Multi-Word Expressions in spoken language: PoliSdict

Daniela Trotta[1], **Teresa Albanese**[1] **Michele Stingo**[2] **Raffaele Guarasci**[3] **Annibale Elia**[1]

[1] Università di Salerno, Salerno, Italy
[2] Network Contacts, Molfetta (BA), Italy
[3] ICAR, Consiglio Nazionale delle Ricerche, Italy

{dtrotta,talbanese}@unisa.it michele.stingo@network-contacts.it raffaele.guarasci@icar.cnr.it elia@unisa.it

Abstract

English. *The term multiword expressions (MWEs) is referred-to a group of words with a unitary meaning, not inferred from that of the words that compose it, both in current use and in technical-specialized languages. In this paper, we describe PoliSdict an Italian electronic dictionary composed of multi-word expressions (MWEs) automatically extracted from a multimodal corpus grounded on political speech language, currently being developed at the "Maurice Gross" Laboratory of the Department of Political Sciences, Social and Communication of the University of Salerno, thanks to a loan from the company Network Contacts. We introduce the methodology of creation and the first results of a systematic analysis which considered terminological labels, frequency labels, recurring syntactic patterns, further proposing an associated ontology.*

Italiano. *Con il termine polirematica si fa generalmente riferimento ad un gruppo di parole con significato unitario, non desumibile da quello delle parole che lo compongono, sia nell'uso corrente sia in linguaggi tecnico-specialistici. In questo contributo viene presentato PoliSdict un dizionario elettronico in lingua italiana composto da espressioni polirematiche occorrenti nel parlato spontaneo estratte a partire da un corpus multimodale di dominio politico in lingua italiana in corso di ampliamento presso il Laboratorio "Maurice Gross" del Dipartimento di Scienze Politiche, Sociali e della Comunicazione dell'Università degli Studi di Salerno, grazie a un finanziamento della società Network Contacts. Viene presentata la metodologia di creazione ed i primi risultati di un'analisi sistematica che ha considerato etichette terminologiche, marche d'uso e pattern ricorrenti, proponendo infine un'ontologia associata.*

1 Introduction

The term multi-word expressions (MWEs) includes a wide range of constructions such as noun compounds, adverbials, binomials, verb particles constructions, collocations, and idioms (Vietri, 2014). D'Agostino & Elia (1998) consider MWUs part of a continuum in which combinations can vary from a high degree of variability of co-occurrence of words (combinations with free distribution), to the absence of variability of co-occurrence[1]. They identify four different types of combinations of phrases or sentences, namely (i) with a high degree of variability of co-occurrence among words; (ii) with a limited degree of variability of co-occurrence among words; (iii) with no or almost no variability of co-occurrence among words; (iv) with no variability of co-occurrence among words. The essential role played by MWEs in Natural Language Processing (NLP) and linguistic analysis in general has been long recognised, as confirmed by then numerous dedicated workshops and special issues of journals discussing this subject in recent years (CSL, 2005; JLRE, 2009), and this appears more clear if we consider as the detection of MWEs represents a real issue in several NLP tasks such as semantic parsing and machine translation (Fellbaum, 2011). According to Chiari (2012) regarding the Italian language a line of great

[1] Concerning compositionality, the study of Nunberg et al. (1994) is noteworthy. This study undermines the issue of compositionality, as widely emphasized in Vietri (2014).

interest is represented by the works of Annibale Elia and Simonetta Vietri (Elia, D'Agostino et al 1985, Vietri 1986, D'Agostino and Elia 1998, Vietri 2004). Finally the discussion concerning the MWEs in Italian lexicography has been systematized in the GRADIT (De Mauro 1999) which records 132.000 different MWEs, whose collection was coordinated by Annibale Elia at the Department of Communication Sciences of the University of Salerno. This research is part of the larger project BIG 4 M.A.S.S. conducted by the company Network Contacts[2] in collaboration with the Department of Social Politics and Communication, which received funding to develop semantic and syntactic modules of Italian.

2 Related work

In the last twenty years or so MWEs have been an increasingly important concern for NLP. MWEs have been studied for decades in phraseology under the term phraseological unit. But in the early 1990s, MWEs received increasing attention in corpus-based computational linguistics and NLP. Early influential work on MWEs includes Smadja (1993), Dagan and Church (1994), Wu (1997), Daille (1995), Wermter and Chen (1997), McEnery et al. (1997), and Michiels and Dufour (1998). These studies address the automatic treatment of MWEs and their applications in practical NLP and information systems. An important research contribution is the Multiword Expression Project carried out at Stanford University, which began in 2001 to investigate means to encode a variety of MWEs in precision grammars [3] . Other major work has been conducted at Lancaster University, which resulted in a large collection of semantically annotated English, Finnish and Russian MWE dictionary resources for a semantic annotation tool (Rayson et al. 2004; Lo¨fberg et al. 2005; Piao et al. 2005; Mudraya et al. 2006). Since then, many advances have been made, either looking at MWEs in general (Zhang et al., 2006; Villavicencio et al., 2007), or focusing on specific MWE types, such as collocations (Pearce, 2002), phrasal verbs (Baldwin, 2005; Ramisch et al., 2008) or compound nouns (Keller et al., 2002). A popular type-independent alternative to MWE identification is to use statistical AMs (Evert and Krenn, 2005; Zhang et al., 2006; Villavicencio et al., 2007). Concerned MWE identification and extraction from monolingual corpora, Kim and Baldwin (2006) proposed a method for automatically identifying English verb particle constructions (VPCs), Pecina (2009) reported an evaluation of a set of lexical association measures based on the Prague Dependency Treebank and the Czech National Corpus, Strik et al. (2010) investigated the possible ways of automatically identifying Dutch MWEs in speech corpora. Related to lexical representation of MWEs in a lexicon and a syntactic treebank, Gregoire (2010) discusses the design and implementation of a Dutch Electronic Lexicon of Multiword Expressions (DuELME), which contains over 5,000 Dutch multiword expressions. Bejc˘ek and Stranak (2010) describe the annotation of multiword expressions found within the Prague Dependency Treebank. In NLP, MWEs in spoken language have been studied in the field of automatic speech recognition, generally with the aim of establishing to what extent modeling such expressions can help reducing word error rate (Strik and Cucchiarini 1999). So a review of related work about MWEs highlights the lack of electronic dictionaries of Italian MWEs for spoken language, hence the idea of creating an *ad hoc* dictionary starting from a resource of political domain. That being said, it should be specified here that this study represents an initial experiment on a relatively small sample, since a larger balanced corpus would be necessary for a broader coverage. Political discourse offers interesting cues for analysis and experimentation (Frank, 1996; Dixon, 2002; Callander & Wilkie, 2007; Osborne, 2014). In recent years, political speech has earned much attention (Guerini et al., 2008; 2013; Esposito et al., 2015) for purposes, ranging from analysis of communication strategies (Muelle, 1973; Wilson, 1990; Wilson, 2011), persuasive Natural Language Processing, politicians' rhetoric (Stover & Ibroscheva, 2017) and virality of information diffusion (Caliandro & Balina, 2015). Regarding MWs resources for Italian we may mention recent contributions such as PANACEA (Platform for Automatic, Normalized Annotation and Cost-Effective Acquisition of Language Resources for Human Language

[2] Network Contacts, is one of the national leader players in the areas of BPO (business process outsourcing), CRM (customer relationship management), Digital Interaction and Call&Contact Center services. Over the years, it has built numerous partnership with some of the most recognized national academic players, such as the University of Salerno, so as to face stimulating research challenges in the fields of Artificial Intelligence and Natural Language Processing.

[3] For more information cfr. http://mwe.stanford.edu

Techologies) that includes Italian word n-grams and Italian word/tag/lemma n-grams in the "Labour" (LAB) domain (Bel at al., 2012) and also PARSEME-IT Corpus, an annotated Corpus of Verbal Multiword Expressions in Italian (Monti et al., 2017).

3 PoliSdict

According to Gross (1999) the lexicographic data available in machine-readable format are printed dictionaries, electronic dictionaries and corpora. In particular dictionaries are built for being used by programs, with their content made of alphanumerical codes which represent the grammatical data that can be reasonably formalized at this moment in time. The creation and management of the electronic dictionary of MWEs in Italian spoken language took place through four main steps:

- lexical acquisition from corpus
- lexicon-based identification of MWEs
- information extraction
- identification of most recurrent PoS patterns

The first step concerns the *lexical acquisition*. We automatically extract MWEs starting from PoliModalCorpus (Trotta et al., 2018), a political domain corpus for Italian language currently composed of transcriptions[4] of 59 face-to-face interviews (14:00:00 hours) held during the political talk show "In mezz'ora in più" (from 24 September 2017 to 14 January 2018) and 18 speeches (7:02:39 hours) held during the election campaign for regional elections (from December 24th 2014 to March 4th 2015) by the then candidate Vincenzo De Luca[5]. The dimension of the individual corporus is indicated below (Tab. 1).

	Type	**Token**	**TTR**
PoliModalCorpus	11,231	158,543	0.07
De Luca Corpus	7,225	56,672	0.12
Total	18,456	215,251	0.08

Table 1 - Corpus statistics overview

In a second step – exploiting the theoretical backgroung offered by the Lexicon-Grammar[6] framework - we identified the MWEs by processing the corpus in Nooj[7] (Elia et al., 2010) and using the Compound-Word Electronic Dictionaries (DELAC-DELACF) (De Bueriis & Elia, 2008), which includes compound words and sequences formed by two or more words which jointly construct single units of meaning, thanks to which it was also possible to attribute a terminological label to each identified MWEs. It has to be noticed that in this step our efforts focused on the extraction of nominal compounds, leaving the extraction and integration of adverbial and adjectival compounds for future research. In a third phase the extracted MWEs were manually verified using the GRADIT (De Mauro, 2000). This operation has allowed us to identify 356 MWEs compared to 882 identified by DELAC-DELACF and to attribute to each compound expression the respective frequency label documented by the GRADIT. In a fourth phase a structural analysis of the extracted MWEs was carried out and the most recurring part of speech patterns were identified. Therefore the terminological labels[8] are distributed as follows: <econ> 112, <fig> 37, <dige> 36, <pol> 21, <med> 17[9]. Even though we extracted the MWEs from interviews of political kind, the MWEs tagged with the <pol> (political) labels are only 21. Following the most recurrent frequency label we found were: TS[10] (167) (i.e. *abuso di ufficio*), CO[11] (136) (i.e. *arredo urbano*), CO - TS (30) (i.e. *istituto di credito*). The methodological approach of the Lexicon-grammar has also restricted the taxonomic

[4] Using a semi-supervised speech-to-text methodology (Google API + manual transcription).

[5] It should be specified here that our is an initial experiment on a relatively small sample, since a larger balanced corpus would be necessary for a broader coverage.

[6] Gross (1975) shows that every verb has a unique behavior, characterized by different properties and constraints. In general, no ether verb has an identical syntactic paradigm. Consequently, the properties of each verbal construction must be represented in a lexicon-grammar.

[7] NooJ is a knowledge-based NLP tool based on huge hand-crafted linguistic resources, i.e. Dictionaries, derivational grammars. (Vietri, 2014).

[8] Being an essentially terminological dictionary, DELAC-DELACF assigns one or more terminology labels to each single entry, based on the areas of knowledge in which a specific compound has been attested. Currently the domains are 173 and the most populated is that of medicine.

[9] The terminological labels with a frequency lower than 17 are not mentioned.

[10] Technical-specialist use (107,194 words have this acronym and are known above all in relation to specific contexts of science or technology, eg amicina).

[11] Common use (as many as 47.060 words are used and understood and understood, regardless of profession or origin, to anyone with a higher level of education, eg allusivo).

analysis of compound polysematic words today they are naturally combined with the notion of compound nouns set by Gross and which can be described as "the sequence of their grammatical categories, in the same way as for adverbs" (Gross, 1986). Starting from this point of view, we may indicate how the most recurring patterns in our dictionary were respectively: *N + A* - valid for 218 words (like *lavori forzati* ecc), *N di N* (82) (i.e. *economia di scala*), *N + N* (30) (i.e. *estratto conto*), *N prep N* (22) (i.e. *ministero del lavoro*), *N a N* (2) (i.e. *corpo a corpo*), *N da N* (2) (i.e. *macchina da guerra*). Notice that, since in this study we are dealing with nominal MWEs the syntactic head of the compounds is always represented by the name in patterns like *N + A* and *A + N*, *N + N*, while in more complex patterns, as *N a N* and the like, we found controversial the identification of a single word as syntactic head. Since our primary interest was to identify and systematically arrange the extracted knowledge from a lexicographic point of view, we decided to deepen the syntactic analysis (which is to say the explicitation of the syntactic heads and the syntactic category of each MWE) during research steps to be included in near future research. Starting from the information extracted so far we have then created an electronic dictionary where to each MWE are associated information about gender and number, part of speech pattern, frequency labels, and terminological label. The dictionary was created using the XML as markup language following the TEI standard[12] and adding the tags `<mark>` in order to include the frequency tags indicated by the GRADIT and `<label>` to indicate the knowledge domain in which the word is attested, indicated to the DELAC-DELACF dictionaries). The choice of exploiting this markup language is motivated by its extreme generalization and flexibility (Pierazzo, 2005) and in order to represent the MWEs in a common format and to enable linkage (Calzolari et al., 2002). The adopted formalism uses the following tags:

- `<entry>`: contains a single structured entry in any kind of lexical resource, such as a dictionary or lexicon
- `<form>`: (form information group) groups all the information on the written and spoken forms of one headword
- `<gramGrp>`: (grammatical information group) groups morpho-syntactic information about a lexical item, e.g. pos, gen, number
- `<mark>`: frequency label from GRADIT
- `<label>`: terminological label from DELAC-DELACF

The dictionary therefore appears as follows:

```
<entry>
  <form>
    <orth>abuso d'ufficio</orth>
    <type>multiword expression</type>
 </form>
 <gramGrp>
  <gram type= "pos">NdiN</gram>
   <gram type="gen">m</gram>
   <gram type="num">s</gram>
 </gramGrp>
 <mark>TS</mark>
 <label>dige</label>
</entry>

<entry>
  <form>
    <orth>agente atmosferico</orth>
    <type>multiword expression</type>
 </form>
 <gramGrp>
  <gram type= "pos">NA</gram>
   <gram type="gen">m</gram>
   <gram type="num">s</gram>
 </gramGrp>
 <mark>TS</mark>
 <label>meteor</label>
</entry>
```

4 Ontologic expansion of the xml dictionary

Following the creation of the dictionary we also decided to organize the knowledge retrieved from the exploited datasets as an ontological dictionary which is actually under construction and that will be freely avilable under Creative Commons License (CC+BY-NC-ND). The choice to build such a linguistic resource is grounded on the idea that a formal representation of the MWEs may not only help software agents in the automatic recognition of compound words within written/oral texts, but can still enhance the resolution of referential expression such as *Primo Ministro*, *Santo Padre* and the like, which is to say of those frozen expressions that bear pragmatic references pointing to subject/object

[12] P5: Guidelines for Electronic Text Encoding and Interchange, Version 3.4.0. Last updated on 23rd July 2018, revision 1fa0b54.

that are likely to change over medium/short periods of time. In order to perform a deeper pragmatic disambiguation of MWEs we exploited the descriptive capability of the Ontology Web Language (OWL), a standard markup language provided by the World Wide Web (W3C) Consortium for the formalization of vocabularies of terms covering specific domains of knowledge. Following the W3C guidelines we shaped the electronic dictionary so that to each MWE a set of description classes and linking relationship are attached, according to the lexicon-grammar analysis previously performed and transposed into the ontology. Here is an example of the metadata scheme provided for the compound expression *campagna elettorale:*

- **Class "DELAC-DELACF Label"**: <pol> (politic)
- **Class "GRADIT" Label**: CO (Common)
- **Class "Syntactic Pattern"**: N(oun) + A(djective)
- **Data property "Corpus frequency"**: 52
- **Data property "Occurrence"**: *Berlusconi comincia la sua* ***campagna elettorale*** *andando in Tunisia a commemorare Craxi, che ne pensa di questa decisione?*
- **Data property "DBpedia redirection link"**: http://it.dbpedia.org/resource/Campagna_elettorale/html

As we can notice the first three classes plus the first two data properties directly derive from the linguistic analysis and their ontological formalisation may serve as powerful search filters in case of description logic queries submitted over the electronic dictionary. To what concerns the DBpedia redirection link property class, this derives from the Italian section of DBpedia project (Auer *et al.*, 2007) and will serve as core mechanism for the pragmatic resolution of the compound expression. It should be further noticed that the mapping effort between the extracted MWEs and DBpedia virtually put the work in progress ontology on the fifth and last level of Berner Lee's Open Data scale, which is to say on the level reserved for web semantic compliant resources additionally providing redirection links to other web datasets for the contextualisation of the described knowledge, following the initial proposal of (Bizer *et al.*, 2008).

5 Future work

In this work we described the initial steps for the development and formalization of PoliSdict, an electronic dictionary of spoken language MWEs. We illustrated the methdology used to build the resource and the preliminary results that we obtained from a systematic analysis. For what is related to future research we consider necessary exploiting standard association measures (like mutual information or log-likelihood ratio) to get an index of cohesion within the identified expressions and compare the use and collocations of MWEs between corpora of written and spoken language in order to understand which of them are the most used. Considering this study as an initial experiment on a relatively small sample, a larger balanced corpus would be necessary for a broader coverage, therefore we intend to proceed with the expansion of the corpus and the associated dictionary. Following we will make the described resources freely accessible by means of graphical interface, so as to offer the possibility to browse and explore data, also allowing the free use of the source codes for research purposes under Creative Commons License (CC+BY-NC-ND.

6 Aknowledgments

We would like to thank Network Contacts s.r.l. for their willingess to help us with valuable research insights and for the support during the writing of this paper. We would also like to thank the anonymous reviewers for their helpful suggestions.

References

Baldwin, T., & Villavicencio, A. (2002, August). Extracting the unextractable: A case study on verb-particles. In *proceedings of the 6th conference on Natural language learning-Volume 20* (pp. 1-7). Association for Computational Linguistics.

Bejček, E., & Straňák, P. (2010). Annotation of multiword expressions in the Prague dependency treebank. *Language Resources and Evaluation*, *44*(1-2), 7-21.

Bel, N., Poch, M., & Toral, A. (2012). *PANACEA (Platform for Automatic, Normalised Annotation and Cost-Effective Acquisition of Language Resources for Human Language Technologies)*. In

Proceedings of the 16th Annual Conference of the European Association for Machine Translation, Trento, Italy.

Calzolari, N., Fillmore, C. J., Grishman, R., Ide, N., Lenci, A., MacLeod, C., & Zampolli, A. (2002, May). Towards Best Practice for Multiword Expressions in Computational Lexicons. In LREC.

Chiari, I. (2012). Collocazioni e polirematiche nel lessico musicale italiano. *Lingua, letteratura e cultura italiana". Atti del convegno Internazionale*, *50*, 165-190.

CSL. 2005. Special issue on Multiword Expressions of Computer Speech & Language, volume 19.

D'Agostino, E., & Elia, A. (1998). Il significato delle frasi: un continuum dalle frasi semplici alle forme polirematiche. *AA. VV, Ai limiti del linguaggio. Bari: Laterza*, 287-310.

Dagan, I., & Church, K. (1994, October). Termight: Identifying and translating technical terminology. In *Proceedings of the fourth conference on Applied natural language processing* (pp. 34-40). Association for Computational Linguistics.

Daille, B. (1995). Combined approach for terminology extraction: lexical statistics and linguistic filtering.

De Bueriis, G., & Elia, A. (2008). Lessici elettronici e descrizioni lessicali, sintattiche, morfologiche ed ortografiche. *Plectica, Salerno*.

De Mauro, T. (1999). Gradit. *Torino: UTET*, *1*.

Maienborn, C., von Heusinger, K., & Portner, P. (Eds.). (2011). *Semantics: An international handbook of natural language meaning* (Vol. 1). Walter de Gruyter.

Grégoire, N. (2010). DuELME: a Dutch electronic lexicon of multiword expressions. *Language Resources and Evaluation*, *44*(1-2), 23-39.

Gross, G. (2018). Thématisation des compléments circonstanciels. In *Le poids des mots. Hommage à Alicja Kacprzak;*. Wydawnictwo Uniwersytetu Łódzkiego.

Gross, M. (1986, August). Lexicon-grammar: the representation of compound words. In *Proceedings of the 11th conference on Computational linguistics* (pp. 1-6). Association for Computational Linguistics.

Gross, M. (1999). A bootstrap method for constructing local grammars. In *Proceedings of the Symposium on Contemporary Mathematics* (pp. 229-250). University of Belgrad.

JLRE. 2009. Special issue on Multiword Expressions of the Journal of Language Resources and Evaluation, volume to appear.

Kim, S. N., & Baldwin, T. (2006, April). Automatic identification of English verb particle constructions using linguistic features. In *Proceedings of the Third ACL-SIGSEM Workshop on Prepositions* (pp. 65-72). Association for Computational Linguistics.

Kim, S. N., & Baldwin, T. (2006, April). Automatic identification of English verb particle constructions using linguistic features. In *Proceedings of the Third ACL-SIGSEM Workshop on Prepositions* (pp. 65-72). Association for Computational Linguistics.

McEnery, T., Langé, J. M., Oakes, M., & Véronis, J. (1997). The exploitation of multilingual annotated corpora for term extraction. *Corpus annotation---linguistic information from computer text corpora*, 220-230.

Michiels, A., & Dufour, N. (1998). DEFI, a tool for automatic multi-word unit recognition, meaning assignment and translation selection. In *Proceedings of the first international conference on language resources & evaluation* (pp. 1179-1186).

Monti J., di Buono M.P., Sangati, F. (2017) PARSE-ME-It Corpus *An annotated Corpus of Verbal Multiword Expressions in Italian*. In: CLIC-It 2017 Proceedings - Rome 11-13 December 2017.

Mudraya, O., Babych, B., Piao, S., Rayson, P., & Wilson, A. (2006). Developing a Russian semantic tagger for automatic semantic annotation. *Corpus Linguistics 2006*, 290-297.

Nunberg, G., Sag, I. A., & Wasow, T. (1994). Idioms. *Language*, *70*(3), 491-538.

Pearce, D. (2002, May). A Comparative Evaluation of Collocation Extraction Techniques. In *LREC*.

Pecina, P. (2010). Lexical association measures and collocation extraction. *Language resources and evaluation*, *44*(1-2), 137-158.

Piao, S., Archer, D., Mudraya, O., Rayson, P., Garside, R., McEnery, T., & Wilson, A. (2005). A large semantic lexicon for corpus annotation. *Corpus Linguistics 2005*.

Pierazzo, E. (2005). *La codifica dei testi: un'introduzione*. Carocci editore.

Ramisch, C., Schreiner, P., Idiart, M., & Villavicencio, A. (2008, June). An evaluation of methods for the extraction of multiword expressions. In *Proceedings of the LREC Workshop-Towards a Shared Task for Multiword Expressions (MWE 2008)* (pp. 50-53).

Rayson, P., Archer, D., Piao, S., & McEnery, A. M. (2004). The UCREL semantic analysis system.

Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computational linguistics*, *19*(1), 143-177.

Strik, H., & Cucchiarini, C. (1999). Modeling pronunciation variation for ASR: A survey of the literature. *Speech Communication*, *29*(2-4), 225-246.

Strik, H., Hulsbosch, M., & Cucchiarini, C. (2010). Analyzing and identifying multiword expressions in spoken language. *Language resources and evaluation*, *44*(1-2), 41-58.

Trotta, D., Albanese, T., Elia, A., *Polimodalcorpus: verso la costruzione del primo corpus multimodale di dominio politico in italiano;* Proceedings of the XXVIII Ass.I.Term International Conference, Salerno, 2018.

Vietri, S. (2004). *Lessico-grammatica dell'italiano. Metodi, descrizioni e applicazioni* (p. 304). UTET Università.

Vietri, S. (1985). *Lessico e sintassi delle espressioni idiomatiche: una tipologia tassonomica dell'italiano*. Liguori.

Vietri, S. (2014). *Idiomatic constructions in Italian: a lexicon-grammar approach* (Vol. 31). John Benjamins Publishing Company.

Villavicencio, A., Kordoni, V., Zhang, Y., Idiart, M., & Ramisch, C. (2007). Validation and evaluation of automatically acquired multiword expressions for grammar engineering. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

Wermter, S., & Chen, J. (1997). Cautious steps towards hybrid connectionist bilingual phrase alignment. In *Recent Advances in Natural Language Processing* (Vol. 97).

Wu, D. (1997). Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational linguistics*, *23*(3), 377-403.

Zhang, Y., Kordoni, V., Villavicencio, A., & Idiart, M. (2006, July). 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.