SEJF -a Grammatical Lexicon of Polish Multi-Word Expressions

Monika Czerepowicka, Agata Savary

SEJF - a Grammatical Lexicon of Polish Multi-Word Expressions

Monika Czerepowicka*, **Agata Savary**†

*University of Warmia and Mazury in Olsztyn, Faculty of Humanities, Poland, czerepowicka@gmail.com
†Université François Rabelais Tours, France, agata.savary@univ-tours.fr

Abstract

We present SEJF, a lexical resource of Polish nominal, adjectival and adverbial multi-word expressions. It consists of an intensional module with about 4,700 multi-word lemmas assigned to 160 inflection graphs, and an extensional module with 88,000 automatically generated inflected forms annotated with grammatical tags. We show the results of its coverage evaluation against an annotated corpus. The resource is freely available under the Creative Commons BY-SA license.

1. Introduction

Multi-word expressions (MWEs) are linguistic objects containing two or more words and showing some degree of non-compositionality. For instance, the meaning of *to kick the bucket* (i.e. to die) cannot be predicted from the meanings of its components, while the singular number of a *cross-roads* is not inherited from the component which should normally be its headword (*roads*). MWEs encompass versatile objects: compounds (*all of a sudden*, *air brake*), complex terms (*random access memory*), multi-word named entities (*European Union*), light-verb constructions (*to take a decision*), idioms (*to kick the bucket*), proverbs (*fortune favors the bold*), etc. Basic facts about MWEs are that they are pervasive in natural language texts, they show idiosyncratic behavior at the level of segmentation, morphology, syntax, semantics or pragmatics, they are concerned by sparseness problems and they are underrepresented in language resources and tools. In morphologically rich, e.g. Slavic, languages MWEs pose additional challenges due to the high number of morpho-syntactic variants under which they occur in texts.

In this paper we focus on Polish compounds. We present SEJF (pl. Słownik Elektroniczny Jednostek Frazeologicznych), a grammatical lexicon of Polish MWEs containing over 4,700 compound nouns, adjectives and adverbs, where inflectional and word-order variation is described via fine-grained graph-based rules. It is provided under two forms – intensional (lemmas and inflection rules) and extensional (list of morphologically annotated variants) – and is available[1] under the terms of the Creative Commons BY-SA license[2].

2. Data Sources

One of the major data sources for the SEJF lexicon was the National Corpus of Polish[3] (NKJP, Narodowy Korpus Języka Polskiego) (Przepiórkowski et al., 2012). The tagsets of both resources are compliant which should facilitate the future use of SEJF in corpus studies.

The NKJP corpus was also used as a source of illustration and verification of research hypotheses. On the basis of concordance lists we verified the forms of the paradigms of almost each MWE included in the lexicon. We also used the corpus to find new, previously undescribed, MWEs thanks to automatic MWE extraction methods developed by the Wrocław University of Technology (Broda et al., 2007). Each of the extracted MWE candidates was manually validated by the lexicographer.

Phraseological units were also acquired from theoretical and lexicographical studies of contemporary Polish. A group of about 1,500 nominal compounds, analyzed by (Kosek, 2008), was the first to be encoded in the dictionary. Some adjectival units were drawn from a dictionary of comparisons (Bańko, 2004). Adverbial units were acquired from two other monographs: (Wojdak, 2008) and (Czerepowicka, 2006).

3. Formalism and Tool

The grammatical description of MWEs in SEJF was done within Toposław (Marciniak et al., 2011), a lexicographic framework offering a user-friendly graphical interface over three core components:

- Morfeusz (Woliński, 2006) – a morphological analyzer and generator of Polish simple words, containing full paradigms of over 250 thousand lemmas.
- Multiflex (Savary, 2009) – a formal language and a tool based on graphs, which describes each inflected form of a MWE as a specific combination of its components. The relation from MWEs to graphs is one-to-many: each MWE (no matter how complex it is) has one particular graph assigned to it, while one graph can describe any number of MWEs.
- A graph editor stemming from Unitex[4], a multilingual corpus processor.

While Morfeusz is Polish-specific, the two other components have also been applied to Serbian, Greek and Macedonian, as mentioned section 7.. Thus, Toposław as a whole is adaptable to another language, provided that a morphological module for simple words in this language exists and that some interface constraints between this module and Multiflex are fulfilled – cf. (Savary, 2009).

[1] http://zil.ipipan.waw.pl/SEJFs

[2] http://creativecommons.org/licenses/by-sa/3.0/

[3] http://clip.ipipan.waw.pl/NationalCorpusOfPolish

[4] http://www-igm.univ-mlv.fr/~unitex/

The description of a MWE in Toposław is a multistage procedure. Firstly, the lexicographer assigns the MWE to the appropriate morphosyntactic class equivalent to one of the 33 *flexemes* (inflectionally motivated POSs) used in the NKJP corpus. Secondly, the MWE is segmented into words and separators, whereas the latter are considered full-fledged components that can further be referred to in inflection graphs. Thirdly, each component word is automatically assigned a list of all lemmas and morphological tags stemming from Morfeusz, thus all possible homonyms are distinguished. The lexicographer manually disambiguates each word by choosing the right interpretation. Fig. 1 shows the nominal MWE *advocat diabła* 'devil's advocate', which has been segmented into three components, including a space. The first component is marked by the lexicographer as admitting inflection. The last one obtains four morphological interpretations, the third of which is correct.

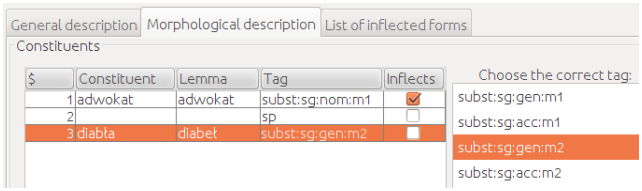

Figure 1: Segmentation and morphosyntactic annotation of the nominal MWE *adwokat diabła* 'devil's advocate' in Toposław. The following codes are used: accusative case (`acc`), genitive case (`gen`), masculine animate gender (`m2`), masculine human gender (`m1`), singular (`sg`), space (`sp`), and substative (`subst`).

In the last step, the lexicographer manually chooses an existing inflection graph (or creates a new one if needed) describing inflected forms of the current MWE entry. Fig. 2 shows the inflection graph `NC-O_N` (cf. tab 2 for the meaning of the `NC`, `O` and `N` codes) for the entry from fig. 1. Graph paths are applied from left to right and the numbered boxes in them correspond to constituents. The formulae inside boxes consist of constituents' indexes and equations on morphological constants and variables. These equations impose constraints on the inflection, variation and agreement of constituents. Here, the formula ⟨*$1:Case=$c;Nb=$n*⟩ says that the first component (here: *adwokat*) inflects freely for case and number. The formulae appearing below paths determine the features of the inflected forms of the whole MWE as a function of the features of its constituents. Here, each form resulting from the unique path inherits its gender from the first constituent and has the conforming case and number (⟨*$1:Gen=$1.Gen;Case=$c;Nb=$n*⟩). Variables like *$c* or *$n* are freely defined by the user and subject to unification, i.e. if they reoccur on the same path the respective components must agree (cf. section 5. and fig. 4).

When applying the graph in fig. 2 to the entry in fig. 1, we automatically obtain the list of all inflected forms and their morphological tags, as shown in fig. 3.

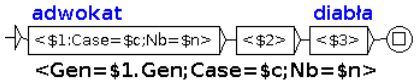

Figure 2: Inflection graph `NC-O_N` for the nominal MWE *adwokat diabła* 'devil's advocate'

adwokat diabła ⊠ adwokat diabła:subst:sg:nom:m1
adwokaci diabła ⊠ adwokat diabła:subst:pl:nom:m1
adwokata diabła ⊠ adwokat diabła:subst:sg:gen:m1
adwokatów diabła ⊠ adwokat diabła:subst:pl:gen:m1
adwokatowi diabła ⊠ adwokat diabła:subst:sg:dat:m1
adwokatom diabła ⊠ adwokat diabła:subst:pl:dat:m1
adwokata diabła ⊠ adwokat diabła:subst:sg:acc:m1
adwokatów diabła ⊠ adwokat diabła:subst:pl:acc:m1
adwokatem diabła ⊠ adwokat diabła:subst:sg:inst:m1
adwokatami diabła ⊠ adwokat diabła:subst:pl:inst:m1
adwokacie diabła ⊠ adwokat diabła:subst:sg:loc:m1
adwokatami diabła ⊠ adwokat diabła:subst:pl:loc:m1
adwokacie diabła ⊠ adwokat diabła:subst:sg:voc:m1
adwokaci diabła ⊠ adwokat diabła:subst:pl:voc:m1

Figure 3: Inflection paradigm for the nominal MWE *adwokat diabła* 'devil's advocate'

4. Contents of the Lexicon

Tab. 1 shows the current state of SEJF. Complete entries are those whose components' inflection is fully handled by Morfeusz and Multiflex, thus the generation of the inflected forms for these entries could be fully performed. Problematic entries are those containing components which are unknown or wrongly handled.

On average, compound nouns have over 12 inflected forms – most of them inflect for case (with 7 case values) and some inflect for number (2 values). Compound adjectives are much more productive, with as many as almost 100 inflected forms on average, due to the case, number and gender inflection (with 9 gender values – 3 masculine, 1 feminine, 2 neuter and 3 plurale tantum ones – according to the Morfeusz tagset). Compound adverbs do not inflect, while among other compounds – selected conjunctions, particles and numerals – only the last ones inflect. The inflection graphs are mostly rather simple: 152 of them contain only one path representing inflection and, possibly, agreement of components. Eight remaining graphs (assigned to 154 MWEs in total) contain two paths, which account mainly for a flexible word order. Tab. 2 shows the most frequently assigned inflection graphs, the corresponding syntactic structures and examples of the assigned entries. A large majority of them consists of a noun and an

	MWU lemmas		Inflected forms	Graphs
	Complete	Problematic		
Nouns	3,705	188	46,021	115
Adjectives	422	33	41,984	30
Adverbs	608	0	608	8
Others	40	1	113	5
ALL	4,775	222	88,726	158

Table 1: Contents of the lexicon

Graph	Syntactic structure	Comment	MWE examples	Assigned MWEs
NC-O_O-1+	S Adj	inflection for number	*koń trojański* 'Trojan horse'	1,153
NC-O_O-1	Adj S	inflection for number	*aksamitna rewolucja* 'velvet revolution'	556
NC-O_O-2t	S Adj	fixed number	*dobra osobiste* 'personal belongings'	426
NC-O_O-1t	Adj S	fixed number	*czarna magia* 'black magic'	396
NC-O_N	S Sgen	inflection for number	*adwokat diabła* 'devil's advocate'	351

Table 2: Distribution of the most frequently assigned inflection graphs. The following codes are used: nominal compound (NC), variable component (O), invariable component (N), substantive (S), substantive in genitive (Sgen), and adjective (Adj).

agreeing adjective in both orders.

5. Interesting Problems

The Toposław suite allows to successfully encode most of the nominal Polish MWEs however not all of them. For instance masculine human gender nouns are challenging in the sense that they exhibit not only the regular case and gender inflection but also have alternative depreciative forms in plural nominative and vocative which take the masculine animate gender m2 (e.g. *adwokaty* instead of *adwokaci* 'advocates'). Because of the unusual gender, these forms constitute a separate flexeme (of type *depr*, cf. the NKJP tagset[5]). Since Toposław does not currently allow to gather several flexemes of the same lemma in one lexeme, generating depreciative forms for masculine human nominal compounds (e.g. *adwokaty diabła* 'devil's advocates') is blocked.

Another reason of a deficient description of the inflection paradigms is the (inevitable) incompleteness of Morfeusz. Neologisms such as *rozporkowy* (relative adjective for a trousers' fly) are not encoded, therefore compounds such as *afera rozporkowa* (lit. 'fly affair'⇒'a sexual scandal') cannot be automatically inflected.

Challenging examples which Toposław allows us to cover include variable word order (*automatyczna sekretarka*, *sekretarka automatyczna* 'answering machine') or fluctuation of the grammatical gender. For instance, the nominal unit *czerwony pająk* (lit. 'red spider'⇒'communist') is exocentric in that its noun component *pająk* 'spider' is in masculine human animate gender (m2), while the whole compound, denoting a person, has the masculine human (m1) behavior. As shown on the upper path in fig. 4, while the case and number of the whole MWE are conforming to the ones of the (inflected) noun and adjective, it's gender is not inherited from component 3 but given by the constant value m1. The major difference in inflection paradigms of masculine human and animate nouns is in the plural accusative form. It is equal to the plural genitive for m1 (*czerwonych pająków*) and to the plural nominative for m2 (*czerwone pająki*). The second path in fig. 4 accounts for the m2-to-m1 shift: the accusative plural masculine human form of the whole compound is obtained by combining the genitive rather than the accusative forms of the two components.

The inflection paradigm generated by the graph in fig.4 for *czerwony pająk* is shown in fig. 5.

[5] http://nkjp.pl/poliqarp/help/ense2.html

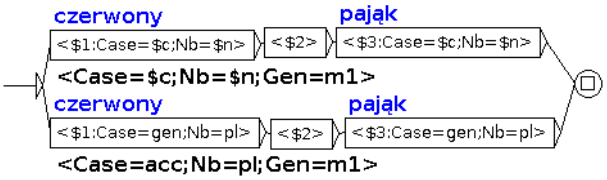

Figure 4: Inflection graph NC-O_N describing a masculine gender fluctuation in *czerwony pająk* 'red spider'⇒'communist'

czerwony pająk ⊠ czerwony pająk:subst:sg:nom:m1
czerwone pająki ⊠ czerwony pająk:subst:pl:nom:m1
czerwonego pająka ⊠ czerwony pająk:subst:sg:gen:m1
czerwonych pająków ⊠ czerwony pająk:subst:pl:gen:m1
czerwonemu pająkowi ⊠ czerwony pająk:subst:sg:dat:m1
czerwonym pająkom ⊠ czerwony pająk:subst:pl:dat:m1
czerwonego pająka ⊠ czerwony pająk:subst:sg:acc:m1
czerwone pająki ⊠ czerwony pająk:subst:pl:acc:m1
czerwonym pająkiem ⊠ czerwony pająk:subst:sg:inst:m1
czerwonymi pająkami ⊠ czerwony pająk:subst:pl:inst:m1
czerwonym pająku ⊠ czerwony pająk:subst:sg:loc:m1
czerwonych pająkach ⊠ czerwony pająk:subst:pl:loc:m1
czerwony pająku ⊠ czerwony pająk:subst:sg:voc:m1
czerwone pająki ⊠ czerwony pająk:subst:pl:voc:m1
czerwonych pająków ⊠ czerwony pająk:subst:pl:acc:m1

Figure 5: Inflection paradigm for the nominal MWE *czerwony pająk* 'red spider'⇒'communist'

6. Evaluation

In order to perform an evaluation of the lexicon we prepared a corpus of general Polish language texts manually annotated with contiguous MWEs. It consists of documents extracted from the manually annotated subcorpus of the National Corpus of Polish. This subcorpus does not contain full texts but only randomly selected paragraphs thereof, and for the sake of our evaluation we chose the 125 longest extracts of different press genres: newspapers, magazines, periodicals, popular science, etc. The annotation schema was rather simple: contiguous sequences of words judged as multi-word expressions of the general Polish language were to be tagged as belonging to one of the following categories: compound noun (CN), foreign compound noun (CNF), compound adjective (CA), foreign compound adjective (CAF), compound adverb (CADV), foreign compound adverb (CADVF) or other MWE (Pol-

Document extracts	Tokens	Annotated MWEs					
		Occurrences					Unique forms
		Nouns	Adjectives	Adverbs	Others	All	
125	234,891	9,468	174	1,087	303	11,032	9,580

Table 3: Contents of the evaluation corpus

	Corpus MWEs found in the lexicon	
	Occurrences	Lemmas
Nouns	598 (6%)	353
Adjectives	7 (4%)	6
Adverbs	364 (33%)	96
All	969 (9%)	455

Table 4: Lexicon coverage evaluated against the corpus

ish, foreign, erroneously spelled – OTH)[6]. The annotator was a native Polish speaker, expert in linguistics, neutral with respect to the project, i.e. uninvolved in the creation of the lexicon. Tab. 3 shows the contents of the resulting evaluation corpus. For the purpose of the evaluation, some categories were merged or eliminated so as to obtain the three final categories to which the lexicon was dedicated: nouns (CN and CNF), adjectives (CA and CAF) and adverbs (CADV and CADVF).

The evaluation results are presented in tab. 4. Note that only about 10% (455 out of 4,775) of all lemmas contained in the lexicon have their inflected forms in the corpus, which confirms the sparseness issues typical for MWEs. The coverage of the evaluation corpus by the lexicon is reasonably high for adverbs (33%) but rather low for nouns and adjectives. The total coverage attains 9%. Two main reasons may underlie this score. Firstly, the lexicon focuses mainly on the most idiomatic, semantically opaque or strongly institutionalized compounds, while the corpus annotator had a much broader understanding of a MWE and marked many relatively weakly lexicalized phrases and collocations (e.g. *wiejska droga* 'country road', *bliski śmierci* 'close to death'). Secondly, the lexicon size was delimited by the scope of the funding project and its development should clearly continue, given that similar resources for other languages easily attain several dozens of thousands of compound lemmas.

7. Related work

Although MWEs are still under-represented in language resources and tools, some efforts have been put towards bridging this gap from the e-lexicographical point of view in several languages. The community around Intex[7], NooJ[8] and Unitex has a long e-lexicographic tradition related to compounds, with dictionaries of compounds created for French (Silberztein, 1993), English (Savary, 2000), Greek (Kyriacopoulou et al., 2002) and others. Lexicons similar to SEJF, following the Multiflex paradigm, exist or are under construction for Serbian (Krstev et al., 2010), Greek (Foufi, 2013), and Macedonian (Rafajlovska and Zdravkova, 2015). Various e-lexicographic frameworks were developed for the creation of MWE e-lexicons notably in Turkish (Oflazer et al., 2004), Basque (Alegria et al., 2004), Dutch (Grégoire, 2010), Serbian (Stanković et al., 2011) and Hebrew (Al-Haj et al., 2014), the third one also covers verbal MWEs.

On the Polish ground, SEJF is one of three grammatical lexicons of Polish multi-word units built under Toposław. The two other resources are: (i) SAWA[9] (Marciniak et al., 2009), a grammatical lexicon of Warsaw urban proper names (streets, squares, bus stops, and other objects linked to the Warsaw communication network), (ii) SEJFEK[10] (Savary et al., 2012), a grammatical lexicon of Polish economic terminology containing over 11,000 specialized nominal compounds. Complementary formalisms for inflectional paradigms of Polish MWUs have been presented in (Graliński et al., 2010) and (Broda et al., 2007).

8. Conclusions and perspectives

We have presented the construction of SEJF, an electronic grammatical lexicon of Polish nominal, adjectival and adverbial MWEs. It is one of the first steps towards a systematic and extensive description of such units, applicable to automatic text processing in Polish. While the coverage of compound adverbs offered by SEJF is reasonable, its contents in terms of compound nouns and adjectives should be extended, as show by the evaluation results. Additional corpora can underlie this further work, including those available via Sketch Engine[11] with collocation support (Radziszewski et al., 2011).

More challenging cases of Polish MWEs, such as nominal and adjectival units with open slot complements, and verbal MWEs, still await a satisfactory description proposal in Polish. One of the steps recently undertaken is the paradigmatic (i.e. relating to constraints of the inflection paradigm of the head verb) description of verbal MWEs within the Verbel project[12]. The inflectional behavior of Polish verbs is very complex, with as many as 12 different flexemes corresponding to one lemma. Since a lexeme has been, so far, the basic description unit in Toposław, using it in its present state to describe verbal MWEs would require 12 lexicon entries for each MWE. This can be overcome if Toposław offers new functionalities allowing us to automatically gather flexemes into lexemes and raise the de-

[6] Some economical sublanguage terms were also annotated but those judged as not belonging to the general Polish language were eliminated during the evaluation.

[7] http://intex.univ-fcomte.fr/

[8] http://www.nooj4nlp.net/pages/nooj.html

[9] http://zil.ipipan.waw.pl/SAWA

[10] http://zil.ipipan.waw.pl/SEJFEK

[11] https://www.sketchengine.co.uk/ske.cgi?page=acd&article=a&language=Polish

[12] http://uwm.edu.pl/verbel

scription onto the level of the latter. This work has already been initiated (Czerepowicka et al., 2014).

9. Acknowledgements

This work has been supported by three projects: (i) Nekst[13], funded by the European Regional Development Fund and the Polish Ministry of Science and Higher Education, (ii) CESAR[14] - a European project (CIP-ICT-PSP-271022), part of META-NET, (iii) IC1207 COST action PARSEME[15].

10. References

Al-Haj, Hassan, Alon Itai, and Shuly Wintner, 2014. Lexical representation of multiword expressions in morphologically-complex languages. *International Journal of Lexicography*, 27(2):130–170.

Alegria, Iñaki, Olatz Ansa, Xabier Artola, Nerea Ezeiza, Koldo Gojenola, and Ruben Urizar, 2004. Representation and Treatment of Multiword Expressions in Basque. In *Proceedings of the ACL'04 Workshop on Multiword Expressions*.

Bańko, Mirosław, 2004. *Słownik porównań*. Warsaw: Polish Scientific Publishers PWN.

Broda, Bartosz, Magdalena Derwojedowa, and Maciej Piasecki, 2007. Recognition of structured collocations in an inflective language. In *Proceedings of the International Multiconference on Computer Science and Information Technology — 2nd International Symposium Advances in Artificial Intelligence and Applications (AAIA'07)*.

Czerepowicka, Monika, 2006. *Opis powierzchniowoskładniowy wyrażeń niestandardowych typu «na lewo», «do dziś», «po trochu», «na zawsze» we współczesnym języku polskim*. Warszawa: Akademicka Oficyna Wydawnicza EXIT.

Czerepowicka, Monika, Iona Kosek, and Sebastian Przybyszewski, 2014. O projekcie elektronicznego słownika odmiany frazeologizmów czasownikowych. *Polonica*, XXXIV:115–123.

Foufi, Vassiliki, 2013. Les noms composés A(A)N du Grec Moderne et leurs variantes. In Fryni Kakoyianni Doa (ed.), *Penser le lexique-grammaire : perspectives actuelles*. Paris, France: Editions Honoré Champion.

Graliński, Filip, Agata Savary, Monika Czerepowicka, and Filip Makowiecki, 2010. Computational Lexicography of Multi-Word Units: How Efficient Can It Be? In *Proceedings of the COLING-MWE'10 Workshop, Beijing, China*.

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

Kosek, Iwona, 2008. *Fleksja i składnia nieciągłych imiennych jednostek leksykalnych*. Olsztyn: Publishing House of the University of Warmia and Mazury.

Krstev, Cvetana, Ranka Stankovic, Ivan Obradovic, Dusko Vitas, and Milos Utvic, 2010. Automatic construction of a morphological dictionary of multi-word units. In *Proceedings of IceTAL 2010*, volume 6233 of *Lecture Notes in Computer Science*.

Kyriacopoulou, Tita, Safia Mrabti, and Anastasia Yannacopoulou, 2002. Le dictionnaire électronique des noms composés en grec moderne. *Lingvisticæ Investigationes*, 25(1):7–28.

Marciniak, Małgorzata, Joanna Rabiega-Wiśniewska, Agata Savary, Marcin Woliński, and Celina Heliasz, 2009. Constructing an Electronic Dictionary of Polish Urban Proper Names. In *Recent Advances in Intelligent Information Systems*. Exit.

Marciniak, Małgorzata, Agata Savary, Piotr Sikora, and Marcin Woliński, 2011. Toposław - a lexicographic framework for multi-word units. *Lecture Notes in Computer Science*, 6562:139–150. Springer.

Oflazer, Kemal, Özlem Çetonoğlu, and Bilge Say, 2004. Integrating Morphology with Multi-word Expression Processing in Turkish. In *Second ACL Workshop on Multiword Expressions, July 2004*.

Przepiórkowski, Adam, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk (eds.), 2012. *Narodowy Korpus Języka Polskiego*. Warsaw: Wydawnictwo Naukowe PWN.

Radziszewski, Adam, Adam Kilgarriff, and Robert Lew, 2011. Polish Word Sketches. In *Proceedings of the 5th Language & Technology Conference*. Poznań, Poland.

Rafajlovska, Aneta and Katerina Zdravkova, 2015. Représentation des expressions composées en macédonien en tant qu'entrées lexicales en Unitex. In *Actes de la Traitement Automatique des Langues Slaves*. Caen, France: Association pour le Traitement Automatique des Langues.

Savary, Agata, 2000. Recensement et description des mots composés - méthodes et applications. PhD Thesis. Université de Marne-la-Vallée.

Savary, Agata, 2009. Multiflex: a Multilingual Finite-State Tool for Multi-Word Units. *Lecture Notes in Computer Science*, 5642:237–240.

Savary, Agata, Bartosz Zaborowski, Aleksandra Krawczyk-Wieczorek, and Filip Makowiecki, 2012. SEJFEK - a Lexicon and a Shallow Grammar of Polish Economic Multi-Word Units. In *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon*. Mumbai, India: The COLING 2012 Organizing Committee.

Silberztein, Max, 1993. Les groupes nominaux productifs et les noms composés lexicalisés. *LingvisticæInvestigationes*, 17(2).

Stanković, Ranka, Ivan Obradović, Cvetana Krstev, and Duško Vitas, 2011. Production of morphological dictionaries of multi-word units using a multipurpose tool. In *Proceedings of the Computational Linguistics-Applications Conference*. Jachranka, Poland.

Wojdak, Piotr, 2008. *Przysłówki polisegmentalne w modelu składniowym polszczyzny*. Szczecin: Publishing House of the University of Szczecin.

Woliński, Marcin, 2006. Morfeusz – a Practical Tool for the Morphological Analysis of Polish. In *Proceedings of IIS:IIPWM'06*. Springer.

[13] www.ipipan.waw.pl/nekst

[14] www.meta-net.eu/projects/cesar

[15] www.parseme.eu