Statistically-Driven Alignment-Based Multiword Expression Identification for Technical Domains

Helena de Medeiros Caseli[◇], **Aline Villavicencio**[♣♠], **André Machado**[♣], **Maria José Finatto**[♡]

[◇]Department of Computer Science, Federal University of São Carlos (Brazil)
[♣]Institute of Informatics, Federal University of Rio Grande do Sul (Brazil)
[♠]Department of Computer Sciences, Bath University (UK)
[♡]Institute of Language and Linguistics, Federal University of Rio Grande do Sul (Brazil)
helenacaseli@dc.ufscar.br, avillavicencio@inf.ufrgs.br,
ammachado@inf.ufrgs.br, mfinatto@terra.com.br

Abstract

Multiword Expressions (MWEs) are one of the stumbling blocks for more precise Natural Language Processing (NLP) systems. Particularly, the lack of coverage of MWEs in resources can impact negatively on the performance of tasks and applications, and can lead to loss of information or communication errors. This is especially problematic in technical domains, where a significant portion of the vocabulary is composed of MWEs. This paper investigates the use of a statistically-driven alignment-based approach to the identification of MWEs in technical corpora. We look at the use of several sources of data, including parallel corpora, using English and Portuguese data from a corpus of Pediatrics, and examining how a second language can provide relevant cues for this tasks. We report results obtained by a combination of statistical measures and linguistic information, and compare these to the reported in the literature. Such an approach to the (semi-)automatic identification of MWEs can considerably speed up lexicographic work, providing a more targeted list of MWE candidates.

1 Introduction

A multiword expression (MWE) can be defined as any word combination for which the syntactic or semantic properties of the whole expression cannot be obtained from its parts (Sag et al., 2002). Examples of MWEs are phrasal verbs (*break down, rely on*), compounds (*police car, coffee machine*), idioms (*rock the boat, let the cat out of the bag*). They are very numerous in languages, as Biber et al. (1999) note, accouting for between 30% and 45% of spoken English and 21% of academic prose, and for Jackendoff (1997) the number of MWEs in a speaker's lexicon is of the same order of magnitude as the number of single words. However, these estimates are likely to be underestimates if we consider that for language from a specific domain the specialized vocabulary is going to consist largely of MWEs (*global warming, protein sequencing*) and new MWEs are constantly appearing (*weapons of mass destruction, axis of evil*).

Multiword expressions play an important role in Natural Language Processing (NLP) applications, which should not only identify the MWEs but also be able to deal with them when they are found (Fazly and Stevenson, 2007). Failing to identify MWEs may cause serious problems for many NLP tasks, especially those envolving some kind of semantic processing. For parsing, for instance, Baldwin et al. (2004), found that for a random sample of 20,000 strings from the British National Corpus (BNC) even with a broad-coverage grammar for English (Flickinger, 2000) missing MWEs accounted for 8% of total parsing errors. Therefore, there is an enormous need for robust (semi-)automated ways of acquiring lexical information for MWEs (Villavicencio et al., 2007) that can significantly extend the coverage of resources. For example, one can more than double the number of verb-particle constructions (VPCs) entries in a dictionary, such as the Alvey Natural Language Tools (Carroll and Grover, 1989), just extracting VPCs from a corpus like the BNC (Baldwin, 2005). Furthermore, as MWEs are language dependent and culturally motivated, identifying the adequate translation of MWE occurrences is an important challenge for machine translation methods.

In this paper, we investigate experimentally the use of an alignment-based approach for the identification of MWEs in technical corpora. We look at the use of several sources of data, including par-

Proceedings of the 2009 Workshop on Multiword Expressions, ACL-IJCNLP 2009, pages 1–8,
Suntec, Singapore, 6 August 2009. ©2009 ACL and AFNLP

allel corpora, using English and Portuguese data from a corpus of Pediatrics, and examining how a second language can provide relevant cues for this tasks. In this way, cost-effective tools for the automatic alignment of texts can generate a list of MWE candidates with their appropriate translations. Such an approach to the (semi-)automatic identification of MWEs can considerably speed up lexicographic work, providing a more targeted list of MWE candidates and their translations, for the construction of bilingual resources, and/or with some semantic information for monolingual resources.

The remainder of this paper is structured as follows. Section 2 briefly discusses MWEs and some previous works on methods for automatically extracting them. Section 3 presents the resources used while section 4 describes the methods proposed to extract MWEs as a statistically-driven by-product of an automatic word alignment process. Section 5 presents the evaluation methodology and analyses the results and section 6 finishes this paper with some conclusions and proposals for future work.

2 Related Work

The term Multiword Expression has been used to describe a large number of distinct but related phenomena, such as phrasal verbs (e.g. *come along*), nominal compounds (e.g. *frying pan*), institutionalised phrases (e.g. *bread and butter*), and many others (Sag et al., 2002). They are very frequent in everyday language and this is reflected in several existing grammars and lexical resources, where almost half of the entries are Multiword Expressions.

However, due to their heterogeneous characteristics, MWEs present a tough challenge for both linguistic and computational work (Sag et al., 2002). Some MWEs are fixed, and do not present internal variation, such as *ad hoc*, while others allow different degrees of internal variability and modification, such as *touch a nerve* (*touch/find a nerve*) and *spill beans* (*spill several/musical/mountains of beans*). In terms of semantics, some MWEs are more opaque in their meaning (e.g. *to kick the bucket* as *to die*), while others have more transparent meanings that can be inferred from the words in the MWE (e.g. *eat up*, where the particle *up* adds a completive sense to *eat*). Therefore, providing appropriate methods for the automatic identification and treatment of these phenomena is a real challenge for NLP systems.

A variety of approaches has been proposed for automatically identifying MWEs, differing basically in terms of the type of MWE and language to which they apply, and the sources of information they use. Although some work on MWEs is type independent (e.g. (Zhang et al., 2006; Villavicencio et al., 2007)), given the heterogeneity of MWEs much of the work looks instead at specific types of MWE like collocations (Pearce, 2002), compounds (Keller and Lapata, 2003) and VPCs (Baldwin, 2005; Villavicencio, 2005; Carlos Ramisch and Aline Villavicencio and Leonardo Moura and Marco Idiart, 2008). Some of these works concentrate on particular languages (e.g. (Pearce, 2002; Baldwin, 2005) for English and (Piao et al., 2006) for Chinese), but some work has also benefitted from asymmetries in languages, using information from one language to help deal with MWEs in the other (e.g. (na Villada Moirón and Tiedemann, 2006; Caseli et al., 2009)).

As basis for helping to determine whether a given sequence of words is in fact an MWE (e.g. *ad hoc* vs *the small boy*) some of these works employ linguistic knowledge for the task (Villavicencio, 2005), while others employ statistical methods (Pearce, 2002; Evert and Krenn, 2005; Zhang et al., 2006; Villavicencio et al., 2007) or combine them with some kinds of linguistic information such as syntactic and semantic properties (Baldwin and Villavicencio, 2002; Van de Cruys and na Villada Moirón, 2007) or automatic word alignment (na Villada Moirón and Tiedemann, 2006).

Statistical measures of association have been commonly used for this task, as they can be democratically applied to any language and MWE type. However, there is no consensus about which measure is best suited for identifying MWEs in general. Villavicencio et al. (2007) compared some of these measures (mutual information, permutation entropy and χ^2) for the type-independent detection of MWEs and found that Mutual Information seemed to differentiate MWEs from non-MWEs, but the same was not true of χ_2. In addition, Evert and Krenn (2005) found that for MWE identification the efficacy of a given measure depends on factors like the type of MWEs being targeted for identification, the domain and size of the corpora used, and the amount of low-frequency data excluded by adopting a threshold. Nonetheless, Villavicencio et al. (2007), discussing the influence of the corpus size and nature over the methods, found that these different measures have a high level of agreement about MWEs, whether in carefully constructed corpora or in more heterogeneous web-based ones. They also discuss the results obtained from adopting approaches like these for extending the coverage of resources, arguing that grammar coverage can be significantly increased if MWEs are properly identified and treated (Villavicencio et al., 2007).

Among the methods that use additional information along with statistics to extract MWE, the one proposed by na Villada Moirón and Tiedemann (2006) seems to be the most similar to our approach. The main difference between them is the way in which word alignment is used in the MWE extraction process. In this paper, the word alignment is the basis for the MWE extraction process while Villada Moirón and Tiedemann's method uses the alignment just for ranking the MWE candidates which were extracted on the basis of association measures (log-likelihood and salience) and head dependence heuristic (in parsed data).

Our approach, as described in details by Caseli et al. (2009), also follows to some extent that of Zhang et al. (2006), as missing lexical entries for MWEs and related constructions are detected via error mining methods, and this paper focuses on the extraction of generic MWEs as a by-product of an automatic word alignment. Another related work is the automatic detection of non-compositional compounds (NCC) by Melamed (1997) in which NCCs are identified by analyzing statistical translation models trained in a huge corpus by a time-demanding process.

Given this context, our approach proposes the use of alignment techniques for identifying MWEs, looking at sequences detected by the aligner as containing more than one word, which form the MWE candidates. As a result, sequences of two or more consecutive source words are treated as MWE candidates regardless of whether they are translated as one or more target words.

3 The Corpus and Reference Lists

The Corpus of Pediatrics used in these experiments contains 283 texts in Portuguese with a total of 785,448 words, extracted from the *Jornal de Pediatria*. From this corpus, the Pediatrics Glossary, a reference list containing multiword terms and recurring expressions, was semi-automatically constructed, and manually checked.[1] The primary aim of the Pediatrics Glossary, as an online resource for long-distance education, was to train, qualify and support translation students on the domain of pediatrics texts.

The Pediatrics Glossary was built from the 36,741 ngrams that occurred at least 5 times in the corpus. These were automatically cleaned or removed using some POS tag patterns (e.g. removing prepositions from terms that began or ended with them). In addition, if an ngram was part of a larger ngram, only the latter appeared in the Glossary, as is the case of *aleitamento materno* (*maternal breastfeeding*) which is excluded as it is contained in *aleitamento materno exclusivo* (*exclusive maternal breastfeeding*). This post-processing resulted in 3,645 ngrams, which were manually checked by translation students, and resulted in 2,407 terms, with 1,421 bigrams, 730 trigrams and 339 ngrams with n larger than 3 (not considered in the experiments presented in this paper).

4 Statistically-Driven and Alignment-Based methods

4.1 Statistically-Driven method

Statistical measures of association have been widely employed in the identification of MWEs. The idea behind their use is that they are an inexpensive language and type independent means of detecting recurrent patterns. As Firth famously said *a word is characterized by the company it keeps* and since we expect the component words of an MWE to occur frequently together, then these measures can give an indication of MWEness. In this way, if a group of words co-occurs with significantly high frequency when compared to the frequencies of the individual words, then they may form an MWE. Indeed, measures such as Pointwise Mutual Information (PMI), Mutual Information (MI), χ_2, log-likelihood (Press et al., 1992) and others have been employed for this task, and some of them seem to provide more accurate predictions of MWEness than others. In fact, in a comparison of some measures for the type-independent detection of MWEs, MI seemed

[1] Available in the TEXTQUIM/UFRGS website: `http://www.ufrgs.br/textquim`

to differentiate MWEs from non-MWEs, but the same was not true of χ_2 (Villavicencio et al., 2007). In this work we use two commonly employed measures for this task: PMI and MI, as implemented in the Ngram Statistics Package (Banerjee and Pedersen, 2003).

From the Portuguese portion of the Corpus of Pediatrics, 196,105 bigram and 362,663 trigram MWE candidates were generated, after filtering ngrams containing punctuation and numbers. In order to evaluate how these methods perform without any linguistic filtering, the only threshold employed was a frequency cut-off of 2 occurrences, resulting in 64,839 bigrams and 54,548 trigrams. Each of the four measures were then calculated for these ngrams, and we ranked each n-gram according to each of these measures. The average of all the rankings is used as the combined measure of the MWE candidates.

4.2 Alignment-Based method

The second of the MWE extraction approaches to be investigated in this paper is the alignment-based method. The automatic word alignment of two parallel texts — a text written in one (source) language and its translation to another (target) language — is the process of searching for correspondences between source and target words and sequences of words. For each word in a source sentence equivalences in the parallel target sentence are looked for. Therefore, taking into account a word alignment between a source word sequence S ($S = s_1 \ldots s_n$ with $n \geq 2$) and a target word sequence T ($T = t_1 \ldots t_m$ with $m \geq 1$), that is $S \leftrightarrow T$, the alignmet-based MWE extracion method assumes that: (a) S and T share some semantic features, and (b) S may be a MWE.

In other words, the alignment-based MWE extraction method states that the sequence S will be a MWE candidate if it is aligned with a sequence T composed of one or more words (a $n : m$ alignment with $n \geq 2$ and $m \geq 1$). For example, the sequence of two Portuguese words *aleitamento materno* — which occurs 202 times in the corpus used in our experiments — is a MWE candidate because these two words were joined to be aligned 184 times with the word *breastfeeding* (a $2 : 1$ alignment), 8 times with the word *breastfed* (a $2 : 1$ alignment), 2 times with *breastfeeding practice* (a $2 : 2$ alignment) and so on.

Thus, notice that the alignment-based MWE extraction method does not rely on the conceptual asymmetries between languages since it does not expect that a source sequence of words be aligned with a single target word. The method looks for the sequences of source words that are frequently joined together during the alignment despite the number of target words involved. These features indicate that the method priorizes precision in spite of recall.

It is also important to say that although the sequences of source and target words resemble the phrases used in the phrase-based statistical machine translation (SMT), they are indeed a refinement of them. More specifically, although both approaches rely on word alignments performed by `GIZA++`[2] (Och and Ney, 2000), in the alignment-based approach not all sequences of words are considered as phrases (and MWE candidates) but just those with an alignment $n : m$ ($n >= 2$) with a target sequence. To confirm this assumption a phrase-based SMT system was trained with the same corpus used in our experiments and the number of phrases extracted following both approaches were compared. While the SMT extracted 819,208 source phrases, our alignment-based approach (without applying any part-of-speech or frequency filter) extracted only 34,277. These results show that the alignment-based approach refines in some way the phrases of SMT systems.

In this paper, we investigate experimentally whether MWEs can be identified as a by-product of the automatic word alignment of parallel texts. We focus on Portuguese MWEs from the Corpus of Pediatrics and the evaluation is performed using the bigrams and trigrams from the Pediatrics Glossary as gold standard.

To perform the extraction of MWE candidates following the alignment-based approach, first, the original corpus had to be sentence and word aligned and Part-of-Speech (POS) tagged. For these preprocessing steps were used, respectively: a version of the Translation Corpus Aligner (TCA) (Hofland, 1996), the statistical word aligner `GIZA++` (Och and Ney, 2000) and the morphological analysers and POS taggers from `Apertium`[3] (Armentano-Oller et al., 2006).

[2] `GIZA++` is a well-known statistical word aligner that can be found at: `http://www.fjoch.com/GIZA++.html`

[3] `Apertium` is an open-source machine translation engine and toolbox available at: `http://www.apertium.org`.

From the preprocessed corpus, the MWE candidates are extracted as those in which two or more words have the same alignment, that is, they are linked to the same target unit. This initial list of MWE candidates is, then, filtered to remove those candidates that: (a) match some sequences of POS tags or words (patterns) defined in previous experiments (Caseli et al., 2009) or (b) whose frequency is below a certain threshold. The remaining units in the candidate list are considered to be MWEs.

Several filtering patterns and minimum frequency thresholds were tested and three of them are presented in details here. The first one (F1) is the same used during the manual building of the reference lists of MWEs: (a) patterns beginning with Article + Noun and beginning or finishing with verbs and (b) with a minimum frequency threshold of 5.

The second one (F2) is the same used in the (Caseli et al., 2009), mainly: (a) patterns beginning with determiner, auxiliary verb, pronoun, adverb, conjunction and surface forms such as those of the verb *to be* (*are*, *is*, *was*, *were*), relatives (*that*, *what*, *when*, *which*, *who*, *why*) and prepositions (*from*, *to*, *of*) and (b) with a minimum frequency threshold of 2.

And the third one (F3) is the same as (Caseli et al., 2009) plus: (a) patterns beginning or finishing with determiner, adverb, conjunction, preposition, verb, pronoun and numeral and (b) with a minimum frequency threshold of 2.

5 Experiments and Results

Table 1 shows the top 5 and the bottom 5 ranked candidates returned by PMI and the alignment-based approach. Although some of the results are good, especially the top candidates, there is still considerable noise among the candidates, as for instance *jogar video game* (lit. *play video game*). From table 1 it is also possible to notice that the alignment-based approach indeed extracts Pediatrics terms such as *aleitamento materno* (*breastfeeding*) and also other possible MWE that are not Pediatrics terms such as *estados unidos* (*United States*).

In table 2 we show the precision (number of correct candidates among the proposed ones), recall (number of correct candidates among those in reference lists) and F-measure ($(2 * precision * recall)/(precision + recall)$) figures for the association measures using all the candidates (on the first half of the table) and using the top 1,421 bigram and 730 trigram candidates (on the second half). From these latter results, we can see that the top candidates produced by these measures do not agree with the Pediatrics Glossary, since there are only at most 18.37% bigram and 6.03% trigram MWEs among the top candidates, as ranked by MI and PMI respectively. Interestingly, MI had a better performance for bigrams while for trigrams PMI performed better.

PMI	alignment-based
Online Mendelian Inheritance	faixa etária
Beta Technology Incorporated	aleitamento materno
Lange Beta Technology	estados unidos
Oxido Nitrico Inalatorio	hipertensão arterial
jogar video game	leite materno
...	...
e um de	couro cabeludo
e a do	bloqueio lactíferos
se que de	emocional anatomia
e a da	neonato a termo
e de nao	duplas mães bebês

Table 1: Top 5 and Bottom 5 MWE candidates ranked by PMI and alignment-based approach

`pt` MWE candidates	PMI	MI
# proposed bigrams	64,839	64,839
# correct MWEs	1403	1403
precision	2.16%	2.16%
recall	98.73%	98.73%
F	4.23%	4.23%
# proposed trigrams	54,548	54,548
# correct MWEs	701	701
precision	1.29%	1.29%
recall	96.03%	96.03%
F	2.55%	2.55%
# proposed bigrams	1,421	1,421
# correct MWEs	155	261
precision	10.91%	18.37%
recall	10.91%	18.37%
F	10.91%	18.37%
# proposed trigrams	730	730
# correct MWEs	44	20
precision	6.03%	2.74%
recall	6.03%	2.74%
F	6.03%	2.74%

Table 2: Evaluation of MWE candidates - PMI and MI

On the other hand, looking at the alignment-based method, 34,277 `pt` MWE candidates were extracted and Table 3 sumarizes the number of candidates filtered following the three filters described in 4.2: F1, F2 and F3.

To evaluate the efficacy of the alignment-based method in identifying multiword terms of Pediatrics, an automatic comparison was performed using the Pediatrics Glossary. In this auto-

pt MWE candidates	F1	F2	F3
# filtered by POS patterns	24,996	21,544	32,644
# filtered by frequency	9,012	11,855	1,442
# final Set	269	878	191

Table 3: Number of pt MWE candidates filtered in the alignment-based approach

pt MWE candidates	F1	F2	F3
# proposed bigrams	250	754	169
# correct MWEs	48	95	65
precision	19.20%	12.60%	38.46%
recall	3.38%	6.69%	4.57%
F	5.75%	8.74%	8.18%
# proposed trigrams	19	110	20
# correct MWEs	1	9	4
precision	5.26%	8.18%	20.00%
recall	0.14%	1.23%	0.55%
F	0.27%	2.14%	1.07%
# proposed bi/trigrams	269	864	189
# correct MWEs	49	104	69
precision	18.22%	12.04%	36,51%
recall	2.28%	4.83%	3.21%
F	4.05%	6.90%	5.90%

Table 4: Evaluation of MWE candidates

matic comparision we considered the final lists of MWEs candidates generated by each filter in table 3. The number of matching entries and the values for precision, recall and F-measure are showed in table 4.

The different values of extracted MWEs (in table 3) and evaluated ones (in table 4) are due to the restriction of considering only bigrams and trigrams in the Pediatrics Glossary. Then, longer MWEs — such as *doença arterial coronariana prematura* (*premature coronary artery disease*) and *pequenos para idade gestacional* (*small for gestational age*) — extracted by the alignment-based method are not being considered at the moment.

After the automatic comparison using the Pediatrics Glossary, an analysis by human experts was performed on one of the derived lists — that with the best precision values so far (from filter F3). The human analysis was necessary since, as stated in (Caseli et al., 2009), the coverage of reference lists may be low, and it is likely that a lot of MWE candidates that were not found in the Pediatrics Glossary are nonetheless true MWEs. In this paper only the pt MWE candidates extracted using filter F3 (as described in section 4.2) were manually evaluated.

From the 191 pt MWE candidates extracted after F3, 69 candidates (36.1% of the total amount) were found in the bigrams or trigrams in the Glossary (see table 4). Then, the remaining 122 candidates (63.9%) were analysed by two native-speakers human judges, who classified each of the 122 candidates as true, if it is a multiword expression, or false, otherwise independently of being a Pediatrics term. For the judges, a sequence of words was considered a MWE mainly if it was: (1) a proper name or (2) a sequence of words for which the meaning cannot be obtained by compounding the meanings of its component words.

The judgments of both judges were compared and a disagreement of approximately 12% on multiwords was verified. This disagreement was also measured by the kappa (K) measure (Carletta, 1996), with k = 0.73, which does not prevent conclusions to be drawn. According to Carletta (1996), among other authors, a value of k between 0.67 and 0.8 indicates a good agreement.

In order to calculate the percentage of true candidates among the 122, two approaches can be followed, depending on what criteria one wants to emphasize: precision or coverage (not recall because we are not calculating regarding a reference list). To emphasize the precision, one should consider as genuine MWEs only those candidates classified as true by both judges, on the other hand, to emphasize the coverage, one should consider also those candidates classified as true by just one of them. So, from 191 MWE candidates, 126 (65.97%) were classified as true by both judges and 145 (75.92%) by at least one of them.

6 Conclusions and Future Work

MWEs are a complex and heterogeneous set of phenomena that defy attempts to capture them fully, but due to their role in communication they need to be properly accounted for in NLP applications and tasks.

In this paper we investigated the identification of MWEs from technical domain, testing statistically-driven and alignment-based approaches for identifying MWEs from a Pediatrics parallel corpus. The alignment-based method generates a targeted precision-oriented list of MWE candidates, while the statistical methods produce recall-oriented results at the expense of precision. Therefore, the combination of these methods can produce a set of MWE candidates that is both more precise than the latter and has more coverage than the former. This can significantly speed up lexicographic work. Moreover, the results obtained show that in comparison with the manual extraction of MWEs, this approach can provide also a general set of MWE candidates in addition to the manually selected technical terms.

Using the alignment-based extraction method we notice that it is possible to extract MWEs that are Pediatrics terms with a precision of 38% for bigrams and 20% for trigrams, but with very low recall since only the MWEs in the Pediatrics Glossary were considered correct. However, after a manual analysis carried out by two native speakers of Portuguese we found that the percentage of true MWEs considered by both or at least one of them were, respectively, 65.97% and 75.92%. This was a significative improvement but it is important to say that, in this manual analysis, the human experts classified the MWEs as true independently of them being Pediatrics terms. So, as future work we intend to carry out a more carefull analysis with experts in Pediatrics to evaluate how many MWEs candidates are also Pediatrics terms.

In addition, we plan to investigate a weighted combination of these methods, favouring those that have better precision. Finally, we also intend to apply the results obtained in to the semi-automatic construction of ontologies.

Acknowledgments

We would like to thank the TEXTQUIM/UFRGS group for making the Corpus of Pediatrics and Pediatrics Glossary available to us. We also thank the financial support of FAPESP in bulding the parallel corpus. This research has been partly funded by the FINEP project COMUNICA.

References

Carme Armentano-Oller, Rafael C. Carrasco, Antonio M. Corbí-Bellot, Mikel L. Forcada, Mireia Ginestí-Rosell, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Gema Ramírez-Sánchez, Felipe Sánchez-Martínez, and Miriam A. Scalco. 2006. Open-source Portuguese-Spanish machine translation. In R. Vieira, P. Quaresma, M.G.V. Nunes, N.J. Mamede, C. Oliveira, and M.C. Dias, editors, *Proceedings of the 7th International Workshop on Computational Processing of Written and Spoken Portuguese, (PROPOR 2006)*, volume 3960 of *Lecture Notes in Computer Science*, pages 50–59. Springer-Verlag, May.

Timothy Baldwin and Aline Villavicencio. 2002. Extracting the Unextractable: A Case Study on Verb-particles. In *Proceedings of the 6th Conference on Natural Language Learning (CoNLL-2002)*, pages 98–104, Taipei, Taiwan.

Timothy Baldwin, Emily M. Bender, Dan Flickinger, Ara Kim, and Stephan Oepen. 2004. Road-testing the English Resource Grammar over the British National Corpus. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pages 2047–2050, Lisbon, Portugal.

Timothy Baldwin. 2005. The deep lexical acquisition of English verb-particles. *Computer Speech and Language, Special Issue on Multiword Expressions*, 19(4):398–414.

Satanjeev Banerjee and Ted Pedersen. 2003. The Design, Implementation and Use of the Ngram Statistics Package. In *In Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, pages 370–381.

Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Grammar of Spoken and Written English*. Longman, Harlow.

Jean Carletta. 1996. Assessing agreement on classification tasks: the kappa statistics. *Computational Linguistics*, 22(2):249–254.

Carlos Ramisch and Aline Villavicencio and Leonardo Moura and Marco Idiart. 2008. Picking them up and Figuring them out: Verb-Particle Constructions, Noise and Idiomaticity. In *Proceedings of the 12th Conference on Computational Natural Language Learning (CoNLL 2008)*, pages 49–56.

John Carroll and Claire Grover. 1989. The derivation of a large computational lexicon of English from LDOCE. In Bran Boguraev and Ted Briscoe, editors, *Computational Lexicography for Natural Language Processing*, pages 117–134. Longman, Harlow, UK.

Helena M. Caseli, Carlos Ramisch, Maria G. V. Nunes, and Aline Villavicencio. 2009. Alignment-based extraction of multiword expressions. *Language Resources and Evaluation*, to appear.

Stefan Evert and Brigitte Krenn. 2005. Using small random samples for the manual evaluation of statistical association measures. *Computer Speech and Language*, 19(4):450–466.

Afsaneh Fazly and Suzanne Stevenson. 2007. Distinguishing Subtypes of Multiword Expressions Using Linguistically-Motivated Statistical Measures. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, pages 9–16, Prague, June.

Dan Flickinger. 2000. On building a more efficient grammar by exploiting types. *Natural Language Engineering*, 6(1):15–28.

Knut Hofland. 1996. A program for aligning English and Norwegian sentences. In S. Hockey, N. Ide, and G. Perissinotto, editors, *Research in Humanities Computing*, pages 165–178, Oxford. Oxford University Press.

Ray Jackendoff. 1997. Twistin' the night away. *Language*, 73:534–59.

Frank Keller and Mirella Lapata. 2003. Using the Web to Obtain Frequencies for Unseen Bigrams. *Computational Linguistics*, 29(3):459–484.

I. Dan Melamed. 1997. Automatic Discovery of Non-Compositional Compounds in Parallel Data. In *eprint arXiv:cmp-lg/9706027*, pages 6027–+, June.

Bego na Villada Moirón and Jorg Tiedemann. 2006. Identifying idiomatic expressions using automatic word-alignment. In *Proceedings of the Workshop on Multi-word-expressions in a Multilingual Context (EACL-2006)*, pages 33–40, Trento, Italy.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the ACL*, pages 440–447, Hong Kong, China, October.

Darren Pearce. 2002. A Comparative Evaluation of Collocation Extraction Techniques. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, pages 1–7, Las Palmas, Canary Islands, Spain.

Scott S. L. Piao, Guangfan Sun, Paul Rayson, and Qi Yuan. 2006. Automatic Extraction of Chinese Multiword Expressions with a Statistical Tool. In *Proceedings of the Workshop on Multi-word-expressions in a Multilingual Context (EACL-2006)*, pages 17–24, Trento, Italy, April.

William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. 1992. *Numerical Recipes in C: The Art of Scientific Computing. Second edition.* Cambridge University Press.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing (CICLing-2002)*, volume 2276 of *(Lecture Notes in Computer Science)*, pages 1–15, London, UK. Springer-Verlag.

Tim Van de Cruys and Bego na Villada Moirón. 2007. Semantics-based Multiword Expression Extraction. In *Proceedings of the Workshop on A Broader Prespective on Multiword Expressions*, pages 25–32, Prague, June.

Aline Villavicencio, Valia Kordoni, Yi Zhang, Marco Idiart, and Carlos Ramisch. 2007. Validation and Evaluation of Automatically Acquired Multiword Expressions for Grammar Engineering. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1034–1043, Prague, June.

Aline Villavicencio. 2005. The Availability of Verb-Particle Constructions in Lexical Resources: How Much is Enough? *Journal of Computer Speech and Language Processing*, 19(4):415–432.

Yi Zhang, Valia Kordoni, Aline Villavicencio, and Marco Idiart. 2006. Automated Multiword Expression Prediction for Grammar Engineering. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 36–44, Sydney, Australia, July. Association for Computational Linguistics.