A New Approach for Idiom Identification Using Meanings and the Web*

Rakesh Verma
Computer Science Dept.
University of Houston
Houston, TX, 77204, USA
rverma@uh.edu

Vasanthi Vuppuluri
Computer Science Dept.
University of Houston
Houston, TX, 77204, USA
vvuppuluri@uh.edu

Abstract

There is a great deal of knowledge available on the Web, which represents a great opportunity for automatic, intelligent text processing and understanding, but the major problems are finding the legitimate sources of information and the fact that search engines provide page statistics not occurrences. This paper presents a new, domain independent, general-purpose idiom identification approach. Our approach combines the knowledge of the Web with the knowledge extracted from dictionaries. This method can overcome the limitations of current techniques that rely on linguistic knowledge or statistics. It can recognize idioms even when the complete sentence is not present, and without the need for domain knowledge. It is currently designed to work with text in English but can be extended to other languages.

1 Introduction

Automatically extracting phrases from the documents, be they structured, un-structured or semistructured has always been an important yet challenging task. The overall goal is to create a easily machine-readable text to process the sentences. In this paper we focus on identifying idioms from text. An idiom is a phrase made up of a sequence of two or more words that has properties that are not predictable from the properties of the individual words or their normal mode of combination. Recognition of idioms is a challenging problem with wide applications. Some examples of idioms are 'yellow journalism,' 'kick the bucket,' and 'quick fix'. For example, the meaning of 'yellow journalism' cannot be derived from the meanings of 'yellow' and 'journalism.'

*Research supported in part by NSF grants CNS 1319212, DUE 1241772 and DGE1433817

Idioms play an important role in Natural Language Processing (NLP). They exist in almost all languages and are hard to extract as there is no algorithm that can precisely outline the structure of an idiom. Idioms are important for natural language generation, parsing, and significantly influence machine translation and semantic tagging. Idioms could be also useful in document indexing, information retrieval, and in text summarization or question-answering approaches that rely on extracting key words or phrases from the document to be summarized, e.g., (Barrera and Verma, 2011; Barrera and Verma, 2012; Barrera et al., 2011). Efficiently extracting idioms significantly improves many areas of NLP. But most of the idiom extraction techniques are biased in a way that they focus on a specific domain or make use of statistical techniques alone, which results in poor performance. The technique in this paper makes use of knowledge from the Web combined with knowledge from dictionaries in deciding if a phrase is a idiom rather than solely depending on frequency measures or following rules of a specific domain. The Web has been attractive to NLP researchers because it can solve the sparsity issue and also its update latency is lower than for dictionaries, but its disadvantages are noise, lack of a good method for finding reliable sources and the coarseness of page statistics. Dictionaries are more reliable but they have higher update latency. Our work tries to minimize the disadvantages and maximize the advantages when combining these resources.

1.1 Contribution

This paper proposes a new idiom identification technique, which is general, domain independent and unsupervised in the sense that it requires no labeled datasets of idioms. The major problem with existing approaches is that most of them are supervised, requiring manually annotated data,

Proceedings of Recent Advances in Natural Language Processing, pages 681–687,
Hissar, Bulgaria, Sep 7–9 2015.

and many of them impose syntactic restrictions, e.g., verb-particle, noun-verb, etc. Our technique makes use of carefully extracted reliable knowledge from the Web and dictionaries. Moreover, our technique can be extended to languages other than English, provided similar resources are available. Although our approach uses meanings, with the advancement of the web, more and more phrase definitions are becoming available on the web and thus the reliance on dictionaries can be reduced or even eliminated. However, in many cases, even though the definition of a phrase may be available, the phrase itself is not necessarily labeled as an idiom so we cannot just do a simple lookup of a phrase and mark it as an idiom.

The rest of the paper is organized as follows. Section 2 presents previous work on idiom extraction and classification. In Section 3 we present our approach in detail. Section 4 presents the datasets and in Section 5 we present the experiments and comparisons. We conclude in Section 6.

2 Related Work

There is considerable work on extracting multiword expressions (MWEs), a superclass of idioms, e.g., (Zhang et al., 2006); (Villavicencio et al., 2007); (Li et al., 2008); (Spence et al., 2013); (Ramisch, 2014); (Marie and Constant., 2014); (Schneider et al., 2014); (Kordoni and Simova, 2014); (Yulia and Wintner, 2014). We do not cover this work here since our focus is on idioms.

Because of its importance, several researchers have investigated idiom identification. As mentioned in (Muzny and Zettlemoyer, 2013), prior work on this topic can be categorized into two streams: *phrase classification* in which a phrase is always idiomatic or literal, e.g., (Gedigian et al., 2006); (Shutova et al., 2010), or *token classification* in which each occurrence of a phrase is classified as either idiomatic or literal, e.g., (Birke et al., 2006); (Katz and Eugenie, 2006); (Li and Sporleder, 2009); (Fabienne et al., 2010); (Caroline et al., 2010); (Peng et al., 2014). Most work on the phrase classification stream imposes syntactic restrictions. Verb/Noun restriction is imposed in (Fazly et al., 2009) and (Diab and Pravin, 2009); subject/verb and verb/direct-object restrictions are imposed in (Shutova et al., 2010) and verb-particle restriction is imposed in (Ramisch et al., 2008). Portions of the American National Corpus were tagged for idioms composed of verb-noun constructions, prepositional phrases, and subordinate clauses in (Laura et al., 2010).

To our knowledge, there are only a few general approaches for idiom identification in the phrase classification stream (Muzny and Zettlemoyer, 2013); (Feldman and Peng, 2013) and most of the techniques are supervised. A supervised technique for automatically identifying idiomatic dictionary entries with the help of online resources like Wiktionary is discussed in (Muzny and Zettlemoyer, 2013). There are three lexical features and five graph-based features in this technique, which model whether phrase meanings are constructed compositionally. The dataset consists of phrases, definitions, and example sentences from the English-language Wiktionary dump from November 13th, 2012. The lexical and graph-based features when used together yield F-scores of 40.1% and 62.0% when tested on the same dataset, once without annotating the idiom labels and once after providing the annotated labels. This approach when combined with the Lesk word sense disambiguation algorithm and a Wiktionary label default rule, yields an F-score of 83.8%.

An unsupervised idiom extraction technique using Principal Component Analysis (PCA) treating idioms as semantic outliers and a supervised technique based on Linear Discriminant Analysis (LDA) was described by (Feldman and Peng, 2013). The idea of treating idioms as outliers was tested on 99 sentences extracted from the British National Corpus (BNC) social science (non-fiction) section, containing 12 idioms, 22 dead metaphors and 2 living metaphors. The idea of idiom detection based on LDA was tested on 2,984 Verb-Noun Combination (VNC) tokens extracted from BNC described in (Fazly et al., 2009). These 2,984 tokens are translated into 2,550 sentences of which 2,013 are idiomatic sentences and 537 are literal sentences. A variety of results were presented for PCA for different false positive rates ranging from 1 to 10% (one Table with rates of 16-20%). For idioms only, the detection rates range from 44% at 1% false positive rate to 89% at 10% false positive rate.

Some of the work in the token classification stream, e.g., (Peng et al., 2014), relies on a list of potentially idiomatic expressions. Such a list can be generated using our technique.

3 Idiom Extraction Model

We now present the details of our approach for extracting idioms, which is implemented in Python and called IdiomExtractor. We focus on the meaning of the word *idiom*, i.e., "*properties of individual words in a phrase differ from the properties of the phrase in itself.*" Hence, we look at what individual words in a phrase mean and what the phrase means as a whole. If the meaning of phrase is different from what the individual words in the phrase try to convey then by definition of the word *idiom*, that phrase is a idiom.

Steps involved in the process of idiom extraction are as follows:

3.1 Definition Extraction

This step is the most important step in determining if a phrase is a idiom. The definitions of the phrase (D_p) and individual words as per the Part-of-Speech (POS) *whenever possible*, in the phrase are obtained, $\{D_{W1}, D_{W2}, ..., D_{Wj}\}$. In some case a dictionary may not have definitions for a word for the given POS, in which case definition of the word is obtained without taking POS into consideration. For obtaining definitions, we use WordNet, WordNik dictionary API and Bing search API. Here,

$D_p = \{D_1, D_2, D_3, ..., D_k\}$
$D_{W1} = \{D_{11}, D_{12}, D_{13}, ..., D_{1n}\}$
$D_{W2} = \{D_{21}, D_{22}, D_{23}, ..., D_{2m}\}$, and so on.

3.2 Recreating Definitions

Once we have the definitions of each word and those of the phrase, each of the definition is POS tagged using the NLTK POS tagger and only the words whose POS tag is from {noun, verb} are considered and the definitions are recreated after stemming the words using the Snowball Stemmer[1] as, RD_p and $\{RD_{W1}, RD_{W2}, ..., RD_{Wn}\}$ with only those words present. This constraint stems from our observations of several idioms, which showed that idioms in general have at least one of the mentioned POS tags in-order for the phrase to have a meaning. Here,

$RD_p = \{RD_1, RD_2, RD_3, ..., RD_k\}$
$RD_{W1} = \{RD_{11}, RD_{12}, RD_{13}, ..., RD_{1n}\}$
$RD_{W2} = \{RD_{21}, RD_{22}, RD_{23}, ..., RD_{2m}\}$, and so on.

Now, each of the word in the original phrase is replaced with its definitions which results in a set of new phrases P as follows:

$P = \{RD_{11}RD_{12}...RD_{j1}, RD_{12}RD_{21}...RD_{j1}, RD_{1n}RD_{2m}...RD_{jl}\}$

To avoid any confusion regarding how the procedure is implemented an example is provided below.

3.3 Subtraction

Each of the phrases present in P is subtracted from each of the recreated definition in RD_p and the result is stored in set S.

3.4 Idiom Result

There are two options the user can choose in deciding if a phrase is a idiom. They are:

– By Union
– By Intersection

By Union: This is a lenient way of deciding if a phrase is a idiom. Here, if at least one word survives the subtraction step above, then that phrase is declared to be a idiom.

By Intersection: This is a stricter way of deciding if a phrase is a idiom. Here, a phrase is a idiom if and only if at least one word survives all of the subtraction operations.

Example - Definition extraction

D_p = Definition of 'forty winks' = {sleeping for a short period of time (usually not in bed)}
D_{W1} = Definitions of 'forty' as a 'Noun' = {the cardinal number that is the product of ten and four}
D_{W2} = Definitions of 'winks' as a 'Noun' = {a very short time (as the time it takes the eye to blink or the heart to beat), closing one eye quickly as a signal, a reflex that closes and opens the eyes rapidly}

Example - Recreating definitions

RD_p = {sleep period time bed}
RD_{W1} = {number product ten}
RD_{W2} = {time time eye blink heart beat, eye signal, reflex}
P = {number product ten time time eye blink heart beat, number product ten eye signal, number product ten reflex}. Note that we do not eliminate duplicate words such as the word "time" in RD_{W2}, since they really do not affect the idiom extraction,

[1] http://snowball.tartarus.org/download.php

procedure IDIOM EXTRACTION
  for $phrase$ p in $phrases$ extracted **do**
    D_p = Definition of phrase p
    RD_p = Recreated definitions of phrase p
    for _word in phrase **do**
      D_{wi} = Definition of the _word
      RD_{wi} = Recreated definitions of the _word
    Recreating definition phrases, P
    $P = \{RD_{11}RD_{12}...RD_{j1}, RD_{12}RD_{21}...RD_{j1}, RD_{1n}RD_{2m}...RD_{jl}\}$
    Subtraction. $S = RD_p - P$
    idiom result: by Union.
    if S is non-empty **then**
      phrase p is an idiom
    idiom result: by Intersection
    if at least one word survives all subtractions **then**
      phrase p is an idiom

Figure 1: Idiom Extraction Algorithm

but future versions of the software will optimize this aspect.

Example - Subtraction

S = {sleep period time bed} - {number product ten time time eye blink heart beat, number product ten eye signal, number product ten reflex}
= {sleep period bed, sleep period time bed, sleep period time bed}

Count of each word that after subtraction = {sleep: 3, period: 3, time: 2, bed: 3}

The idiom extraction steps can easily be understood with an example as follows:

Example - idiom Result
By Union: Since S is a non-empty set, the phrase 'forty winks' is a idiom
By Intersection: At least one word in S is present as many times as those of recreated definitions. Hence 'forty winks' is a idiom.

4 Datasets

For the experiments in this paper, we used different datasets extracted from englishclub.com and Oxford Dictionary of Idioms and VNC corpus. The datasets and their extraction process is explained here.

4.1 Idiom Example Sentences Dataset

Dataset-1: An idiom dataset is obtained from englishclub.com[2]. From the website, 198 idioms are randomly chosen and 198 example sentences that exemplify those 198 idioms are used. These 198 example sentences that are manually extracted serve as our dataset. This dataset facilitates the evaluation of false positive rate of our technique.

4.2 Oxford Dictionary of Idioms Dataset

Dataset-2: This dataset is a collection of idioms obtained from the Oxford Dictionary of idioms. The text file consisting of 176 idioms is the input for IdiomExtractor. This dataset facilitates the evaluation of recall and false negative rate of our approach.
Preprocessing and Sanitization:
PDFMiner[3] was used to extract text as XML from the PDF version of Oxford Dictionary of Idioms and then a Python script was used to extract idioms from the .xml file into a text file. Also, any non-ASCII characters are ignored while writing the idioms to the text file.

4.3 VNC Dataset

Dataset-3: VNC-tokens are obtained from (Fazly et al., 2009). This dataset consists of 53 unique

[2] https://www.englishclub.com/ref/Idioms/ (02/23/2015)
[3] http://www.unixuser.org/~euske/python/pdfminer/ (11/28/2014)

(%)	IdiomExtractor (Union)	IdiomExtractor (Intersection)	AMALGr	Expected maximum
Recall	82.30	67.17	31.50	100.00
Precision	65.90	95.50	14.82	100.00
F-score	73.25	78.69	20.16	100.00

Table 1: Idiom extraction: IdiomExtractor Vs. AMALGr on Dataset-1

(%)	IdiomExtractor (Union)	IdiomExtractor (Intersection)	AMALGr	Expected maximum
Recall	100.00	90.90	67.61	100.00
Precision	100.00	100.00	67.23	100.00
F-score	100.00	95.23	67.42	100.00

Table 2: Idiom extraction: IdiomExtractor Vs. AMALGr on Dataset-2

tokens which were tagged as idiomatic or literal. Irrespective of what the tag was we considered all the tokens as input for our software. We evaluate the recall and false negative rate of our software with the help of this dataset.

5 Performance Evaluation

5.1 IdiomExtractor's Performance

Depending on the number of idioms whose definitions were obtained, the maximum possible recall, precision and F-score are calculated for each of three datasets and the values are tabulated under the 'Expected maximum' column.

On Dataset-1: IdiomExtractor has an F-score of 73.25% by Union approach and 78.69% by Intersection approach. Recall and Precision is documented in Table 3.4. Definitions of all 198 idioms in this dataset are obtained from englishclub.com.

On Dataset-2: IdiomExtractor has an F-score of 95.23% by Intersection approach and 100.00% by Union approach. Recall and Precision is documented in the Table 3.4. For this experiment, we used Oxford Dictionary of Idioms to obtain definitions of 176 idioms.

On Dataset-3: IdiomExtractor has an F-score of of 90.72% by Intersection approach and an F-score of 95.04% by Union approach. In this experiment, we used idiom definitions obtained from two Internet sources[4,5] and individual word definitions are obtained from WordNet dictionary.

[4] `http://idioms.thefreedictionary.com/`
[5] `http://dictionary.reference.com/`

5.2 IdiomExtractor Vs. AMALGr

We compare our idiom extraction module with AMALGr from (Schneider et al., 2014) since their definition of MWE "lexicalized combinations of two or more words that are exceptional enough to be considered as single units in the lexicon" aligns with our definition of a idiom and since the authors kindly made their software available.[6] AMALGr requires SAID[7] corpus to be purchased from Linguistic Data Consortium (LDC) (which we purchased) to train the software along with other training data sets. AMALGr requires input text to be represented as two tab separated tokens per line, with the first token being a word from the input and the second token being the part of speech of the word, followed by an empty line when the sentence ends.

When tested on Dataset-1, F-score of IdiomExtractor is 50% more when compared to the F-score of AMALGr. We believe that IdiomExtractor's performance can further be improved if efficient phrasal dictionaries were available for research purposes. Results are documented in Table 3.4.

Reason for low precision of AMALGr: AMALGr joins individual words of MWEs either with an underscore (strong MWE) or tilde (weak MWE). In certain cases, not all words of all the idioms are joined together with either of the special characters and parts of idioms were tagged as MWEs. For example, 'ugly duckling', 'settle a score' weren't tagged as MWEs. An example where part of an idiom is tagged as a MWE is "punch someones lights out." These are declared

[6] Not everyone we contacted was willing to share idiom extraction software.
[7] `https://catalog.ldc.upenn.edu/LDC2003T10` (02/03/2015)

(%)	IdiomExtractor (Union)	IdiomExtractor (Intersection)	AMALGr	Expected maximum
Recall	90.56	83.01	54.71	90.56
Precision	100.00	100.00	100.00	100.00
F-score	95.04	90.72	70.73	95.04

Table 3: Idiom extraction: IdiomExtractor Vs. AMALGr on Dataset-3

as false positives since we were looking for an exact match for the idiom. This caused a drop in the precision.

When tested on Dataset-2, out of 176 idioms, 119 are tagged as idioms by AMALGr (including both strong and weak idioms as described in (Schneider et al., 2014)) with Recall = 67.61%, Precision = 67.23%, F-score = 67.42%, which, when compared to the performance of IdiomExtractor's Union approach is 32.39% less. Results are documented in Table 3.4.

When tested on Dataset-3, out of 55 VNC-tokens, 29 are tagged as MWEs (strong MWEs and weak MWEs combined). In comparison with IdiomExtractor, the recall from AMALGr is 28.30% less than that of IdiomExtractor, which is 83.01%. IdiomExtractor failed to catch 5 VNC-tokens whose definitions were not provided.

6 Conclusion

In this paper we have presented a new approach for idiom extraction that is both domain and language independent, and does not require labeling of idioms. Our approach is effective as demonstrated on two datasets and in a direct comparison with the supervised approach AMALGr.

One problem with our approach is that the current resources available to us do not contain meanings of all of the idiom phrases. However, we believe that with advancement in technology we would be able to do a much better job of obtaining the phrase definitions in the near future.

One direction for future work is to compare with the set {noun, verb, adjective, adverb} when recreating definitions.

References

Araly Barrera and Rakesh Verma, *Combining Syntax and Semantics for Automatic Extractive Single-document Summarization*, ACM SAC, Document Engineering Track, 2011, Taiwan.

Araly Barrera and Rakesh Verma, *Combining Syntax and Semantics for Automatic Extractive Single-document Summarization*, CICLING, LNCS 7182, 366-377, 2012, New Delhi, India.

Araly Barrera, Rakesh Verma and Ryan Vincent, *SemQuest: University of Houston's Semantics-based Question Answering System*, Text Analysis Conference, 2011.

Julia Birke and Anoop Sarkar. *A Clustering Approach for Nearly Unsupervised Recognition of Nonliteral Language.* EACL 2006, 11st Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, April 3-7, 2006, Trento, Italy.

Marie Candito and Matthieu Constant. *Strategies for Contiguous Multiword Expression Analysis and Dependency Parsing.* Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers : 743-753.

Mona T. Diab and Pravin Bhutada. *Verb noun construction MWE token supervised classification.* In Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications, pp. 17-22. Association for Computational Linguistics, 2009.

Afsaneh Fazly, Paul Cook and Suzanne Stevenson. *Unsupervised Type and Token Identification of Idiomatic Expressions.* Computational Linguistics 35, no. 1 (2009): 61-103.

Anna Feldman and Jing Peng. *Automatic detection of idiomatic clauses.* Computational Linguistics and Intelligent Text Processing - 14th International Conference, CICLing 2013, Samos, Greece, March 24-30, 2013, Proceedings, Part I.

Fabienne Fritzinger, Marion Weller and Ulrich Heid. *A Survey of Idiomatic Preposition-Noun-Verb Triples on Token Level.* Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta.

Matt Gedigian, John Bryant, Srini Narayanan and Branimir Ciric. *Catching metaphors.* In Proceedings of the Third Workshop on Scalable Natural Language Understanding, pp. 41-48. Association for Computational Linguistics, 2006.

Spence Green, Marie-Catherine de Marneffe and Christopher D. Manning. *Parsing models for identifying multiword expressions.* Computational Linguistics 39.1 (2013): 195-227.

Graham Katz and Eugenie Giesbrecht. *Automatic identification of non-compositional multi-word expressions using latent semantic analysis.* In Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties, pp. 12-19. Association for Computational Linguistics, 2006.

Valia Kordoni and Iliana Simova. *Multiword Expressions in Machine Translation.* Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014.

Street Laura, Nathan Michalov, Rachel Silverstein, Michael Reynolds, Lurdes Ruela, Felicia Flowers, Angela Talucci, Priscilla Pereira, Gabriella Morgon, Samantha Siegel, Marci Barousse, Antequa Anderson, Tashom Carroll and Anna Feldman. *Like Finding a Needle in a Haystack: Annotating the American National Corpus for Idiomatic Expressions.* Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta.

Linlin Li and Caroline Sporleder. *Classifier combination for contextual idiom detection without labeled data.* Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, 6-7 August 2009, Singapore, A meeting of SIGDAT, a Special Interest Group of the ACL.

Linlin Li and Caroline Sporleder. *Linguistic Cues for Distinguishing Literal and Non-Literal Usages.* COLING 2010, 23rd International Conference on Computational Linguistics, Posters Volume, 23-27 August 2010, Beijing, China.

Ru Li, Lijun Zhong and Jianyong Duan. *Multiword Expression Recognition Using Multiple Sequence Alignment.* ALPIT 2008, Proceedings of The Seventh International Conference on Advanced Language Processing and Web Information Technology, Dalian University of Technology, Liaoning, China, 23-25 July 2008.

Grace Muzny and Luke S. Zettlemoyer *Automatic Idiom Identification in Wiktionary.* Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL.

Jing Peng, Anna Feldman and Ekaterina Vylomova. *Classifying Idiomatic and Literal Expressions Using Topic Models and Intensity of Emotions*. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL.

Carlos Ramisch, Aline Villavicencio, Leonardo Moura and Marco Idiart. *Picking them up and figuring them out: Verb-particle constructions, noise and idiomaticity.* Proceedings of the Twelfth Conference on Computational Natural Language Learning, CoNLL 2008, Manchester, UK, August 16-17, 2008: 49-56.

Carlos Ramisch. *Multiword Expressions Acquisition: A Generic and Open Framework.* Theory and Applications of Natural Language Processing. Springer. 2015.

Nathan Schneider, Emily Danchik, Chris Dyer and Noah A. Smith. *Discriminative lexical semantic segmentation with gaps: running the MWE gamut.* Transactions of the Association for Computational Linguistics 2 (2014): 193-206.

Ekaterina Shutova, Sun Lin and Anna Korhonen. *Metaphor Identification Using Verb and Noun Clustering.* COLING 2010, 23rd International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2010, Beijing, China 1002–1010, 2010.

Caroline Sporleder, Linlin Li, Philip Gorinski and Xaver Koch. *Idioms in Context: The IDIX Corpus* Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta.

Yulia Tsvetkov and Shuly Wintner. *Identification of multiword expressions by combining multiple linguistic information sources.* Computational Linguistics 40, no. 2 (2014): 449-468.

Aline Villavicencio, Valia Kordoni, Yi Zhang, Marco Idiart and Carlos Ramisch. *Validation and Evaluation of Automatically Acquired Multiword Expressions for Grammar Engineering.* EMNLP-CoNLL 2007, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 28-30, 2007, Prague, Czech Republic.

Yi Zhang, Valia Kordoni, Aline Villavicencio and Marco Idiart. *Automated multiword expression prediction for grammar engineering.* In Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties, pp. 36-44. Association for Computational Linguistics, 2006.