2002. REVERE: support for requirements synthesis from documents. Information Systems Frontiers Journal 4 (3), 343–353.

Tanaka, T., Baldwin, T., 2003. Noun-noun compound machine translation: a feasibility study on shallow processing. In: Proceedings of the ACL-03 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment, Sapporo, Japan, pp. 17–24.

Tschichold, C., 2000. Multi-word Units in Natural Language Processing. Olms, New York, 2000.

Wermter, S., Chen, J., 1997. Cautious steps towards hybrid connectionist bilingual phrase alignment. In: Proceedings of the Conference on Recent Advances in Natural Language Processing, Sofia, Bulgaria, pp. 364–368.

Wu, D., 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. Computational Linguistics 23 (3), 377–401.