SEJF - A Grammatical Lexicon of Polish Multiword Expressions

Monika Czerepowicka[1(✉)] and Agata Savary[2]

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

Abstract. We present SEJF, a lexical resource of Polish nominal, adjectival and adverbial multiword expressions. It consists of an intensional module with about 4,700 multiword 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

Multiword 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*), multiword 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 morphosyntactic variants under which they occur in texts.

In this paper we focus on Polish compounds. We present SEJF (pl. Słownik Elektroniczny Jednostek Frazeologicznych 'Electronic Dictionary of Phraseological Units'), 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

Z. Vetulani et al. (Eds.): LTC 2015, LNAI 10930, pp. 59–73, 2018.
https://doi.org/10.1007/978-3-319-93782-3_5

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) [19]. 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 [5]. 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 [12], was the first to be encoded in the dictionary. Some adjectival units were drawn from a dictionary of comparisons [3]. Adverbial units were acquired from two other monographs: [6,31].

3 Formalism and Tool

The grammatical description of MWEs in SEJF was done within Toposław [16], a lexicographic framework offering a user-friendly graphical interface over three core components:

- Morfeusz [32] – a morphological analyzer and generator of Polish simple words, containing full paradigms of over 250 thousand lemmas.
- Multiflex [25] – 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 in Sect. 8. 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. [25].

[1] http://zil.ipipan.waw.pl/SEJF.
[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. Figure 1 shows the nominal MWE *adwocat 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.

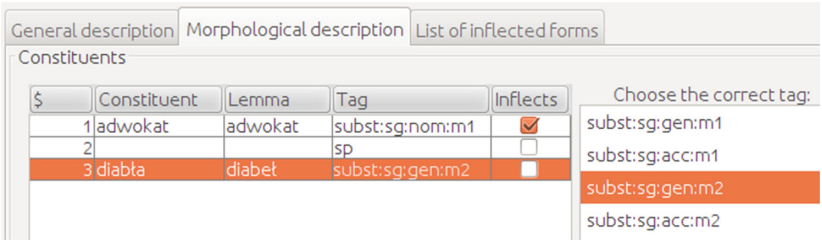

Fig. 1. Segmentation and morphosyntactic annotation of the nominal MWE *adwokat diabła* 'devil's advocate' in Toposław. The following codes are used: accusative case (`acc`), genitive case (`gen`), masculine animate gender (`m2`), masculine human gender (`m1`), singular (`sg`), space (`sp`), and substantive (`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. Figure 2 shows the inflection graph `NC-O_N` (cf. Table 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. Sect. 5 and Fig. 4).

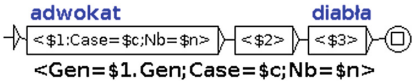

Fig. 2. Inflection graph `NC-O_N` for the nominal MWE *adwokat diabła* 'devil's advocate'.

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.

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
adwokatach diabła ⊠ adwokat diabła:subst:pl:loc:m1
adwokacie diabła ⊠ adwokat diabła:subst:sg:voc:m1
adwokaci diabła ⊠ adwokat diabła:subst:pl:voc:m1

Fig. 3. Inflection paradigm for the nominal MWE *adwokat diabła* 'devil's advocate'.

4 Contents of the Lexicon

Table 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.

Table 1. Contents of the lexicon.

	MWE 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

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. Table 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 agreeing adjective in both orders.

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`).

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

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 which are stylistically marked and show the speaker's pejorative attitude to the persons named by the multiword noun. Grammatically speaking, depreciative forms differ from the regular ones in plural nominative and vocative, namely they 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, as in *automatyczna sekretarka*, *sekretarka automatyczna* (lit. *automatic secretary*) '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.

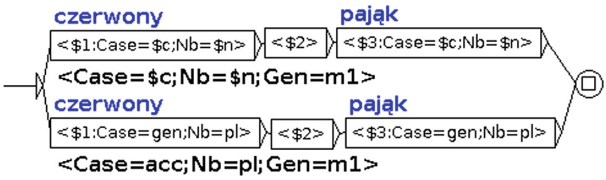

Fig. 4. Inflection graph `NC-0_N` describing a masculine gender fluctuation in *czerwony pająk* (lit. *red spider*) 'communist'.

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

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

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
adwokatach diabła ⊠ adwokat diabła:subst:pl:loc:m1
adwokacie diabła ⊠ adwokat diabła:subst:sg:voc:m1
adwokaci diabła ⊠ adwokat diabła:subst:pl:voc:m1
adwokaty diabła ⊠ adwokat diabła:subst:pl:nom:m2
adwokaty diabła ⊠ adwokat diabła:subst:pl:voc:m2

Fig. 5. Inflection paradigm for the nominal MWE *czerwony pająk* (lit. *red spider*) 'communist'.

Table 3. Contents of the evaluation corpus.

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

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 multiword 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 (Polish, 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. Table 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 Table 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%)

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

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.

Table 4. Lexicon coverage evaluated against the 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

7 Application to Automatic Treebank Annotation

SEJF, as a high-quality grammatical resource, can be used in a variety of NLP applications. Notably, its utility for automatic treebank annotation was recently tested by [26]. The task was to project 3 available resources of Polish MWEs, including SEJF, on a Polish constituency treebank, Składnica [30], which contained no initial MWE annotations. This task is important since MWE-annotated treebanks are scarce and constitute bottlenecks in the MWE-oriented research.

The extensional version of SEJF, containing the 88,000 morphosyntactic variants of MWEs, as in Figs. 3 and 5, was used in the experiments. The SEJF entries were transformed into queries and evaluated against the treebank. As a result, the treebank subtrees containing continuous sequences of leaves corresponding to the SEJF entries, and respecting the relevant morphological constraints, were automatically marked. The automatic projection was followed by a manual validation. The SEJF-specific outcome of this process is shown in Table 5.

The true positives (TP) correspond to the MWEs from SEJF correctly identified in the treebank by the automatic projection. The most frequently repeated MWEs in this group are adverbials like *przede wszystkim* (lit. *before all*) 'mainly' (25 occ.), *na pewno* (lit. *on sure*) 'certainly' (12 occ.), *na miejscu* (lit. *on place*) 'instantaneously/relevant' (12 occ.), *po prostu* (lit. *on simple*) 'simply' (12 occ.), *od razu* (lit. *from time*) 'immediately' (9 occ.), etc. False positives (FP) are errors resulting from bugs in the mapping procedure. The compositional readings (CRead) are cases like those in examples (1)–(4), sometimes included in larger

Table 5. Results of projecting the SEJF entries on the Składnica treebank, including true positives (TP), false positives (FP), compositional readings (CRead), and idiomaticity rate (IRate).

	Nouns		Adverbs		Others		All categories	
	Occ.	Lemmas	Occ.	Lemmas	Occ.	Lemmas	Occ.	Lemmas
TP	209	154	153	67	6	4	368	225
FP	0	0	4	4	1	1	5	5
CRead	17	12	19	11	0	0	36	23
All occ.	226	165	176	78	7	3	409	248
IRate	0.92	n/a	0.89	n/a	1	n/a	0.91	n/a

MWEs, as in (3)–(4). The idiomaticity rate [8], i.e., the ratio of occurrences with idiomatic reading to all correctly recognized occurrences, is relatively high, especially for nominal MWEs. The MWE with the highest number of compositional readings is *na miejscu* (lit. *on place*) 'instantaneously/relevant' as in example (2) (6 occ.). Note that the same MWE also has a high number of idiomatic occurrences (12).

(1) ... w **drugiej połowie** XIX wieku
'... in the **second half** of the 19th century'
coinciding MWE: *druga połowa* (lit. *second half*) 'one's husband or wife'

(2) ... był **na miejscu** zdarzenia
'... he was **at the place** of the event'
coinciding MWE: *na miejscu* (lit. *on place*) 'instantaneously/relevant'

(3) ... od czasu **do czasu** zazdrościła przyjaciółkom
'... from time **to time** she envied her friends'
coinciding MWE: *do czasu* (lit. *to time*) 'temporarily'

(4) Zmiany dokonane w Oplu Fronterze wyszły mu **na dobre**
(lit. 'Changes operated in Opel Fronter went out **for the good**.')
'Changes operated in Opel Fronter turned to its advantage.
coinciding MWE: *na dobre* (lit. *for the good*) 'permanently'

These results show that SEJF can be successfully applied to automatic treebank annotation, due to the fine-grained grammatical descriptions contained in this resource, and to the high idiomaticity rate of Polish MWEs. Automatic disambiguation, i.e. distinguishing idiomatic from compositional readings, remains a challenge in cases when a MWE does not occur although it might (i.e. most morphosyntactic constraints it imposes are fulfilled). Note, however, that cases like (1)–(2) can be resolved if the MWE entry is enriched with information on its valency, i.e. its allowed or prohibited non-lexicalized modifiers. Efforts towards rich syntactic encoding of this kind have notably been undertaken in valence dictionaries with phraseological components [20], and synergies between such formalisms and SEJF-like e-dictionaries are being investigated.

8 Related Work

Although MWEs are still under-represented in language resources and tools, efforts have been put towards bridging this gap from the e-lexicographical point of view in many languages, as discussed in [15]. 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 [28], English [24], Greek [14] and others. Lexicons similar to SEJF, following the Multiflex paradigm, exist or are under construction for Serbian [13], Greek [9], and Macedonian [22]. Various e-lexicographic frameworks were developed for the creation of MWE e-lexicons notably in Turkish [18], Basque [2], Dutch [11], Serbian [29] and Hebrew [1], the last one also covers verbal MWEs.

On the Polish ground, SEJF is one of three grammatical lexicons of Polish multiword units built under Toposław. The two other resources are: (i) SAWA[9] [17], a grammatical lexicon of Warsaw urban proper names (streets, squares, bus stops, and other objects linked to the Warsaw communication network), (ii) SEJFEK[10] [27], a grammatical lexicon of Polish economic terminology containing over 11,000 specialized nominal compounds. Complementary formalisms for inflectional paradigms of Polish MWEs have been presented in [5,10].

9 Conclusions and Future Work

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, including richly annotated corpora such as treebanks. While the coverage of compound adverbs offered by SEJF is reasonable, its contents in terms of compound nouns and adjectives should be extended, as shown by the evaluation results. Additional corpora can underlie this further work, including those available via Sketch Engine[11] with collocation support [21].

As mentioned in Sect. 5 the description of nominal MWEs in masculine human gender is not fully satisfactory with Toposław, due to the impossibility to generate the depreciative forms of these expressions. These problems can be overcome with a recent follow-up of Toposław, called Werbosław, which allows the user to gather several flexemes of the same lemma in one lexeme.

More precisely, according to [23], a lexeme is understood as an abstract unit of language containing all forms connected with the same lexical meaning. For instance, *adwokaci diabła* 'devil's advocates' in human masculine (`m1`) and *adwokaty diabła* 'devil's advocates (depr.)' in human animate (`m2`), belong to the

[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/pltenten-corpus/.

same lexeme. A lexeme can further subdivide into several flexemes [4], i.e. morphosyntactically homogeneous sets of forms belonging to the same lexeme. Since a substantive (`subst`) is defined in the NKJP-Morfeusz tagset as a class which inflects for case an number and *has* (invariable) gender, *adwokaty/adwokaci diabła* in two different genders cannon belong to the same nominal flexeme. Thus, only the forms in `m1` are classified into the flexeme of class `subst`. The forms in `m2` are separated in another flexeme of class `depr` (depreciative form), defined as inflecting for case and having number and gender.

Toposław is flexeme-oriented, therefore these two flexemes would have to be described separately, which would be unnatural, since they both share the same lemma *adwokat diabła* 'devil' advocate'. Werbosław, conversely, is lexeme-oriented. Each of its individual entries is a lexeme whose class has to be selected by the lexicographer in the initial stage of the description, as shown in Fig. 6.

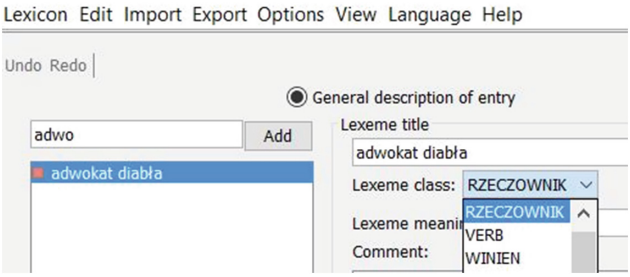

Fig. 6. Selecting the class (RZECZOWNIK 'noun', VERB, etc.) of the lexeme *adwokat diabła* 'devil's advocate' in Werbosław.

A lexeme is a unit of a higher order as compared to a flexeme. Therefore, the next step is to define the list of flexemes associated with a given lexeme, as shown in Fig. 7.

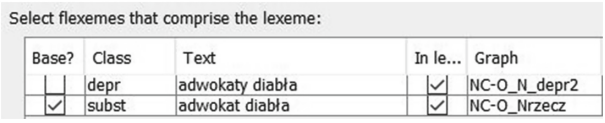

Fig. 7. Selecting the flexemes (here: `depr` and `subst`) associated with the lexeme (here `RZECZOWNIK` 'noun') *adwokat diabła* 'devil's advocate' in Werbosław.

The description of each of the flexemes follows the same steps as in Toposław, i.e. consists of analyzing each component morphosyntactically and selecting the right inflection graph. Fig. 8 shows the graph for the depreciative flexeme of *adwokat diabła* 'devil's advocate'. Recall that the depreciative forms only show in the nominative and vocative case in plural, i.e. Morfeusz only generates these

two forms for a depreciative noun. Therefore, the number in the graph can be fixed to plural (`Nb=pl`) and the case inflection can be unrestricted (`Case=$c`).

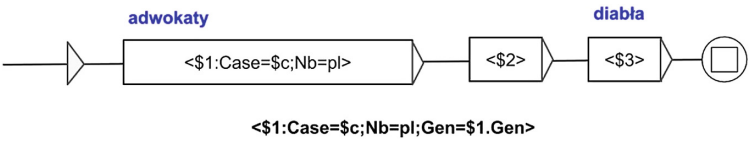

Fig. 8. Inflection graph `NC-O_N_depr2` describing the depreciative flexeme *adwokaty diabła* 'devil's advocates' of the nominal lexeme *adwokat diabła* 'devil's advocate'.

As a result, the full description of the lexeme yields an enhanced list of the inflected forms shown in Fig. 9. Note the occurrence of the two depreciative forms in the `m2` gender, as opposed to the paradigm obtained with Toposław in Fig. 3.

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
adwokaty diabła ⊠ adwokat diabła:subst:pl:nom:m2
adwokaty diabła ⊠ adwokat diabła:subst:pl:voc:m2

Fig. 9. Inflection paradigm of the nominal lexeme *adwokat diabła* 'devil's advocate', containing regular and depreciative forms.

An even more challenging behavior is exhibited by Polish verbs, where a single lexeme consists of up to 12 different flexemes. For instance, the non-past flexemes (`fin`) like *robi* 'does' inflect for number and person, and have aspect; the past flexeme (`praet`) like *robił* 'did' inflect for number, gender and agglutination, and have aspect; the gerunds (`ger`) like `robienie` 'doing' inflect for number, case and negation, and have gender and aspect; etc. Thanks to flexeme-to-lexeme shift operated in Werbosław, verbal mutiword expressions, such as *odwracać kota ogonem* (lit. *to turn the cat with its tail to the front*) 'to distort the facts', can now be conveniently described. Such expressions are being currently addressed within the Verbel project[12] [7].

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

Note finally that the descriptive framework of Toposław and Werbosław does not account for more complex syntactic phenomena such as diathesis change and long-distance dependencies. Therefore, verbal MWEs can only be described from the point of view of their inflectional and word-order variants. Other syntactic variants (passivisation, nominalisation, internal modification, etc.) call for an expressive power close to full-fledged syntactic formalisms. The same applies to nominal, adjectival and adverbial MWEs with open slots, such as *[czyjaś] prawa ręka* (lit. *[someone's] right hand*) '[someone's] main assistant'. Despite these shortcomings, we hope to have shown that our proposals prove useful for the description of large classes of MWEs whose frequency in a corpus is usually rather high.

Acknowledgements. This work has been supported by three projects: (i) Nekst(http://www.ipipan.waw.pl/nekst), funded by the European Regional Development Fund and the Polish Ministry of Science and Higher Education, (ii) CESAR (http://clip.ipipan.waw.pl/CESAR) - a European project (CIP-ICT-PSP-271022), part of META-NET, (iii) IC1207 COST action PARSEME(http://www.parseme.eu).

References

1. Al-Haj, H., Itai, A., Wintner, S.: Lexical representation of multiword expressions in morphologically-complex languages. Int. J. Lexicogr. **27**(2), 130–170 (2014)
2. Alegria, I., Ansa, O., Artola, X., Ezeiza, N., Gojenola, K., Urizar, R.: Representation and treatment of multiword expressions in Basque. In: Proceedings of the ACL 2004 Workshop on Multiword Expressions, pp. 48–55 (2004)
3. Bańko, M.: Słownik porównań. Polish Scientific Publishers PWN, Warsaw (2004)
4. Bień, J.S.: Koncepcja słownikowej informacji morfologicznej i jej komputerowej weryfikacji. Rozprawy Uniwersytetu Warszawskiego 383 (1991)
5. Broda, B., Derwojedowa, M., Piasecki, M.: 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 2007), pp. 237–246 (2007)
6. Czerepowicka, M.: Opis powierzchniowoskładniowy wyrażeń niestandardowych typu «na lewo», «do dziś», «po trochu», «na zawsze» we współczesnym języku polskim. Akademicka Oficyna Wydawnicza EXIT, Warszawa (2006)
7. Czerepowicka, M., Kosek, I., Przybyszewski, S.: O projekcie elektronicznego słownika odmiany frazeologizmów czasownikowych. Polonica **34**, 115–123 (2014)
8. El Maarouf, I., Oakes, M.: Statistical measures for characterising MWEs. In: IC1207 COST PARSEME 5th General Meeting (2015). http://typo.uni-konstanz.de/parseme/index.php/2-general/138-admitted-posters-iasi-23-24-september-2015
9. Foufi, V.: Les noms composés A(A)N du Grec Moderne et leurs variantes. In: Kakoyianni Doa, F. (ed.) Penser Le Lexique-Grammaire : Perspectives Actuelles. Editions Honoré Champion, Paris (2013)
10. Graliński, F., Savary, A., Czerepowicka, M., Makowiecki, F.: Computational lexicography of multi-word units. How efficient can it be? In: Proceedings of the COLING-MWE 2010 Workshop, Beijing, China (2010)

11. Grégoire, N.: DuELME: a Dutch electronic lexicon of multiword expressions. Lang. Resour. Eval. **44**(1–2), 23–39 (2010)
12. Kosek, I.: Fleksja i składnia nieciągłych imiennych jednostek leksykalnych. Publishing House of the University of Warmia and Mazury, Olsztyn (2008)
13. Krstev, C., Stanković, R., Obradović, I., Vitas, D., Utvić, M.: Automatic construction of a morphological dictionary of multi-word units. In: Loftsson, H., Rögnvaldsson, E., Helgadóttir, S. (eds.) NLP 2010. LNCS (LNAI), vol. 6233, pp. 226–237. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-14770-8_26
14. Kyriacopoulou, T., Mrabti, S., Yannacopoulou, A.: Le dictionnaire électronique des noms composés en grec moderne. Lingvist. Investig. **25**(1), 7–28 (2002)
15. Losnegaard, G.S., Sangati, F., Escartín, C.P., Savary, A., Bargmann, S., Monti, J.: Parseme survey on MWE resources. In: Chair, N.C.C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA), Paris, May 2016
16. Marciniak, M., Savary, A., Sikora, P., Woliński, M.: Toposław – a lexicographic framework for multi-word units. In: Vetulani, Z. (ed.) LTC 2009. LNCS (LNAI), vol. 6562, pp. 139–150. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-20095-3_13
17. Marciniak, M., Rabiega-Wiśniewska, J., Savary, A., Woliński, M., Heliasz, C.: Constructing an electronic dictionary of polish urban proper names. In: Recent Advances in Intelligent Information Systems, pp. 233–246. Exit (2009)
18. Oflazer, K., Çetonoğlu, Özlem., Say, B.: Integrating morphology with multi-word expression processing in Turkish. In: Second ACL Workshop on Multiword Expressions, pp. 64–71 (2004)
19. Przepiórkowski, A., Bańko, M., Górski, R.L., Lewandowska-Tomaszczyk, B. (eds.): Narodowy Korpus Języka Polskiego. Wydawnictwo Naukowe PWN, Warsaw (2012)
20. Przepiórkowski, A., Hajnicz, E., Patejuk, A., Woliński, M.: Extended phraseological information in a valence dictionary for NLP applications. In: Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing (LG-LP 2014), pp. 83–91. Association for Computational Linguistics and Dublin City University, Dublin, Ireland (2014). http://www.aclweb.org/anthology/siglex.html#2014_0
21. Radziszewski, A., Kilgarriff, A., Lew, R.: Polish word sketches. In: Proceedings of the 5th Language and Technology Conference, Poznań, Poland, pp. 237–242, November 2011
22. Rafajlovska, A., Zdravkova, K.: Représentation des expressions composées en macédonien en tant qu'entrées lexicales en Unitex. In: Actes de la Traitement Automatique des Langues Slaves, pp. 1–8. Association pour le Traitement Automatique des Langues, Caen, France, June 2015. http://www.atala.org/taln_archives/TASLA/TASLA-2015/tasla-2015-court-001
23. Saloni, Z.: Klasyfikacja gramatyczna leksemów polskich. Język Polski **54**(1), 3–13 (1974)
24. Savary, A.: Recensement et description des mots composés - méthodes et applications, Ph.D. Thesis. Université de Marne-la-Vallée (2000)
25. Savary, A.: Multiflex: a multilingual finite-state tool for multi-word units. In: Maneth, S. (ed.) CIAA 2009. LNCS, vol. 5642, pp. 237–240. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-02979-0_27

26. Savary, A., Waszczuk, J.: Projecting multiword expression resources on a polish treebank. In: Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing, pp. 20–26. Association for Computational Linguistics, Valencia, Spain April 2017. http://www.aclweb.org/anthology/W17-1404
27. Savary, A., Zaborowski, B., Krawczyk-Wieczorek, A., Makowiecki, F.: 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, pp. 195–214. The COLING 2012 Organizing Committee, Mumbai, India, December 2012. http://www.aclweb.org/anthology/W12-5116
28. Silberztein, M.: Les groupes nominaux productifs et les noms composés lexicalisés. Lingvist. Investig. **17**(2), 405–425 (1993)
29. Stanković, R., Obradović, I., Krstev, C., Vitas, D.: Production of morphological dictionaries of multi-word units using a multipurpose tool. In: Proceedings of the Computational Linguistics-Applications Conference, Jachranka, Poland, pp. 77–84, October 2011
30. Świdziński, M., Woliński, M.: Towards a bank of constituent parse trees for polish. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2010. LNCS (LNAI), vol. 6231, pp. 197–204. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15760-8_26
31. Wojdak, P.: Przysłówki polisegmentalne w modelu składniowym polszczyzny. Publishing House of the University of Szczecin, Szczecin (2008)
32. Woliński, M.: Morfeusz - a practical tool for the morphological analysis of polish. In: Kłopotek, M.A., Wierzchoń, S.T., Trojanowski, K. (eds.) Intelligent Information Processing and Web Mining. AINSC, vol. 35. Springer, Heidelberg (2006). https://doi.org/10.1007/3-540-33521-8_55