A Broad Evaluation of Techniques for Automatic Acquisition of Multiword Expressions

Carlos Ramisch[♣♠], **Vitor De Araujo**[♣], **Aline Villavicencio**[♣]
[♣]Federal University of Rio Grande do Sul (Brazil)
[♠] GETALP — LIG, University of Grenoble, France
{ceramisch, vbuaraujo, avillavicencio}@inf.ufrgs.br

Abstract

Several approaches have been proposed for the automatic acquisition of multiword expressions from corpora. However, there is no agreement about which of them presents the best cost-benefit ratio, as they have been evaluated in distinct datasets and/or languages. To address this issue, we investigate these techniques analysing the following dimensions: expression type (compound nouns, phrasal verbs), language (English, French) and corpus size. Results show that these techniques tend to extract similar candidate lists with high recall ($\sim$ 80%) for nominals and high precision ($\sim$ 70%) for verbals, while the use of association measures for candidate filtering is useful but with some of the measures being more onerous and not significantly better than raw counts. We finish with an evaluation of flexibility and an indication of which technique is recommended for each language-type-size context.

1 Introduction

Taking into account multiword expressions (MWEs) is important to confer naturalness to the output of NLP systems. An MT system, for instance, needs to be aware of idiomatic expressions like *raining cats and dogs* to avoid literal translations.[1] Likewise, a parser needs to deal with verb-particle expressions like *take off from Paris* and with light verb constructions like *take a walk along the river* in order to avoid PP-attachment errors.

Even though the last decade has seen considerable research in the automatic acquisition of MWEs, both in theoretical and in computational linguistics, to date there are few NLP applications integrating explicit MWE treatment. This may be partly explained by the complexity of MWEs: as they are heterogeneous and flexible (Sag et al., 2002), there is no unique push-button approach to identify all types of MWEs in all languages. Existing approaches are either generic but present relatively low precision or they require a large amount of language-specific resources to yield good results.

The goal of this paper is to evaluate approaches for the automatic acquisition of MWEs from corpora (§2), examining as parameters of the experimental context the language (English and French), type of target MWE (verbal and nominal) and size of corpus (small, medium, large). We focus on 4 approaches[2] and the experimental setup is presented in §3. In §4 we evaluate the following acquisition dimensions: quality of extracted candidates and of association measures, use of computational resources and flexibility. Thus, this research presents a comparative investigation of available approaches and indicates the best cost-benefit ratio in a given context (language, type, corpus size), pointing out current limitations and suggesting future avenues of research for the field.

2 MWE Acquisition Approaches

Efforts for the evaluation of MWE acquisition approaches usually focus on a single technique or compare the quality of association measures (AMs) used to rank a fixed annotated list of MWEs. For instance, Evert and Krenn (2005) and Seretan (2008) specifically evaluate and analyse the lexical AMs used in MWE extraction on small samples of bigram candidates. Pearce (2002), systematically evaluates a set of techniques for MWE extraction on a small test set of English collocations. Analogously, Pecina (2005) and Ramisch et al. (2008) present extensive comparisons of individual AMs and of their combination for MWE extraction in Czech, German and English. In this paper, our goal is not to compare individual AMs but acquisition approaches as a whole. There have also been efforts for the extrinsic evaluation of MWEs for NLP applications such as information retrieval (Xu et al., 2010), word sense disambiguation (Finlayson and Kulkarni, 2011) and MT (Carpuat and Diab, 2010).

[1]The equivalent expressions in French would be *raining ropes*, in German *raining young dogs*, in Portuguese *raining Swiss knives*, etc.

[2]We consider only freely available, downloadable and openly documented tools. Therefore, outside the scope of this work are proprietary tools, terminology and lexicography tools, translation aid tools and published techniques for which no available implementation is provided.

One recent initiative aiming at more comparable evaluations of MWE acquisition approaches was in the form of a shared task (Grégoire et al., 2008). However, the present work differs from the shared task in its aims. The latter considered only the ranking of precompiled MWE lists using AMs or linguistic filters at the end of extraction. As for many languages and domains no such lists are available, we are interested in the *whole acquisition process*. In addition, the evaluation results produced for the shared task may be difficult to generalise, as some of the evaluations prioritized the precision of the techniques without considering the recall or the novelty of the extracted MWEs. To date little has been said about the practical concerns involving MWE acquisition, like computational resources, flexibility or availability. With this work, we hope to help filling this gap by performing a broad evaluation considering many different parameters.

We focus on 4 approaches for MWE acquisition from corpora, which follow the general trend in the area of using shallow linguistic (lemmas, POS, stopwords) and/or statistical (counts, AMs) information to distinguishing ordinary sequences (e.g. *yellow dress*, *go to a concert*) from MWEs (e.g. *black box*, *go by a name*). In addition to the brief description below, Section 4.4 underlines the main differences between the approaches.

1. **LocalMaxs**[3] extracts MWEs by generating all possible *n*-grams from a sentence and then filtering them based on the local maxima of the AM's distribution (Silva and Lopes, 1999). It is based purely on word counts and is completely language independent, but it is not possible to directly integrate linguistic information in order to target a specific type of construction.[4]. The evaluation includes both *LocalMaxs Strict* which prioritizes high precision (henceforth `LocMax-S`) and *LocalMaxs Relaxed* which focuses on high recall (henceforth `LocMax-R`). A variation of the original algorithm, SENTA, has been proposed to deal with non-contiguous expressions (da Silva et al., 1999). However, it is computationally costly[5] and there is no freely available implementation.
2. **MWE toolkit**[6] (`mwetk`) is an environment for type and language-independent MWE acquisition, integrating linguistic and frequency information (Ramisch et al., 2010). It generates a targeted list of MWE candidates extracted and filtered according to user-defined criteria like POS sequences and a set of statistical AMs. It is an integrated framework for MWE treatment, providing from corpus preprocessing facilities to the automatic evaluation of the resulting list with respect to a reference. Its input is a corpus annotated with POS, lemmas and dependency syntax, or if these are not available, raw text.
3. **Ngram Statistics Package**[7] (`NSP`) is a traditional approach for the statistical analysis of *n*-grams in texts (Pedersen et al., 2011). It provides tools for counting *n*-grams and calculating AMs, where an *n*-gram is a sequence of *n* contiguous words or *n* words occurring in a window of *w* words in a sentence. While most of the measures are only applicable to bigrams, some of them are also extended to trigrams and 4-grams. The set of available AMs includes robust and theoretically sound measures such as log-likelihood and Fischer's exact test. Although there is no direct support to linguistic information such as POS, it is possible to simulate them to some extent using the same workaround as for `LocMax`.
4. **UCS toolkit**[8] provides a large set of sophisticated AMs. It focuses on high accuracy calculations for bigram AMs, but unlike the other approaches, it starts from a list of candidates and their respective frequencies, relying on external tools for corpus preprocessing and candidate extraction. Therefore, questions concerning contiguous *n*-grams and support of linguistic filters are not dealt with by `UCS`, In our experiments, we will use the list of candidates generated by `mwetk` as input for `UCS`..

	Small	**M**edium	**L**arge
# sentences	5,000	50,000	500,000
# `en` words	133,859	1,355,482	13,164,654
# `fr` words	145,888	1,483,428	14,584,617

Table 1: Number of sentences and of words of each fragment of the Europarl corpus in `fr` and in `en`.

As the focus of this work is on MWE acquisition (identification and extraction), other tasks related to MWE treatment, namely interpretation, classification and applications (Anastasiou et al., 2009), are not considered in this paper. This is the case, for instance, of approaches for dictionary-based in-context MWE token identification requiring an initial dictionary of valid MWEs, like jMWE (Kulkarni and Finlayson, 2011).

3 Experimental Setup

For comparative purposes, we investigate the acquisition of MWEs in two languages, English (`en`) and French

[3] `http://hlt.di.fct.unl.pt/luis/multiwords/index.html`

[4] Although this can be simulated by concatenating words and POS tags together in order to form a token.

[5] It is based on the calculation of all possible *n*-grams in a sentence, which explode in number when going from contiguous to non-contiguous *n*-grams.

[6] `http://mwetoolkit.sourceforge.net`

[7] `http://search.cpan.org/dist/Text-NSP`

[8] `http://www.collocations.de/software.html`

(fr), analysing nominal and verbal expressions in en and nominal in fr,[9] obtained with the following rules:

- **Nominal expressions** en: a noun preceded by a sequence of one or more nouns or adjectives: *European Union*, *clock radio*, *clown anemone fish*.
- **Nominal expressions** fr: a noun followed by either an adjective or a prepositional complement (with the prepositions *de*, *à* and *en*) followed by an optionally determined noun: *algue verte*, *aliénation de bien*, *allergie à la poussière*.
- **Verbal expressions** en: verb-particle constructions formed by a verb (except *be* and *have*) followed by a prepositional particle[10] not further than 5 words after it: *give up*, *switch the old computer off*.

To test the influence of corpus size on performance, three fragments of the en and fr parts of the Europarl corpus v3[11], were used as test corpora: (**S**)mall, (**M**)edium and (**L**)arge, Table 1.

The extracted MWEs were automatically evaluated against the following gold standards: WordNet 3, the Cambridge Dictionary of Phrasal Verbs, and the VPC (Baldwin, 2008) and CN (Kim and Baldwin, 2008) datasets [12] for en; the Lexique-Grammaire [13] for fr. The total number of entries is listed below, along with the number of entries occurring at least twice in each corpus (in parentheses), which was the criterion used to calculate recall in § 4.1:

- Nominal expressions en: 59,683 entries (**S**: 122, **M**: 764, **L**: 2,710);
- Nominal expressions fr: 69,118 entries (**S**: 220, **M**: 1,406, **L**: 4,747);
- Verbal expressions en: 1,846 entries (**S**: 699, **M**: 1,846, **L**: 1,846).

4 Evaluation Results

The evaluation of MWE acquisition is an open problem. While classical measures like precision and recall assume that a complete (or at least broad-coverage) gold standard exists, manual annotation of top-*n* candidates and mean average precision (MAP) are labour-intensive even when applied to a small sample, emphasizing precision regardless of the number of acquired *new* expressions. As approaches differ in the way they allow the description of extraction criteria, we evaluate candidate extraction separately from AMs.

[9] As fr does not present many verb-particle constructions and due to the lack of availability of resource for other types of fr verbal expressions (e.g. light verb constructions), only nominal expressions are considered.

[10] *up, off, down, back, away, in, on.*

[11] http://www.statmt.org/europarl/

[12] The latter are available from http://multiword.sf.net/

[13] http://infolingu.univ-mlv.fr/

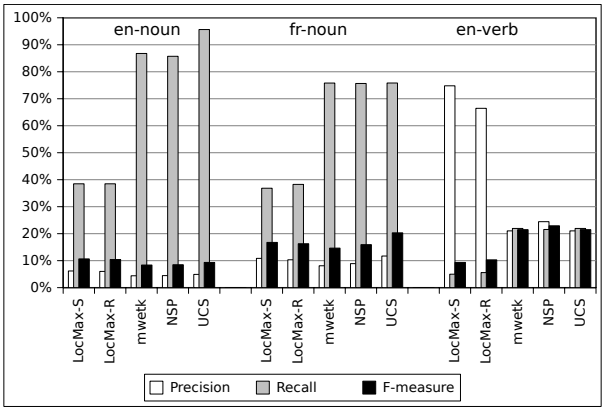

Figure 1: Quality of candidates extracted from medium corpus, comparison across languages/MWE types.

4.1 Extracted Candidates

We consider as *MWE candidates* the initial set of sequences before any AM is applied. Candidate extraction is performed through the application of patterns describing the target MWEs in terms of POS sequences, as described in § 3. To minimise potential cases of noise, candidates occurring only once in the corpus were discarded. We compare the quality of these candidates in terms of (P)recision, (R)ecall and (F)-measure using the gold standard references described in § 3. These measures are underestimations as they assume that candidates not in the gold standard are false MWEs, whereas they may simply be absent due to coverage limitations.

The quality of candidates extracted from the medium-size corpus (M) varies across MWE types/languages, as shown in Figure 1. The candidates for UCS are obtained by keeping only the bigrams in the candidate list returned by the mwetk. For nominal MWEs, the approaches have similar patterns of performance in the two languages, with high recall and low precision yielding an F-measure of around 10 to 15%. The variation between en and fr can be partly explained by the differences in size of the gold standards for each of these languages. Further research would be needed to determine to what degree the characteristics of these languages and the set of extraction patterns influence these results. For verbal expressions, LocMax has high precision (around 70%) but low recall while the other approaches have more balanced P and R values around 20%. This is partly due to the need for simulating POS filters for extraction of verbal MWE candidates with LocMax which included only contiguous *n*-grams in which the first and the last word matched verb+particle pattern and intervening words were manually removed. This step to extract discontiguous MWEs is incorporated in the other approaches.

The techniques differ in terms of extraction strategy: (i) mwetk and NSP allow the definition of linguistic fil-

	S	M	L
		LocMax-S	
P	7.53%	6.18%	4.50%
R	42.62%	38.48%	37.42%
		LocMax-R	
P	7.46%	6.02%	—
R	42.62%	38.48%	—
		P-mwetk	
P	6.50%	4.40%	2.35%
R	83.61%	86.78%	89.23%
		NSP	
P	6.61%	4.46%	2.48%
R	83.61%	85.73%	89.41%
		UCS	
P	6.96%	4.91%	2.77%
R	96.19%	95.65%	96.88%

Table 2: (P)recision and (R)ecall of en nominal candidates, comparison across corpus sizes (S)mall, (M)edium and (L)arge.

ters while LocMax only allows the application of grep-like filters after extraction; (ii) there is no preliminary filtering in mwetk and NSP, they simply return all candidates matching a pattern, while LocMax filters the candidates based on the local maxima criterion; (iii) LocMax only extracts contiguous candidates while the others allow discontiguous candidates. The way mwetk and NSP extract discontiguous candidates differs: the former extracts all verbs with particles no further than 5 positions to the right.NSP extracts bigrams in a window of 5 words, and then filters the list keeping only those in which the first word is a verb and that contain a particle. However, the results are similar, with slightly better values for NSP.

The evaluation of en nominal candidates according to corpus size is shown in Table 2.[14] For all approaches, precision decreases when the corpus size increases as more noise is returned, while recall increases for all except LocMax. This may be due to the latter ignoring smaller *n*-grams when larger candidates containing them become sufficiently frequent, as is the case when the corpus increases.Table 3 shows that the candidates extracted by LocMax are almost completely covered by the candidates extracted by the other approaches. The relaxed version extracts slighly more candidates, but still much less than mwetk, NSP and UCS, which all extract a similar set of candidates. In order to distinguish the performance of the approaches, we need to analyse the AMs they use to rank the candidates.

4.2 Association Measures

Traditionally, to evaluate an AM, the candidates are ranked according to it and a threshold value is applied, below which the candidates are discarded. However, if we average the precision considering all true MWEs as threshold points, we obtain the mean average precision (MAP) of the measure without setting a hard threshold. Due to length limitations, we cannot detail the calculation of AMs; please refer to the documentation of each approach, cited in § 2, for more details.

[14] It was not possible to evaluate LocMax-R on the large corpus as the provided implementation did not support corpora of this magnitude.

	LocMax-S	LocMax-R	mwetk	NSP	UCS	Total verbs
LocMax-S	—	124	124	122	124	124
LocMax-R	4747	—	156	153	156	156
mwetk	4738	4862	—	1565	1926	1926
NSP	4756	4879	14611	—	1565	1629
UCS	4377	4364	13407	13045	—	1926
Total nouns	4760	4884	15064	14682	13418	

Table 3: Intersection of the candidate lists extracted from medium corpus. Nominal candidates en in bottom left, verbal candidates en in top right.

Table 4 presents the MAP values for the tested AMs applied to the candidates extracted from the large corpus (L), where the larger the value, the better the performance. We used as baseline the assignment of a random score and the use of the raw frequency for the candidates. Except for mwetk:t and mwetk:pmi, all MAP values are significantly different from the two baselines, with a two-tailed t test for difference of means assuming unequal sample sizes and variances (p-value < 0.005).

The LocMax:glue AM performs best for all types of MWEs, suggesting local maxima as a good MWE indicator and also the efficacy of the AM used to filter the candidates as it generates highly precise results (considering the difficulty of this task). On the other hand this approach returns a small set of candidates and this may be problematic depending on the task (e.g. for building a wide-coverage lexicon). For mwetk, the best overall AM is the Dice coefficient; the other measures are not consistently better than the baseline, or perform better for one MWE type than for the other. The Poisson-Stirling (ps) measure performed quite well, while the other two measures tested for NSP performed below baseline for some cases. Finally, as we expected, the AMs applied by UCS perform all above baseline and, for nominal MWEs, are comparable to the best AM (e.g. Poisson.pv and local.MI). The MAP for verbal expressions varies a lot for UCS (from 30% to 53%), but none of the measures comes close to the MAP of the glue (87.06%). None of the approaches provides a straightforward method for choosing or combining different AMs.

4.3 Computational resources

In the decision of which measure to adopt, factors like the degree of MWE flexibility and computational perfor-

	en noun	fr noun	en verb
		Baseline	
random	2.749	6.1072	17.2079
freq	4.7478	8.7946	22.7155
		LocMax-S	
glue	6.9901	12.9383	87.0614
		mwetk	
dice	5.7783	9.5419	46.3609
t-test	5.0907	8.6373	26.4185
pmi	2.7589	2.9173	53.5591
log-lik.	3.166	5.5176	45.8837
		NSP	
pmi	2.9902	7.6782	62.1689
ps	5.3985	12.3791	57.6238
tmi	2.108	4.8928	19.8009
		UCS	
z.score	6.1202	11.7657	46.8707
Poisson.pv	6.5858	12.8226	32.7737
MI	5.1465	9.3363	53.5591
relative.risk	5.0999	9.2919	46.6702
odds.ratio	5.0364	9.2104	50.2201
gmean	6.0101	11.524	45.6089
local.MI	6.4294	12.7779	29.9858

Table 4: Mean average precision of AMs in large corpus.

mance of the AM may also be taken into account. For instance, the Dice coefficient can be applied to any length of *n*-gram quite fast while more sophisticated measures like Poisson.pv can be applied only to 2-grams and sometimes use lots of computational resources. Even if one could argue that we can be lenient towards a slow offline extraction process, the extra waiting may not be worth a slight quality improvement. Moreover, memory limitations are an issue if no large computer clusters are available.

In Figure 2, we plotted in log-scale the time in seconds used by each approach to extract nominal and verbal expressions in en, using a dedicated 2.4GHz quad-core Linux machine with 4Gb RAM. For nominal expressions the time needed to extract grows linearly with the size of the corpus, whereas for verbal expressions it seems that time increases faster than the size of the corpus. UCS is the slowest approach for both MWE types while NSP and LocMax-S are the fastest. However, it is important to emphasize that NSP consumed more than 3Gb memory to extract 4- and 5-grams from the large corpus and LocMax-R could not handle the large corpus at all. In theory, all techniques can be applied to arbitrarily large corpora if we used a map-reduce approach (e.g. NSP provides tools to split and join the corpus). However, the goal of this evaluation is to discover the performance of the techniques with no manual optimization. In this sense, the mwetk seems to provide an average trade-off between quality and resources used.

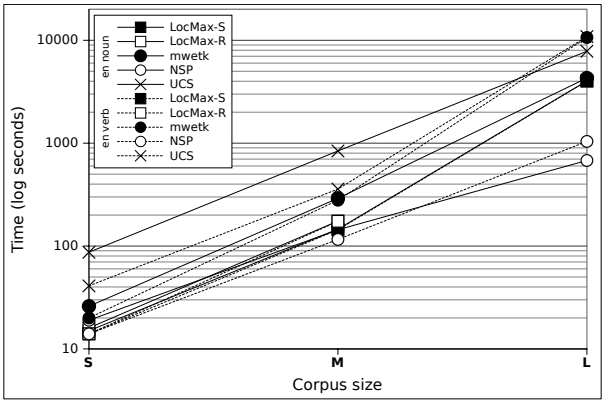

Figure 2: Time (seconds, log scale) to extract en nouns (bold line) and verbs (dashed line) from corpora.

	LocMax	mwetk	NSP	UCS
candidate extraction	Yes	Yes	Yes	No
n-grams with $n > 2$	Yes	Yes	Yes	No
Discontiguous MWE	No	Yes	Yes	—
Linguistic filter	No	Yes	No	No
Robust AMs	No	No	Yes	Yes
Large corpora	Partly	Yes	Yes	No
Availability	Free	Free	Free	Free

Table 5: Summary of tools for MWE acquisition.

4.4 Flexibility

In Table 5 we summarise the characteristics of the evaluated approaches. Among them, UCS does not perform candidate extraction from corpora but accepts as input a list of bigrams and their occurrence counts. While it only supports *n*-grams of size 2, NSP implements some of the AMs for 3 and 4-grams and mwetk and LocMax have no constraint on the number of words. LocMax extracts only contiguous MWEs while mwetk allows the extraction of unrestrictedly distant words and NSP allows the specification of a window of maximum *w* ignored words between each two words of the candidate. Only mwetk integrates linguistic filters on the lemma, POS and syntactic annotation, but this was performed using external tools (sed/grep) for the other approaches with similar results. The AMs implemented by LocMax and mwetk are conceived for any size of *n*-gram and are thus less statistically sound than the clearly designed measures used by UCS and, to some extent, by NSP (Fisher test). The large corpus used in our experiments was not supported by LocMax-R version, but LocMax-S has a version that deals with large corpora, as well as mwetk and NSP. Finally, all of these approaches are freely available for download and documented on the web.

5 Conclusions and future work

We evaluated the automatic acquisition of MWEs from corpora. The dimensions evaluated were type of construction (for flexibility and contiguity), language and corpus size. We evaluated two steps separately: candidate extraction and filtering with AMs. Candidate lists are very similar, with approaches like `mwetk` and `NSP` returning more candidates (they cover most of the nominal MWEs in the corpus) but having lower precision. `LocMax-S` presented a remarkably high precision for verbal expressions. However, the choice of an AM may not only take into account its MAP but also its flexibility and the computational resources used.[15] Our results suggest that the approaches could be combined using machine learning (Pecina, 2005).

In the future, we would like to develop this evaluation further by taking into account other characteristics such as the domain and genre of the source corpus. Such evaluation would be useful to guide future research on specialised multiword terminology extraction, determining differences with respect to generic MWE extraction. We would also like to evaluate other MWE-related tasks (e.g. classification, interpretation) and also dictionary-based identification (Kulkarni and Finlayson, 2011) and bilingual MWE acquisition (Carpuat and Diab, 2010). Finally, we believe that an application-based extrinsic evaluation involving manual validation of candidates would ultimately demonstrate the usefulness of current MWE acquisition techniques.

References

Dimitra Anastasiou, Chikara Hashimoto, Preslav Nakov, and Su Nam Kim, editors. 2009. *Proc. of the ACL Workshop on MWEs: Identification, Interpretation, Disambiguation, Applications (MWE 2009)*, Suntec, Singapore, Aug. ACL.

Timothy Baldwin. 2008. A resource for evaluating the deep lexical acquisition of english verb-particle constructions. In Grégoire et al. (Grégoire et al., 2008), pages 1–2.

Marine Carpuat and Mona Diab. 2010. Task-based evaluation of multiword expressions: a pilot study in statistical machine translation. In *Proc. of HLT: The 2010 Annual Conf. of the NAACL (NAACL 2003)*, pages 242–245, Los Angeles, California, Jun. ACL.

Joaquim Ferreira da Silva, Gaël Dias, Sylvie Guilloré, and José Gabriel Pereira Lopes. 1999. Using localmaxs algorithm for the extraction of contiguous and non-contiguous multiword lexical units. In *Proceedings of the 9th Portuguese Conference on Artificial Intelligence: Progress in Artificial Intelligence*, EPIA '99, pages 113–132, London, UK. Springer.

Stefan Evert and Brigitte Krenn. 2005. Using small random samples for the manual evaluation of statistical association measures. *Comp. Speech & Lang. Special issue on MWEs*, 19(4):450–466.

Mark Finlayson and Nidhi Kulkarni. 2011. Detecting multiword expressions improves word sense disambiguation. In Kordoni et al. (Kordoni et al., 2011), pages 20–24.

Nicole Grégoire, Stefan Evert, and Brigitte Krenn, editors. 2008. *Proc. of the LREC Workshop Towards a Shared Task for MWEs (MWE 2008)*, Marrakech, Morocco, Jun.

Su Nam Kim and Timothy Baldwin. 2008. Standardised evaluation of english noun compound interpretation. In Grégoire et al. (Grégoire et al., 2008), pages 39–42.

Valia Kordoni, Carlos Ramisch, and Aline Villavicencio, editors. 2011. *Proc.of the ACL Workshop on MWEs: from Parsing and Generation to the Real World (MWE 2011)*, Portland, OR, USA, Jun. ACL.

Nidhi Kulkarni and Mark Finlayson. 2011. jMWE: A java toolkit for detecting multi-word expressions. In Kordoni et al. (Kordoni et al., 2011), pages 122–124.

Éric Laporte, Preslav Nakov, Carlos Ramisch, and Aline Villavicencio, editors. 2010. *Proc.of the COLING Workshop on MWEs: from Theory to Applications (MWE 2010)*, Beijing, China, Aug. ACL.

Darren Pearce. 2002. A comparative evaluation of collocation extraction techniques. In *Proc. of the Third LREC (LREC 2002)*, Las Palmas, Canary Islands, Spain, May. ELRA.

Pavel Pecina. 2005. An extensive empirical study of collocation extraction methods. In *Proc. of the ACL 2005 SRW*, pages 13–18, Ann Arbor, MI, USA, Jun. ACL.

Ted Pedersen, Satanjeev Banerjee, Bridget McInnes, Saiyam Kohli, Mahesh Joshi, and Ying Liu. 2011. The ngram statistics package (text::NSP) : A flexible tool for identifying ngrams, collocations, and word associations. In Kordoni et al. (Kordoni et al., 2011), pages 131–133.

Carlos Ramisch, Paulo Schreiner, Marco Idiart, and Aline Villavicencio. 2008. An evaluation of methods for the extraction of multiword expressions. In Grégoire et al. (Grégoire et al., 2008), pages 50–53.

Carlos Ramisch, Aline Villavicencio, and Christian Boitet. 2010. Multiword expressions in the wild? the mwetoolkit comes in handy. In Yang Liu and Ting Liu, editors, *Proc. of the 23rd COLING (COLING 2010) — Demonstrations*, pages 57–60, Beijing, China, Aug. The Coling 2010 Organizing Committee.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proc. of the 3rd CICLing (CICLing-2002)*, volume 2276/2010 of *LNCS*, pages 1–15, Mexico City, Mexico, Feb. Springer.

Violeta Seretan. 2008. *Collocation extraction based on syntactic parsing*. Ph.D. thesis, University of Geneva, Geneva, Switzerland.

Joaquim Silva and Gabriel Lopes. 1999. A local maxima method and a fair dispersion normalization for extracting multi-word units from corpora. In *Proceedings of the Sixth Meeting on Mathematics of Language (MOL6)*, pages 369–381, Orlando, FL, USA, Jul.

Ying Xu, Randy Goebel, Christoph Ringlstetter, and Grzegorz Kondrak. 2010. Application of the tightness continuum measure to chinese information retrieval. In Laporte et al. (Laporte et al., 2010), pages 54–62.

[15] The data used in experiments is available at `http://www.inf.ufrgs.br/~ceramisch/?page=downloads/mwecompare`.