Identification of Multiword Expressions in the brWaC

Rodrigo Augusto Scheller Boos, Kassius Vargas Prestes, Aline Villavicencio

Institute of Informatics
Federal University of Rio Grande do Sul, Brazil
rodrigoschellerboos@gmail.com, kvprestes@inf.ufrgs.br, avillavicencio@inf.ufrgs.br

Abstract

Although corpus size is a well known factor that affects the performance of many NLP tasks, for many languages large freely available corpora are still scarce. In this paper we describe one effort to build a very large corpus for Brazilian Portuguese, the brWaC, generated following the Web as Corpus kool yinitiative. To indirectly assess the quality of the resulting corpus we examined the impact of corpus origin in a specific task, the identification of Multiword Expressions with association measures, against a standard corpus. Focusing on nominal compounds, the expressions obtained from each corpus are of comparable quality and indicate that corpus origin has no impact on this task.

Keywords: Multiword Expressions, Corpora, Web-crawling

1. Introduction

Several studies have found corpus size to be an important factor affecting the performance of many tasks and applications, from spelling correction (Banko and Brill, 2001) and case frame acquisition (Sasano et al., 2009) to information retrieval (Talvensaari, 2008) and machine translation (Brants et al., 2007). For instance, comparing Japanese corpora of size up to 100 billion words, Sasano et al. (2009) got better results for case frame acquisition with larger corpora, and the performance was not saturated even using the full corpus. Brants et al. (2007) used increasing corpus size of up to 2 trillion tokens for training language models for machine translation, with translation quality rising with corpus size.

Large corpora are crucial for tasks like distributional thesaurus construction (Lin, 1998; Baroni and Lenci, 2010), subcategorization frame acquisition (Korhonen et al., 2006) and identification of multiword expressions (MWEs) (Evert and Krenn, 2005; Ramisch et al., 2010; Tsvetkov and Wintner, 2010). Initiatives for constructing very large corpora have increased in recent years, especially using the Web as corpus.

Approaches for using the Web as the basis for building very large corpora employ crawlers to collect sets of texts which are subsequently cleaned to extract only the textual contents and filtered to remove duplicate material and noise from texts that have little human produced content. The WaCky (Web-As-Corpus Kool Yinitiative)[1] approach (Ferraresi et al., 2008) in particular has been used to build very large corpora for languages like English, Italian and German with over a billion words each. This method starts from a list of medium frequency content seed words and produces general purpose corpora which have a good level of content variation and quantity of information. Approaches to collect comparable corpora, on the other hand, use focused crawling, starting from a domain specific page and guiding the crawler based on the proportion of relevant words that can be found in the text of that page and of all the other pages that belong to the same host (Talvensaari, 2008; Granada et al., 2012; Laranjeira et al., 2014). Similarly, parallel corpora have been collected identifying sites with equivalent pages in multiple languages (Barbosa et al., 2012). These initiatives represent an inexpensive way of creating a large resource, especially for languages for which freely available resources of this magnitude are still scarce. In this work we adopt the WaCky method for collecting a very large corpus for Brazilian Portuguese, the brWaC.

As a case study of the quality of the resulting corpus in this paper brWaC is used as basis for the identification of MWEs like noun compounds (*general practitioner*) and multiword terminology (*artificial intelligence, machine learning*). Although MWEs are often employed in general and technical language, their automatic identification based on association measures (AMs) is often limited by their low token frequency in standard corpora. For instance, in the 100M word British National Corpus (Burnard, 2000), 30.85% of 2318 phrasal verbs are found less than 5 times, with 12.77% occurring only once (Villavicencio, 2005). As some of these AMs are sensitive to low frequencies and as MWEs have lower frequencies than single words, for accurate MWE identification it is important not only to have a very large corpus but to maintain corpus quality. In this work we look at the impact of corpus quality in terms of MWE identification, examining possible differences in accuracy of a set of AMs using the brWaC and a standard corpus, the CETENFolha for Portuguese[2] The results obtained confirm that using a very large web-generated corpus for a task like MWE identification it is possible to maintain comparable quality to standard corpora.

This paper is structured as follows: we start with a discussion of Multiword Expressions in section 2 and of the WaCky approach in section 3. The methodology followed in this work is presented in section 4, the evaluation in section 5 and the results obtained in section 6. We finish with some conclusions and future work.

[1] http://wacky.sslmit.unibo.it/doku.php.

[2] http://www.linguateca.pt/CETENFolha/

2. Multiword Expressions

MWEs can be defined as *a sequence of words that acts as a single unit at some level of linguistic analysis* (Calzolari et al., 2002) and characterized by idiosyncratic features, be them lexical, syntactic, semantic, pragmatic and/or statistic, and at one or more of these levels (Kim and Baldwin, 2010). They include compound nouns (*apple juice, federal government*) idioms (*let the cat out of the bag*), phrasal verbs (*break down*) and collocations (*salt and pepper*). The relevance of MWEs for (computational) linguistic related studies can also be assessed when one considers some estimates about their frequency in language. For Biber et al. (1999) between 30% and 45% of spoken English and 21% of academic prose corresponds to MWEs, while for Jackendoff (1997) the number of MWEs in a speakers lexicon is of the same order of magnitude as the number of single words, and new expressions are constantly being coined (*weapons of mass destruction, social media, big data*). However, in spite of the considerable occurrence of MWEs in general and technical language, lexical resources in general have limited coverage particularly for domain specific MWEs, and some of the means adopted for increasing their coverage include techniques for the automatic identification of MWEs from corpora.

A variety of approaches has been proposed for automatically identifying MWEs, differing in terms of the type of MWE and language to which they apply, and the techniques they use. Although some work on MWEs is type independent, given the heterogeneity of MWEs much of the work looks instead at specific types of MWE like collocations (Evert and Krenn, 2005; Pearce, 2002; Smadja, 1993), compounds and terms (Lapata and Lascarides, 2003; Daille, 2012) and VPCs (Baldwin, 2005; Ramisch et al., 2008). Some of these works concentrate on particular languages like English (Pearce, 2002; Lapata and Lascarides, 2003; Baldwin, 2005), German (Evert and Krenn, 2005), Spanish (Moreno-Ortiz et al., 2013), Basque (Gurrutxaga and Alegria, 2013), Persian (Samvelian and Faghiri, 2013) and Portuguese (de Caseli et al., 2010; Antunes and Mendes, 2013; Sanches Duran et al., 2013), with multilingual works using information from one language to help deal with MWEs in the other (Tsvetkov and Wintner, 2010; Daille, 2012).

Techniques for helping to determine whether a given sequence of words is in fact an MWE (*give a gift* vs. *give a speech*) usually include a candidate extraction or generation step and a filtering step (Ramisch, 2012). The former often employs patterns which range from n-grams to part-of-speech (POS) tag sequences (Justeson and Katz, 1995; Baldwin, 2005; de Caseli et al., 2010) and syntactic relations (Seretan, 2008), for a more precise generation of candidates that match the relevant target linguistic variations (e.g. verbs and particles for verb-particle constructions). For candidate filtering statistical association measures (AMs) are often used (Pearce, 2002; Evert and Krenn, 2005; Ramisch et al., 2012). The idea behind the use of AMs is that they are an inexpensive language and type independent means of detecting recurrent patterns, assuming that if a group of words occurs with significantly high relative frequency when compared to the frequencies of the individual words, the more likely it is that they form an MWE. A large selection of AMs is available for MWE identification as discussed by Pecina (2010), and among them some commonly used AMs are χ^2, pointwise mutual information (PMI), log-likelihood and c-value (Frantzi et al., 2000) for terms, along with their combination. Given a set of candidate MWEs these AMs rank them according to the degree of association between the words, and these ranks may differ according to the AMs used. In a comparison of some measures for the type-independent detection of MWEs, Mutual Information seemed to differentiate MWEs from non-MWEs, but the same was not true of Pearsons χ^2 (Villavicencio et al., 2007), while PMI can be sensitive to low frequencies, so that infrequent word pairs are assigned high values and can dominate the top of the rank (Bouma, 2009).

Corpus size plays a crucial role in MWE identification, and if the distribution of words can be described by Zipf's law, this is even more extreme for MWEs (Evert and Krenn, 2005), which are prone to suffer even more acutely from data sparseness. Therefore, for a reliable MWE identification, where true candidates are at the top of the rank, large corpora need to be used. For instance, for idioms Geyken et al. (2004) found that for a sample of 46 idioms a corpus of 100 million tokens was too small, and interesting results required 800 million words. Controlled well-balanced corpora of that magnitude are still rare for many languages, and one alternative is to use the Web as a corpus. To test the impact of doing that, Villavicencio et al. (2007) looked at the influence of size and quality of different corpora like BNC and Yahoo for MWE identification, and found that in terms of language usage, web generated corpora are fairly similar to more carefully built corpora. Moreover, they found a higher agreement between Web corpora (Google and Yahoo) than between the complete BNC and one of its subsets for MWE identification using AMs like χ^2 and pointwise mutual information, suggesting that larger sizes compensated for any possible noise in the data. Besides corpus size other factors that also seem to have an impact on MWE identification are the particular language and MWE type under consideration (Evert and Krenn, 2005).

This work follows Evert and Krenn (2005) and Villavicencio et al. (2007) looking at the impact of corpus size and origin on MWE identification, focusing on nominal compounds in Brazilian Portuguese comparing the web corpus with a traditional one. In particular we investigate the effects of using a web corpus generated following the WaCky method (Baroni et al., 2009).

3. The WaCky approach

WaCky corpora are constructed using the following steps defined by Baroni et al. (2009):

1. **seed URL collection**: to ensure content variety, bigrams generated by randomly selecting content words of medium frequency from the target language were submitted to a search engine. For English 2000 bigrams were used (e.g. *iraq package, soil occurs, elsewhere limit*), for German 1653 and for Italian 1000. For each bigram, a maximum of ten seed URLs were

retrieved. From these, duplicates were removed and the remaining URLs were randomly fed to a crawler restricted to pages in the relevant language.

2. **post-crawl cleaning**: to maintain only pages of medium size and remove duplicates. These are further cleaned to remove code and boilerplate elements and filtered according to function word density (each page with 10 types, 30 tokens and 25% of function words).

3. **near-duplicate detection and removal**: calculating the n-gram overlap in terms of 25 5-grams for each two documents.

4. **annotation** of the corpus with additional information such as part-of-speech tagging and parsing.

4. Construcing the brWaC

For the first step in the WaCky method, we constructed a set of content words to use as seeds for the crawler from a list of word frequencies available for the Brazilian Portuguese corpora in Linguateca[3]. From this list which also contains function words like prepositions, numerals and articles, we removed stopwords[4], and applied both a high and a low frequency thresholds to obtain words with medium frequency, removing those with more than 10,000 or less than 100 occurrences. From this set of words, we built 1000 random pairs to submit as queries to a search engine (Bing API). Following Baroni et al. (2009) we use bigram queries because single word queries may return uninteresting pages like company pages or definitions. Using the search engine we generate a set of URLs and the crawler goes through the links in each page, storing each of the pages visited. However, these contain components beyond the human produced text that we want, such as HTML code and boilerplate.

The second step, the post-crawl cleaning, included boilerplate stripping applying a technique based on the boilerpipe library (Kohlschütter et al., 2010), which uses features like the link density and number of words. We also experimented with additional features, and the one that produced the best results was the density of stopwords. According to Pomikálek (2011) content texts will probably contain at least 25% of stopwords, so we removed every text block that did not satisfy this property.

Step 3, duplicate removal, is crucial since otherwise, the amount of data collected may not reflect content variation. This step consisted of getting 20 5-grams sample of words from each text, and comparing them to the other texts. If there were more than 2 identical 5-grams, we assumed that the texts were duplicates. This duplicate filtering approach was adapted from the one proposed by Broder et al. (1997).

For the last step, corpus annotation, we tokenized, lemmatized and POS tagged the corpus, using the TreeTagger (Schmid, 1994) trained for Portuguese[5], which is one of the fastest Taggers available[6], an important characteristic given the large quantities of texts that need to be processed.

Currently, brWaC is under construction, and the generation of a 52M word subset using a multithreaded crawler took approximately six hours including crawling and post-processing of the retrieved pages, including boilerplate stripping. Additionally, duplicate page removal involved the cost of reading the texts, removing stopwords and generating n-grams, which is linear, but demanding[7]. Comparing all the n-grams from different texts to one another (with a quadratic complexity) to identify the intersection between then, results in a mean cost of $20 * n^2$ ($O(n^2)$ complexity). Details about the subset are given in Table 1, with the number of types and tokens in the subset corpus. For comparative purposes, we also include information about CETEN-Folha, a Brazilian Portuguese corpus, extracted from the newspaper Folha de São Paulo[8].

Table 1: Corpora types and tokens

Corpus	Tokens	Types	MWE candidates
brWaC	52M	875K	12,000
CETENFolha	24M	343K	4,024

5. MWE Identification

MWE identification is performed in 3 stages: corpus pre-processing, candidate generation with n-gram and POS patterns, and statistical filtering (Figure 1). Pre-processing of brWaC is as described in section 4., step 4, and the same pipeline is applied to the CETENFolha (tokenization, lemmatization and POS tagging).

Candidates are generated from n-gram and POS patterns, focusing on bigrams and trigrams, using Text-NSP (Banerjee and Pedersen, 2003), which also provides frequency information. The POS patterns for compounds are defined following Justeson and Katz (1995) in terms of Nouns (N), Adjectives (A) and Prepositions (P), table 2. For instance, the pattern Noun Preposition Noun captures expressions with a preposition connecting two nouns, like rede/N sem/P fio/N (wireless network) which are very common in Brazilian Portuguese. We have also included a lexicon filter in this step to exclude non Brazilian Portuguese words. To do this we use a lexicon built by Muniz (2003), and if one of the words on the MWE candidate is not in the lexicon, we discard the candidate.

At the end of this stage we have a list of all n-grams annotated with their POS and their frequencies.

For filtering we apply a low frequency threshold to remove any candidate with less than 50 occurrences in the corpus.

[3] Linguateca Corpora Frequency List, available at dinis2.linguateca.pt/acesso/tokens/formas.totalbr.txt.

[4] Lists of Portuguese Stopwords available at http://www.linguateca.pt/chave/stopwords/.

[5] The TreeTagger supports a number of languages, including Brazilian Portuguese.

[6] According to evaluation in http://mattwilkens.com/2008/11/08/evaluating-pos-taggers-speed/

[7] From empirical observation it takes approximately 10 million operations per text.

[8] Available from http://www.linguateca.pt/cetenfolha/index_info.html

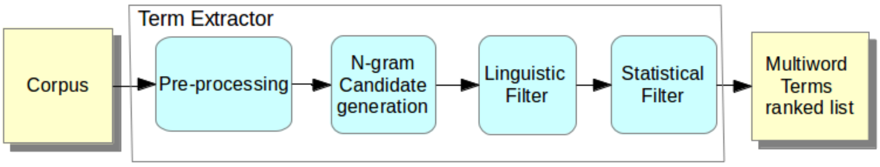

Figure 1: Term Extraction pipeline

Table 2: Compound Patterns

Pattern	example
N N	Nações Unidas (*United Nations*)
N A	governo federal (*federal government*)
N N N	Supremo Tribunal Federal (*Supreme Federal Court*)
N N A	Fundo Monetário Internacional (*International Monetary Fund*)
N A A	produto interno bruto (*gross national product*)
N P N	casa de praia (*beach house*), bolsa de valores (*stock exchange*)

We also use frequency, PMI and c-value, which are commonly used AMs for MWE identification. PMI is calculated as in equation 2 where w_1 and w_2 are the single words that form a MWE candidate and $P(w)$ is the probability of ocurrence of w in the corpus, calculated as in equation 1 where $freq(w)$ is the frequency of w in the corpus and N is the total number of words in the corpus.

$$P(w) = \frac{freq(w)}{N} \tag{1}$$

$$PMI(w_1, w_2) = \frac{P(w_1 w_2)}{P(w_1)P(w_2)} \tag{2}$$

For computing the c-value (Frantzi et al., 2000) we use the frequency of a MWE candidate a and subtract from it the frequencies of the candidates that contain a. It is calculated as in equation 3 where f_a is the frequency of a MWE candidate a, $|a|$ is the number of words of a and T_a is the set of all MWE candidates that contains a.

$$\text{c-value}(a) = \log_2 |a| (f_a - \frac{1}{|T_a|} \sum_{b \in T_a} f_b) \tag{3}$$

6. Results

In this paper we compare MWE identification using a standard corpus and brWaC. A total of 12,000 MWE candidates were extracted from brWaC and 4,024 from CETENFolha (Table 1).

The first aspect we compared was whether MWE candidates had similar distribution profiles in both corpora or if using the Web as a Corpus would have an impact in their distributions. In figure 2 we can see the frequency distribution of the candidates extracted from these corpora. We can see that both corpora generate candidates with similar frequency ranges, following a Zipfian distribution.

The second comparison was whether the use of web data would have an impact on the performance of different AMs. From the AMs tested to rank the MWE candidates, the best results were obtained with c-value for both corpora, closely followed by frequency. As PMI has a bias toward low frequency words, uncommon proper names and other rare combinations were among the top ranked results. Indeed, in relation to the frequency distributions of the first 200 candidates in brWaC ranked with PMI, only 1 of them has a frequency greater than 500 while among those ranked with c-value only 1 candidate has a frequency lower than 500, and the distribution of the top 200 candidates for each of these 2 AMs is in figure 3.

In a manual evaluation performed by specialists of the top 200 candidates ranked by c-value for brWaC and CETENFolha respectively 153 and 137 of them were considered valid MWEs. Figure 4 shows the precision of c-value and frequency for different numbers of evaluated candidates. As expected there is a decrease in precision as n increases for both corpora and measures, but a larger evaluation would be needed for determining how it would evolve. The MWEs from these corpora cover a wide variety of subjects like:

politics and law: governo federal (*federal government*), congresso nacional (*national congress*), Nações Unidas (*United Nations*), direitos autorais (*copyrights*), direitos humanos (*human rights*)

history: idade média (*middle age*), guerra fria (*cold war*)

location names: with countries (Estados Unidos - *United States*), states (*Minas Gerais*) and cities (*Porto Alegre*)

proper names: Adolf Hitler, Albert Einstein, Fernando Henrique Cardoso

events: jogos olímpicos (*olympic games*), copa do mundo (*world cup*)

The rejected MWE's for both corpora include:

common time expressions: ano passado (*last year*), dia seguinte (*next day*)

expression containing verbs - due to PoS tagging errors: ver artigo principal (*see main article*), uso para detalhe (*use for detail*)

incomplete location and proper names: cidade de São (incomplete for *city of São Paulo*), Carlos Alberto (incomplete for *Carlos Alberto Parreira*)

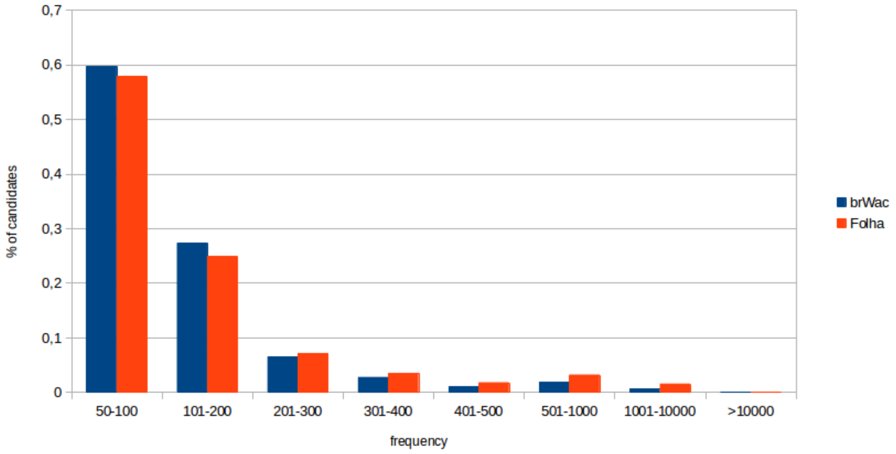

Figure 2: Candidates Frequency Distribution

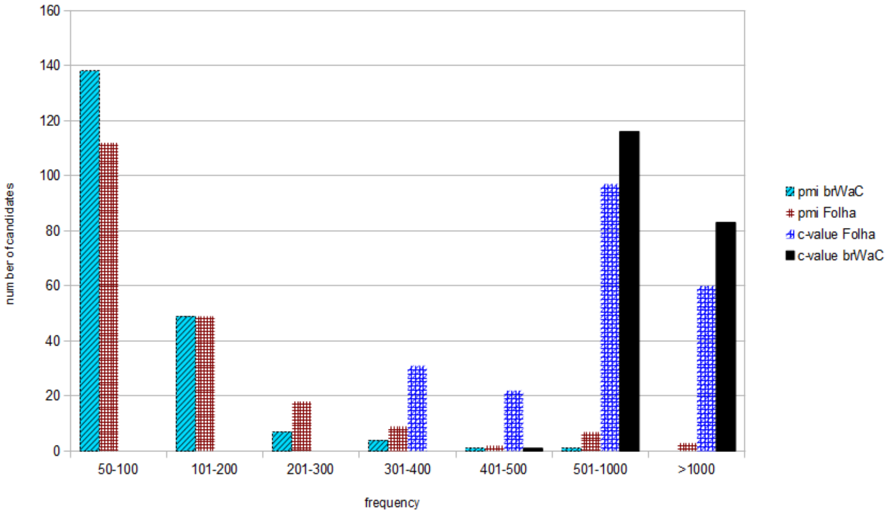

Figure 3: pmi/c-value frequency distribution

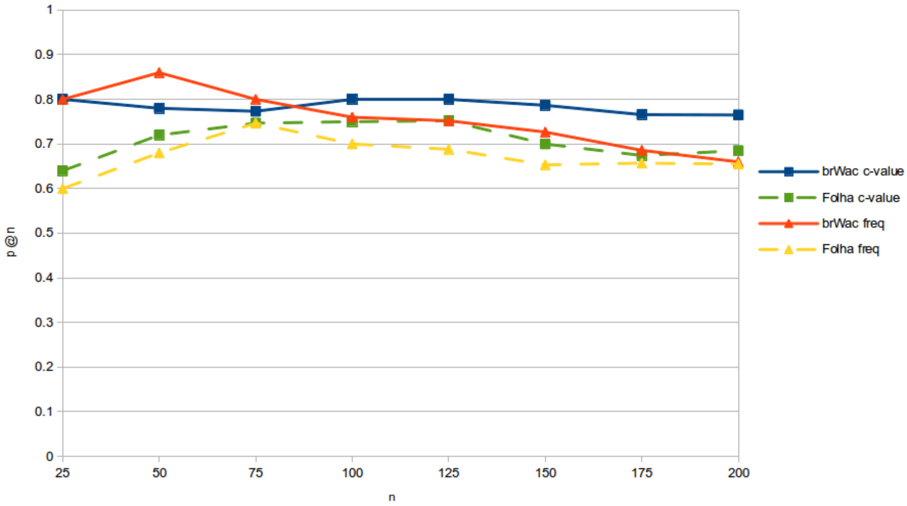

Figure 4: Precision at n

other expressions: norte do rio (*north of the river*), sul do rio (*south of river*), pessoas vivas (*living people*), janela nova (*new window*)

One difference between the MWEs obtained from each corpora is that while CETENFolha is more time-biased containing more names of relevant political figures of the time of publication (1994), brWaC contains more MWEs for general historical events and periods (like segunda guerra mundial - *second world war*) in comparison. Overall, the MWE identification from a Web based corpus suggests that it produces results with comparable quality to standard corpora. Although further manual evaluation of the accuracy of MWE candidates from different parts of the rank is planned for the future, a larger scale automatic evaluation would require gold standard MWE resources, that for a less resourced language like Portuguese are not available, and general lexica may lack MWE coverage[9].

7. Conclusions and Future Work

In this work we discussed the construction of brWaC, applying the approach proposed by (Baroni et al., 2009) for collecting a very large corpus for Brazilian Portuguese. This is an important initiative for a less resourced language like Portuguese, and in a extrinsic evaluation of the quality of the resulting corpus we compared a subset of brWaC with a standard corpus for the identification of MWEs. A manual analysis of the top MWE candidates extracted from these corpora suggests that they generate results with comparable quality. In addition brWaC does not have the intrinsic time bias for current affairs of a newspaper corpus, capturing documents as varied in content and time as the Web. It can also be straightforwardly extended with more data from the dynamically growing Web. Future work includes a manual evaluation of the MWE candidates collected at different stages of completion of the corpus, and a larger scale evaluation with the complete brWaC.

Acknowledgements

This work is partly supported by project CNPq (PRONEM) 003/2011, CNPq 482520/2012-4, 312184/2012-3 and Capes-Cofecub 707/11 .

8. References

Sandra Antunes and Amália Mendes. 2013. Mwe in portuguese: Proposal for a typology for annotation in running text. In *Proceedings of the 9th Workshop on Multiword Expressions*, pages 87–92, Atlanta, Georgia, USA, June.

Timothy Baldwin. 2005. Deep lexical acquisition of verb-particle constructions. In *Comp. Speech & Lang. Special issue on MWEs* (Villavicencio et al., 2005), pages 398–414.

S. Banerjee and T. Pedersen. 2003. The design, implementation, and use of the Ngram Statistic Package. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, pages 370–381, Mexico City, February.

Michele Banko and Eric Brill. 2001. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, ACL '01, pages 26–33. Association for Computational Linguistics.

Luciano Barbosa, Vivek Kumar Rangarajan Sridhar, Mahsa Yarmohammadi, and Srinivas Bangalore. 2012. Harvesting parallel text in multiple languages with limited supervision. In Martin Kay and Christian Boitet, editors, *COLING*, pages 201–214. Indian Institute of Technology Bombay.

Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. Pearson Education Ltd, Harlow, Essex, 1st edition.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. In *Proceedings of the Biennial GSCL Conference*, Tübingen, Germany.

Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. 2007. Large language models in machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 858–867.

Andrei Z. Broder, Steven C. Glassman, Mark S. Manasse, and Geoffrey Zweig. 1997. Syntactic clustering of the web. *Comput. Netw. ISDN Syst.*, 29(8-13):1157–1166, Sep.

L. Burnard. 2000. Users reference guide for the British National Corpus. Technical report, Oxford University Computing Services.

N. Calzolari, C. J. Fillmore, R. Grishman, N. Ide, A. Lenci, C. Macleod, and A. Zampolli. 2002. Towards best practice for multiword expressions in computational lexicons. In *LREC*. European Language Resources Association.

Béatrice Daille. 2012. Building bilingual terminologies from comparable corpora: The TTC TermSuite. In *Proceedings of the 5th Workshop on Building and Using Comparable Corpora with special topic "Language Resources for Machine Translation in Less-Resourced Languages and Domains", co-located with LREC 2012*, Istanbul, Turkey.

Helena Medeiros de Caseli, Carlos Ramisch, Maria das Graças Volpe Nunes, and Aline Villavicencio. 2010. Alignment-based extraction of multiword expressions. In *Lang. Res. & Eval. Special Issue on Multiword expression: hard going or plain sailing* (Rayson et al., 2010), pages 59–77.

Stefan Evert and Brigitte Krenn. 2005. Using small random samples for the manual evaluation of statistical as-

[9] For instance, the NILC lexicon containing 1,649,768 fully inflected entries has only 226 candidates from brWaC and 139 from CETENFolha.

sociation measures. In *Comp. Speech & Lang. Special issue on MWEs* (Villavicencio et al., 2005), pages 450–466.

A. Ferraresi, E. Zanchetta, M. Baroni, and S. Bernardini. 2008. Introducing and evaluating ukwac, a very large web-derived corpus of english. In *In Proceedings of the 4th Web as Corpus Workshop (WAC-4*.

Katerina Frantzi, Sophia Ananiadou, and Hideki Mima. 2000. Automatic recognition of multiword terms: the C-value/NC-value method. *Int. J. on Digital Libraries*, 3(2):115–130.

A. Geyken, A. Sokirko, I. Rehbein, and C. Fellbaum. 2004. What is the optimal corpus size for the study of idioms? In *Proceedings of the Annual meeting of the German Linguistic Society*, Mainz, Germany.

Roger Granada, Lucelene Lopes, Carlos Ramisch, Cassia Trojahn, Renata Vieira, and Aline Villavicencio. 2012. A comparable corpus based on aligned multilingual ontologies. In *Proceedings of the First Workshop on Multilingual Modeling*, MM '12, pages 25–31, Stroudsburg, PA, USA. Association for Computational Linguistics.

Antton Gurrutxaga and Iñaki Alegria. 2013. Combining different features of idiomaticity for the automatic classification of noun+verb expressions in Basque. In *Proceedings of the 9th Workshop on Multiword Expressions*, pages 116–125, Atlanta, Georgia, USA, June.

Ray Jackendoff. 1997. *The Architecture of the Language Faculty*. The MIT Press.

John S. Justeson and Slava M. Katz. 1995. Technical terminology: some linguistic properties and an algorithm for identification in text. *Nat. Lang. Eng.*, 1(1):9–27.

Su Nam Kim and Timothy Baldwin. 2010. How to pick out token instances of english verb-particle constructions. In *Lang. Res. & Eval. Special Issue on Multiword expression: hard going or plain sailing* (Rayson et al., 2010), pages 97–113.

Christian Kohlschütter, Peter Fankhauser, and Wolfgang Nejdl. 2010. Boilerplate detection using shallow text features. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 441–450, New York, NY, USA. ACM.

A. Korhonen, Y. Krymolowski, , and E. J. Briscoe. 2006. A large subcategorization lexicon for natural language processing applications. In *Proceedings of the 5th LREC*, Genova, Italy.

Maria Lapata and Alex Lascarides. 2003. A probabilistic account of logical metonymy. *Comp. Ling.*, 29(2):261–315.

B. Laranjeira, V. P. Moreira, A. Villavicencio, C. Ramisch, and M. J. Finatto. 2014. Comparing the quality of focused crawlers and of the translation resources obtained from them.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 2*, ACL '98, pages 768–774. Association for Computational Linguistics.

Antonio Moreno-Ortiz, Chantal Perez-Hernandez, and Maria Del-Olmo. 2013. Managing multiword expressions in a lexicon-based sentiment analysis system for spanish. In *Proceedings of the 9th Workshop on Multiword Expressions*, pages 1–10, Atlanta, Georgia, USA, June.

Marcelo Caetano Martins Muniz. 2003. Léxicos computacionais: Desafios na construção de um léxico de português brasileiro. Master's thesis, Instituto de Ciências Matemáticas de São Carlos, USP, São Carlos.

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. 2010. Lexical association measures and collocation extraction. In *Lang. Res. & Eval. Special Issue on Multiword expression: hard going or plain sailing* (Rayson et al., 2010), pages 137–158.

Jan Pomikálek. 2011. *Removing Boilerplate and Duplicate Content from Web Corpora [online]*. Thse de doctorat en informatique, Masarykova univerzita, Fakulta informatiky.

Carlos Ramisch, Aline Villavicencio, Leonardo Moura, and Marco Idiart. 2008. Picking them up and figuring them out: Verb-particle constructions, noise and idiomaticity. In Alex Clark and Kristina Toutanova, editors, *Proc. of the Twelfth CoNLL (CoNLL 2008)*, pages 49–56, Manchester, UK, Aug. Coling 2008 Organizing Committee.

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

Carlos Ramisch, Vitor De Araujo, and Aline Villavicencio. 2012. A Broad Evaluation of Techniques for Automatic Acquisition of Multiword Expressions. In *Proceedings of ACL 2012 Student Research Workshop*, pages 1–6, Jeju Island, Korea, Jul. Association for Computational Linguistics.

Carlos Ramisch. 2012. *A generic and open framework for multiword expressions treatment: from acquisition to applications*. Nouvelle thse, University of Grenoble (France) and Federal University of Rio Grande do Sul (Brazil), Grenoble, France, Sep.

Paul Rayson, Scott Piao, Serge Sharoff, Stefan Evert, and Begoña Villada Moirón. 2010. *Lang. Res. & Eval. Special Issue on Multiword expression: hard going or plain sailing*, 44(1-2), Apr.

Pollet Samvelian and Pegah Faghiri. 2013. Introducing PersPred, a syntactic and semantic database for Persian complex predicates. In *Proceedings of the 9th Workshop on Multiword Expressions*, pages 11–20, Atlanta, Georgia, USA, June.

Magali Sanches Duran, Carolina Evaristo Scarton, Sandra Maria Aluísio, and Carlos Ramisch. 2013. Identifying pronominal verbs: Towards automatic disambiguation of the clitic 'se' in Portuguese. In *Proceedings of the 9th Workshop on Multiword Expressions*, pages 93–100, Atlanta, Georgia, USA, June.

Ryohei Sasano, Daisuke Kawahara, and Sadao Kurohashi. 2009. The effect of corpus size on case frame acquisition for discourse analysis. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 521–529, Stroudsburg, PA, USA. Association for Computational Linguistics.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees.

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

Frank A. Smadja. 1993. Retrieving collocations from text: Xtract. *Comp. Ling.*, 19(1):143–177.

Tuomas Talvensaari. 2008. Effects of aligned corpus quality and size in corpus-based clir. In Craig Macdonald, Iadh Ounis, Vassilis Plachouras, Ian Ruthven, and RyenW. White, editors, *Advances in Information Retrieval*, volume 4956 of *Lecture Notes in Computer Science*, pages 114–125. Springer Berlin Heidelberg.

Yulia Tsvetkov and Shuly Wintner. 2010. Extraction of multi-word expressions from small parallel corpora. In *Coling 2010: Posters*, pages 1256–1264, Beijing, China, August. Coling 2010 Organizing Committee.

Aline Villavicencio, Francis Bond, Anna Korhonen, and Diana McCarthy. 2005. *Comp. Speech & Lang. Special issue on MWEs*, 19(4).

Aline Villavicencio, Valia Kordoni, Yi Zhang, Marco Idiart, and Carlos Ramisch. 2007. Validation and evaluation of automatically acquired multiword expressions for grammar engineering. In Jason Eisner, editor, *Proc. of the 2007 Joint Conference on EMNLP and Computational NLL (EMNLP-CoNLL 2007)*, pages 1034–1043, Prague, Czech Republic, Jun. ACL.

Aline Villavicencio. 2005. The availability of verb-particle constructions in lexical resources: How much is enough? In *Comp. Speech & Lang. Special issue on MWEs* (Villavicencio et al., 2005), pages 415–432.