Extraction and description of multi-word lexical units in plWordNet 3.0

Agnieszka Dziob
Wrocław University of Technology
Wrocław, Poland
agnieszka.dziob@pwr.edu.pl

Michał Wendelberger
Wrocław University of Technology,
Wrocław, Poland

Abstract

In this paper, we present methods of extraction of multi-word lexical units (MWLUs) from large text corpora and their description in plWordNet 3.0. MWLUs are filtered from collocations of the structural type Noun+Adjective (NA).

1 Introduction

Our focus in this paper are multi-word lexical units (henceforth, MWLUs), derived from collocations (automatically extracted from corpora). As in the case of many linguistic terms, there is no agreement among scholars on their common defining criteria. Two main approaches are distinguished. The first one treats as collocations all expressions that tend to co-occur in the immediate syntactic neighbourhood (Firth 1957). This approach is followed by the constructors of corpora (cf. Przepiórkowski 2012). The second approach puts the emphasis on the linguistic properties of collocations such as non-compositionality and impossibility of modification and substitution (Evert 2004). In this approach the term *collocation* is close to the term *multi-word expression* (henceforth, MWE), used in computational linguistics for the linkage of words of the established meaning, analysed as a whole (Sag et al. 2002) and to our understanding of the term MWLU. In the present paper we define MWLU by reference to *lexical unit* (henceforth, LU), a central element of a wordnet (Fellbaum 1998), a whole attributed with meaning and morphosyntactic properties (Derwojedowa et al. 2008). Thus, MWLU will be an LU, consisting of more than one word and constituting a semantic and morpho-syntactic whole. It is close in spirit to Maziarz et al. 2015 proposal saying that MWLU is *“built from more than one word, associated with a definite meaning somehow stored in one's mental lexicon and immediately retrieved from memory as a whole”* (Maziarz et al. 2015). Such a definition forces one to perceive MWLUs as having defined structure and semantics which makes the connection "behave like the single individual" (Calzolari et al. 2002).

2 Data preparation

In the work on extracting MWEs, IPI PAN Corpus[1] and the plWordNet corpus of the Wrocław University of Technology (Piasecki et al. 2014) corpora were used. The extraction was carried out using the set of MWeXtractor tools, developed for the purposes of the CLARIN[2] project. MWeXtractor is a package of tools, which was created for the purposes of the construction of MWLU's network in plWordNet and their syntactic description. It is the part of a bigger infrastructure for aimed for the work with text corpora. The package user has the access to the data cloud, where they record their own corpora (or uses the existing corpora available on the open licence). MWeXtractor tools package is available on the open licence. Sketch Engine is a tool for the work with corpora, which allows for the extraction of collocations on the basis of their grammatical relations (Kilgarriff et al. 2004). In many respects Sketch Engine and MWeXtractor do not differ from each other. For the purposes of the development of

[1] http://korpus.pl/

[2] http://clarin-pl.eu/

MWeXtractor package new statistical measures were implemented, described in this Section. Those measures, which are compilations or modifications of the known measures, improved extraction results, described in Sections 2 and 3.

In the first phase, the authors defined initial data (sets of corpora, tagset, WCCL's operators describing relations within a collocation (Radziszewski et al. 2011)). In addition, the order of candidates for MWLU can be changed and the continuity of the elements of a collocation does not have to be preserved. The next stage was a dispersion of collocations, through which candidates whose syntactic traits were regarded interesting, are being promoted. In the MWeXtractor package, apart from available measures that are present in the subject literature, the measures designed for the purposes of the present work and presented in Sections 2.1, 2.2, 2.3 were also implemented.

2.1 W Specific Exponential Correlation

The function W Specific Exponential Correlation is a compilation of a few other associative measures, of Specyfic Exponential Correlation among others described above. She is represented by the following pattern:

$$y = p(x,y)\log_2 \frac{p(x,y)^e}{p(x)p(y)}$$

And for her the described generalization is used the pattern:

$$y = p(x_1, x_2, ..., x_n)\log_2 \frac{p(x_1, x_2, ..., x_n)^e}{\prod_{2=1}^{n} p(x_i)}$$

2.2 W Order

W Order is the function based on the assumption, that for them the chic more peculiar to the given connection in which storage connections are appearing, with it more interesting, more certain collocation. The function is disregarding interpretation of the order of the chic, examining only their number and the frequency distribution in chics and from the frequency riots of the collocation for the given candidate, and studying only their attitude.

$$y = \frac{1}{\prod_{i=1}^{n}(1+\frac{f(S(t)i))}{maks(f(S(t)))+1)}}$$

2.3 W Term Frequency Order

This function W Term Frequency Order includes the frequency of appearing of the candidate which many associative measures are using assessed as good.

$$y = \mathrm{f\ (t)WOrder(t)}$$

Two types of files are final data - files with lists k-best of candidates for MWE, and files with evaluations of these lists. The number of generated files in the ranking is equal ((and + V + C) ∗ R ∗ F), where and, V and are indicating C one by one number of exploited functions of associative, vector associative measures and classifiers, however R and F are one by one a number of rounds and folds of cross validation. Additionally for every file with the ranking generated is being Q of files of the evaluation of this ranking, where Q is a number of exploited functions of the evaluation of lists k-bests.
The final list of extracted collocations also contained collocaltions being already Lexical Units in plWordNet. Last filtering consisted in removing proper names and determined descriptions and these LU's.

2.4 Results

Table 1 presents the 20 bests of extracted collocations (of the k-best list). The list included forms of lemma according part of speech:

String of lemma of corpus
N:*link* A:*zewnętrzny* ('external link')
N:*raz* A:*pierwszy* ('first time')
N:*wojna* A:*światowy* ('word war')
N:*to* A:*sam* ('the same')
N:*samorząd* A:*terytorialny* ('local government')
N:*piłka* A:*nożny* ('football')
N:*porządek* A:*dzienny* ('agenda')
N:*papier* A:*wartościowy* ('security')
N:*sprawa* A:*wewnętrzny* ('affairs')
N:*igrzyska* A:*olimpijski* ('Olimpic Games')

N:*strona* A:*drugi* (‘other side’)
N:*podatek* A:*dochodowy* (‘income tax’)
N:*minister* A:*właściwy* (‘minister responsible’)
N:*finanse* A:*publiczny* (‘public finance’)
N:*rada* A:*nadzorczy* (‘supervisory board’)
N:*opieka* A:*zdrowotny* (‘healt care’)
N:*rok* A:*ubiegły* (‘last year’)
N:*ciąg* A:*daleki* (‘string far’)
N:*działalność* A:*gospodarczy* (‘bussines activity’)
N:*projekt* A:*rządowy* (‘government project’)

Table 1: Bests of extracted collocations

3 Syntactically non-compositional MWE’s

Automatic evaluation was the first phase of verification of the extracted collocations. We verified syntactic non-compositionality for NA-type collocations (noun and a postposed Adjective), for which we defined syntactic idiosyncrasies, attesting the stability of the connection (in such a form) in the corpus. Based on a statistical analysis, we argue that MWLUs syntactic non-compositionality must have the following features:

1. established word order
2. separability.

What we understand by the established word order is the ratio of neutral word order (Adjective in postposition) occurrence in the corpus to the alternative word order (Adjective in preposition). We took the established word order as the main criterion, and if its occurrence was lower than 87.09%, the algorithm suggested abandoning further procedure (Maziarz et al. 2015). In the case of reaching more than 87.09 % of occurrence, the algorithm tested separability defined as the ratio of occurrence in the word order with the Adjective in preposition and postposition divided by at least one other text word to the sum of occurrences in both word orders, but without no text word between elements of the collocation.

Finally, by using this method we extracted 607 collocations – potential MWLUs. From this list, we rejected several proper names and incomplete phrases. The rest of collocations was automatically accepted.

Table 2 shows chosen syntactically non-compositional MWLUs.

gra losowa (‘game of chance’)
energetyka odnawialna (‘renewable energy industry’)
klęska żywiołowa (‘natural disaster’)
kodeks celny (‘customs code’)
linie papilarne (‘fingerprint’)
medycyna weterynaryjna (‘veterinary medicine’)
obszar wiejski (‘rural area’)
oficer prasowy (‘Press officer’)
pole golfowe (‘golf course’)
pojemność skokowa (‘engine displacement’)

Table 2: Syntactially non-compositional MWLU’s

4 Verification of extracted collocations

At this stage, we gave linguists the list of extracted collocations for verification. At the preliminary stage of verification, linguists removed (i) combinations which were proper names (and were eliminated during the automatic verification), (ii) combinations with incomplete phrases or (iii) peculiar metaphorical uses (rare in accessible sources). Next, linguists assessed the remaining combinations in accordance to the following criteria:

1. a word cannot appear outside the given collocation (imprisoned meaning),
2. terminology,
3. paraphraseability,
4. free word order (in case of the type NA) (Maziarz et al. 2015a)

By a phrase “a word cannot appear outside the given collocation” we understood the word, for which a given collocation is specific, i.e. the word does not appear in any other collocation in Polish or it does not appear in predicative position. An example of such a collocation is *linia naboczna* (‘lateral line’).

As “terms”, we recognised these collocations, which are precisely and explicitly specified in one or more sources (Polański et al. 1999). In the case of mathematical-natural sciences, technical sciences, law, econometrics or linguistics one source, e.g. encyclopaedia (specialist), the specialist dictionary or the specialist lexicon, was enough for positive verification of the collocation. In the case of other disciplines (especially social sciences or humanities) to do the positive verification two sources of the types listed above were needed. Universal encyclopedias and normative legal texts (acts, regulations) were treated as sufficient sources for term status confirmation of the selected units (Maziarz et al. 2015a). We also took into account other sources (e.g. scientific texts, institutional regulations) whose status is confirmed by some organization (e.g. scientific unit, association). In such cases, to do the positive verification it was essential for the candidate to occur in two sources.

"Paraphraseability" means the possibility of occurrence of a collocation in transformations, in which the collocation becomes separated, or one of its elements is replaced by another word or phrase, without the change in meaning. At this stage the following transformations were allowed:

1. a subordinate clause instead of an Adjective or a participle: *niebieska teczka = teczka, która jest niebieska* ('blue file = file, which is blue');
2. a noun or a prepositional phrase instead of an Adjective (with the force of semantic transposition): *tekst prawny = tekst prawa* ('legal test = text of law'), *drewniana podłoga = podłoga z drewna* ('wooden floor = floor made of wood');
3. a synonym or a dictionary definition in the place of any element of a collocation: gra zespołowa = zabawa towarzyska, która ma określone zasady, może wymagać rekwizytów[3] (team game = team sociable fun, which has particular rules, can need requisites).

In the case of the NA-type, an additional criterion, i.e. word order, was taken into account. On the basis of corpus data, linguists judged whether it was possible to change word order in a collocation without changes in its meaning. In addition, we decided that for the change in word order to be unacceptable, the ratio of NA word order to AN word order has to be greater than 100:1 (Maziarz et al. 2015a).

5 Applications

MWLUs are collected in the MWE dictionary, in which the following description of candidates is applied:

1. MWE's syntactic scheme,
2. MWE's part of speech,
3. MWE's base form,
4. MWE's syntactic head,
5. base form of each MWE's component,
6. part of speech for each MWE's component.

At present, the dictionary contains 45 thousand MWLUs, mainly of nouns and bigrams. MWLU's are grouped together according to syntactic schemes described according to the WCCL formalism (Radziszewski et al. 2011a). The dictionary is systematically enlarged.

Acknowledgements

Work supported by the Polish Ministry of Education and Science, Project CLARIN-PL, the European Innovative Economy Programme project POIG.01.01.02-14-013/09, and by the EU's 7FP under grant agreement No. 316097 [ENGINE].

References

Nicoletta Calzolari, Charles Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catharine MacLeod, & Antonio Zampolli. 2002. *Towards best practice for multiword expressions in computational lexicons*. W: Proceedings of 3rd International Conference on Language Resources and Evaluation (LREC-2002). Las Palmas, Canary Islands - Spain.

Derwojedowa Magdalena, Szpakowicz Stanisław, Zawisławska Magdalena i Piasecki Maciej. 2008. *Lexical units as the centrepiece of a wordnet.* Proceedings of Intelligent Information Systems, Zakopane Poland. Institute of Computer Science PAS.

Stefan Evert. 2004. *The Statistics of Word Cooccurrences Word Pairs and Collocations*, University of Stuttgart.

Christiane Fellbaum (ed.). 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

John Firth. 1957. *The synopsis of linguistic theory 1930-1955. In studies of linguistic analysis*. The Philological Society, Oxford.

Adam Kilgarriff, Pavel Rychly, Pavrl Smrz, David Tugwell. 2004. *The Sketch Engine*. Proceedings of the 11th EURALEX International Congress. France.

[3] Source: plWordNet (`http://plwordnet.pwr.wroc.pl/wordnet/`)

Marek Maziarz, Stan Szpakowicz, Maciej Piasecki. 2015. *A Procedural Definition of Multi-word Lexical Units*. Proceedings of the International Conference on Recent Advances in Natural Language Processing, Hissar, Bulgaria.

Marek Maziarz, Stanisław Szpakowicz, Maciej Piasecki, and Agnieszka Dziob. 2015a. *Jednostki wielowyrazowe. Procedura sprawdzania leksykalności połaczeń wyrazowych* [‘*Multi-word units. A procedure for testing the lexicality of collocations*’]. Technical Report PRE-11, Faculty of Computer Science and Management, Wroclaw University of Technology.

Maciej Piasecki, Marek Maziarz, Stanisław Szpakowicz, and Ewa Rudnicka. 2014. *PlWordNet as the Cornerstone of a Toolkit of Lexico-semantic Resources.* Proc. 7th International Global Wordnet Conference, Tartu, Estonia, 25-29 January.

Krzysztof Polański (ed.). 1999. *Encyklopedia językoznawstwa ogólnego*. [‘*Encyclopedia of general linguistics*’], Ossoliński National Institute, Wroclaw.

Adam Przepiórkowski. 2004. *The IPI PAN Corpus - preliminary version*. Institute of Computer Sciences, PAS, Warsaw.

Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, Barbara Lewandowska-Tomaszczyk (ed.). 2012. *National Corpus of Polish.* Polish Scientific Publishers PWN, Warsaw.

Adam Radziszewski, Adam Wardyński and Tomasz Śniatowski. 2011. WCCL: A Morpho-syntactic Feature Toolkit. Text, Speech and Dialogue. Volume 6836 of Lecture Notes in Computer Science. Springer.

Adam Radziszewski, Michał Marcińczuk, Adam Wardyński. 2011a. *Specyfikacja języka WCCL* [‘*Specification of WCCL language*’]. Faculty of Computer Science and Management, Wroclaw University of Technology. Source: http://nlp.pwr.wroc.pl/redmine/projects/joskipi/wiki/Specyfikacja.

John Sinclair. 1991. *Corpus, Concordance, Collocation*. Oxford University Press, Oxford.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, Dan Flickinger. 2012. *Multiword Expressions: A Pain in the Neck for NLP*. Proceedings of the 3rd Intenational Conference on Computational Linguistics and Intelligent Text Processing. Mexico City.