Polish corpus of verbal multiword expressions

Agata Savary, Jakub Waszczuk

Polish corpus of verbal multiword expressions

Agata Savary
University of Tours, LIFAT
France
first.last@univ-tours.fr

Jakub Waszczuk
Heinrich Heine Universität Düsseldorf
Germany
waszczuk@phil.hhu.de

Abstract

This paper describes a manually annotated corpus of verbal multi-word expressions in Polish. It is among the 4 biggest datasets in release 1.2 of the PARSEME multiligual corpus. We describe the data sources, as well as the annotation process and its outcomes. We also present interesting phenomena encountered during the annotation task and put forward enhancements for the PARSEME annotation guidelines.

1 Introduction

Multiword expressions (MWEs), such as ***at times***, ***red tape*** or ***take off***, are word combinations with idiosyntactic behaviour, notably non-compositional semantics. Therefore, they constitute a challenge for linguistic modelling and semantically-oriented text processing. Verbal MWEs (VMWEs), like *to* ***bear*** *sth* ***in mind***, are particularly challenging due to their partly regular and partly idiosyncratic morphosyntactic flexibility, and their frequent discontinuity in texts (Savary et al., 2018). These challenges are even harder in languages like Polish, with rich inflectional morphology and a relatively free word order.

In order to bring progress to MWE modelling and processing, the PARSEME initiative has been coordinating multilingual efforts towards annotating VMWEs in corpora and their automatic identification in texts (Ramisch et al., 2018). This paper describes the most recent version of the Polish corpus, which is the 4th biggest dataset in edition 1.2 of the PARSEME suite (Ramisch et al., 2020). We show how the basic definitions from the PARSEME methodology apply to Polish (Sec. 2), we analyse the state of the art in Polish MWE-annotated corpora (Sec. 3), we describe the construction of our corpus (Sec. 4) and its outcomes (Sec. 5). We evoke some challenging phenomena and lessons learned from manual annotation (Sec. 6) and on this basis we put forward some recommendations for enhancing the PARSEME annotation guidelines (Sec. 7). Finally, we conclude and sketch perspectives for future work (Sec. 8).

2 Verbal multiword expressions in Polish

The Polish VMWE dataset is integrated in the PARSEME corpus annotation methodology. The latter increasingly relies on (version 2 of) Universal Dependencies (UD), a *de facto* standard for morphosyntactic annotation (Nivre et al., 2020). Thus, we largely follow the definitions of both initiatives.

Firstly, we differentiate *words* (linguistically motivated units undergoing syntactic relations) from *tokens* (technical items resulting from corpus segmentation). This difference is notably visible in multiword tokens (MWTs), highly productive in some Polish verb forms like *widział|em* 'I saw' (cf. Sec. 6).

We further understand MWEs as combinations of words which: (i) have at least two lexicalized components, i.e. components always realized by the same lexemes, (ii) display lexical, morphological, syntactic or semantic idiosyncrasies. For instance, in ***postawić*** *kogoś* ***w stan*** *gotowości* (lit. 'put sb into state of-readiness') 'to put sb on alert',[1] the object *stan* 'state' must receive a complement (here: *gotowości*

[1] Henceforth, the lexicalized components of MWEs are highlighted in bold, an asterisk (*) means ungrammaticality, while a dash (#) signals a substantial change in meaning with respect to the original expression.

'of-readiness'). This is idiosyncratic since such a complement is not required in non-idiomatic constructions, like *postawić miotłę w kąt* 'put the-broom in the-corner'. However, this complement, even if compulsory, is not lexically fixed: ***postawić*** *kogoś* ***w stan*** *gotowości/pogotowia/oskarżenia/upadłości/etc.* (lit. 'to-put sb into state of-readiness/emergency/accusation/bankruptcy/etc.'). Therefore, only the words ***postawić w stan*** count as lexicalized components.

VMWEs are MWEs whose *canonical form*, i.e. the least syntactically marked form keeping the idiomatic reading, is such that its syntactic head is a verb V and its other lexicalized components form phrases directly dependent on V. This means that a canonical form is a weakly connected graph (i.e. fully connected if directions of the dependencies are disregarded). Consider the example *obiektywna* ***rola****, jaką uczelnie* ***odgrywają*** *w Polsce* 'objective role which universities play in Poland'. Here, the noun *rola* 'role' heads the verb *odgrywają* 'play' rather than vice versa. Since a construction with a relative clause is syntactically more marked than without it, we have to transform it into a canonical form, e.g. *uczelnie* ***odgrywają*** *obiektywną* ***rolę*** *w Polsce* 'universities play an objective role in Poland'. This is why we can consider this candidate as headed by the verb and passing the light verb construction tests.

Five out of the ten VMWE (sub)categories from the PARSEME guidelines v 1.2 are relevant to Polish:

- *Inherently reflexive verbs* (IRV) are combinations of a verb v and a reflexive clitic (RCLI) r, such that at least one of the non-compositionality conditions holds: (i) v never occurs without r, as in ***gapić się*** (lit. 'stare RCLI') 'stare'; (ii) r distinctly changes the meaning of v, like in ***stać się*** (lit. 'stand RCLI') 'become'; (iii) r changes the subcategorization frame of v, like in ***dziwić się*** *takim reakcjom* (lit. 'surprise RCLI such reactions.DAT') 'be surprised by such reactions'[2]
- *Light verb constructions* (LVCs) are combinations of a verb v and a noun n (with an optional preposition) in which v is semantically void or bleached, and n is a predicate, i.e. it is abstract and has semantic arguments. Two subcategories are defined. In an *LVC.full*, v's subject is n's semantic argument. For instance, in ***wezmę odwet*** 'I-will-take revenge' the (pro-dropped) subject of the verb ('I') is the agent of the revenge and the verb adds no meaning to the noun. In an *LVC.cause*, n is no semantic argument of but adds a causative meaning to v. For instance, in *Ela* ***podsunęła*** *Janowi tę* ***myśl*** (lit. 'Ela moved Jan this thought') 'Ela suggested this thought to Jan', Jan might have a thought without any intervention of Ela (i.e. she is not a semantic argument of the thought). But in this precise sentence, Ela is the cause of Jan's thought.
- *Verbal idioms* (VIDs) are verb phrases of various syntactic structures which contain cranberry words or exhibit lexical, morphological or syntactic inflexibility. For instance, in ***nosić*** *kogoś* ***na rękach*** (lit. 'carry sb on hands') 'to give special care to sb', when the noun is inflected in number or replaced by a semantically related word, the idiomatic meaning is lost (#*nosić kogoś na rękę/ramionach* 'carry sb on hand/shoulders').

Another category potentially pertaining to Polish are *inherently adpositional verbs* (IAVs), defined as combinations of a verb v and an adposition a (i.e. a preposition in Polish), such that: (i) v never occurs without a, as in ***polegać na*** *kimś* 'to rely on someone', or (ii) a significantly changes v's meaning, as in ***o*** *co tu* ***chodzi****?* (lit. 'about what here goes') 'what is the matter here?'. IAVs were to be experimentally and optionally annotated in PARSEME corpora since version 1.1. In Polish, we performed this annotation in edition 1.1 but IAVs proved too hard to distinguish from 'regular' verbal valency with the current annotation guidelines. Therefore, we abandoned the IAV annotation in edition 1.2 of the Polish corpus.

3 Multiword expressions in Polish treebanks

In previous work on modelling and annotating Polish MWEs, lexicon, grammar and treebank construction efforts have often been closely related.

Głowińska and Przepiórkowski (2010) and Głowińska (2012) present the manual shallow syntactic annotation of the National Corpus of Polish (NKJP).[3] The whole corpus follows multilayer annotation

[2]When the verb *dziwić* 'surprise' takes a regular non-reflexive object, it admits a complement in instrumental but not in dative (*dziwiła go swoim zachowaniem/*swojemu zachowaniu* 'she-surprised him her behavior.INST/DAT').

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

principles. In particular the layer of syntactic groups (roughly chunks), builds upon the layer of the so-called syntactic words, which in turn builds upon the layer of tokens. The layer of syntactic words includes a number of (mostly) continuous MWEs such as multiword prepositions (***w duchu*** *czegoś* 'in the spirit of sth'), adverbs (***do czysta*** 'completely') or conjunctions (***a zatem*** 'that is'). Those are not explicitly marked as MWEs but can be queried by looking for word nodes which point at least two token nodes.[4] All MWEs delimited in this way are decorated with their parts of speech. Verbal MWEs are not covered. NKJP is released with a shallow grammar developed for its automatic pre-annotation. Among the 1,187 grammar rules, 350 are lexicalized rules describing MWEs.

Fragments of the NKJP corpus have been transformed into the constituency treebank Składnica. On top of the previous morphosyntactic annotation described above, the constituency parser *Świgra* produced candidate trees, which were then manually disambiguated (Świdziński and Woliński, 2010). A recent version of Składnica (Woliński et al., 2018) integrates data from a valency dictionary Walenty.[5] Walenty has a rich phraseological component (Przepiórkowski et al., 2014; Hajnicz et al., 2016) and a semantic layer. On the morphosyntactic level, verbal MWEs are represented as valency frames in which some arguments are lexically fixed, e.g. ***zobaczyć*** *coś* ***na własne oczy*** (lit. 'to see sth on own eyes') 'to see sth for oneself' receives a frame with the head verb *zobaczyć* '*see*', a free subject and object, and a lexicalized complement *na własne oczy* '*on own eyes*'. On the semantic level, adverbial, nominal, adjectival and other MWEs can appear as lexicalized elements of verbal frames, e.g. a multiword adverb ***w trupa*** '*into a dead body*' occurs as a possible lexicalized realization of the semantic role of manner in the verbal frame of *upić się* 'get drunk', the whole combination meaning 'to get totally drunk'. The latest dowloadable Walenty version (from 2016) contains notably over 60,000 syntactic verbal frames, 14,295 of which have lexicalized arguments, i.e. correspond to VMWEs entries. Walenty frames were integrated into the Świgra's grammar, which was then used to enhance Składnica. The latter does not seem to explicitly indicate which tree nodes correspond to lexicalized components of VMWEs from Walenty.

Such efforts of making MWE occurrences in Składnica explicit were undertaken in two Polish UD treebanks. In the Polish Dependency Bank (PDB), Wróblewska (2012) automatically converted the continuous MWEs into dependency chains using the `mwe` relation (pertaining to UD version 1). Later, PDB was enlarged with new texts and converted into UD version 2, with the `fixed` and `flat` dependencies marking morphologically fixed MWEs and named entities, respectively. The number of both types of labels in PDB version 2.5 is 3,850 and 5,525, respectively. Later the whole NKJP corpus[6] was enriched with dependencies, using a parser trained on PDB, and manually correcting major flaws (Wróblewska, 2020). There, the `fixed` and `flat` dependencies most probably follow the same principles as in PDB, but no statistics of these specific labels were available at the time of writing. It is also unclear if any fixed MWEs were marked except those predicted by the parser, i.e. the coverage of MWEs is unclear.

In parallel to the above treebanking efforts involving Świgra, Walenty and UD conversion, similar work was done in the Lexical Functional Grammar framework. Patejuk and Przepiórkowski (2014) developed an LFG grammar of Polish, integrated with Walenty, parsed texts stemming mainly from NKJP, and manually disambiguated them to obtain an LFG treebank. They further performed an automatic conversion of this treebank into the UD version 2 (Przepiórkowski and Patejuk, 2020), including the so-called enhanced dependencies[7] The resulting UD-LFG treebank contains 144 and 884 `fixed` and `flat` dependencies, respectively. Like in PDB, the former are limited mostly to continuous morphosyntactically fixed MWEs, and the latter to named entities, i.e. the information about verbal MWEs from Walenty is not propagated to the treebank, and nominal/adjectival MWEs are neglected.

An effort focused on explicitly marking occurrences of large classes of MWEs in Składnica was undertaken by Savary and Waszczuk (2017). They used 3 resources: (i) Walenty, (ii) the named entity annotation layer of the NKJP corpus, and (iii) SEJF, an electronic lexicon of Polish nominal, adjectival and adverbial MWEs, with 4,700 multiword lemmas, 160 inflection graphs and 88,000 automatically gener-

[4]Such a query will however also return multi-token words which are no MWEs, for instance analytical forms of verbs.

[5]http://zil.ipipan.waw.pl/Walenty

[6]More precisely, the manually annotated 1-million-token subcorpus of NKJP, called NKJP1M is concerned here.

[7]Enhanced dependencies enable overt marking of some relations which are implicit in the basic UD format, notably arguments which are ellipted or shared by conjuncts. A syntactic graph containing enhanced dependencies is not a tree.

ated inflected forms (Czerepowicka and Savary, 2018). These 3 resources were automatically mapped on Składnica, and the outcome was manually validated, which resulted in the SkładnicaMWE treebank with explicit marking of over 1,300 named entities, as well as 450 verbal and 400 nominal/adjectival/adverbial MWEs.[8] Differently from the previous efforts, this time, the treebank remains in its original constituency format, and information about MWEs is added to selected tree nodes as additional features, together with pointers to those lexical nodes which represent lexicalized components of the MWEs. This is in sharp contrast with the UD encoding, where dependencies indicating the MWE status potentially compete with those marking the syntactic relations. SkladnicaMWE is also the first Polish treebank with an explicit marking of verbal, nominal and adjectival MWEs. This resource would be worth extending with entries from VERBEL, a more recent grammatical e-lexicon of verbal MWEs.[9]

In the context of this state of the art, we describe the first attempt towards systematic annotation of Polish verbal MWEs in running text. We do not use any pre-annotation methods so as to avoid bias. The resulting resource is fully integrated into the PARSEME suite of multilingual treebanks annotated for verbal MWEs (Savary et al., 2018; Ramisch et al., 2018; Ramisch et al., 2020). It follows the cross-lingually unified and validated annotation guidelines and the centralized quality insurance methodology.

4 Constructing the Polish VMWE-annotated corpus

All the manual annotations of VMWEs were performed on texts coming from one of three (more or less overlapping) sources (cf. Sec. 3): (i) NKJP1M, a 1-million word manually annotated subcorpus of NKJP; (ii) PCC, Polish Coreference Corpus (Ogrodniczuk et al., 2015); (iii) PDB (cf. Sec. 3). From the first two sources we only took newspaper texts, while PDB provided a mixture of news, periodicals, literature, fiction, popular science, social media, parliamentary debates and manuals. The source corpus and the text genre of each sentence are indicated in its comment, as documented in the corpus repository.[10]

Like all corpora in the PARSEME suite v 1.2, the Polish dataset is released in the `.cupt` format,[11] an instance of the CoNLL-U Plus format[12] defined for annotations built upon UD corpora. Fig. 1 shows the first sentence of a corpus file. The first line is global to the whole corpus and gives the headings of the 11 columns. The first 10 stem from the CoNLL-U format, and the 11th contains the VMWE annotations. Here, tokens 1–2 belong to an IRV ***postarać się*** (lit. 'try RCLI') 'try hard', which overlaps with another IRV encompassing tokens 2–5 ***się pogodzić*** (lit. 'RCLI reconcile') 'make it up (with someone)'.

```
# global.columns = ID FORM LEMMA UPOS XPOS FEATS HEAD DEPREL DEPS MISC PARSEME:MWE
# text = Postaraj się z tym pogodzić.
# source_sent_id = http://hdl.handle.net/11234/1-3105 UD_Polish-PDB/pl_pdb-ud-train.conllu train-s11054
1  Postaraj  postarać  VERB  impt:sg:sec:perf  Aspect=Perf|Mood=Imp|...  1  root  _  _  1:IRV
2  się       się       PRON  part              PronType=Prs|Reflex=Yes   1  expl:pv  _  _  1;2:IRV
3  z         z         ADP   prep:inst:nwok    AdpType=Prep|...          4  case  _  _  *
4  tym       to        PRON  subst:sg:inst:n:ncol  Case=Ins|Gender=Neut|...  1  obl:arg  _  _  *
5  pogodzić  pogodzić  VERB  inf:perf          Aspect=Perf|VerbForm=Inf|...  1  xcomp  _  SpaceAfter=No  2
6  .         .         PUNCT interp            PunctType=Peri            1  punct  _  _  *
```

Figure 1: First sentence of a corpus, with two overlapping VMWEs.

Henceforth, the first 10 columns of a `.cupt` file will be referred to as morphosyntactic annotation. By morphological annotation alone we mean columns 3–6 (LEMMA, UPOS, XPOS and FEATS) and by syntactic annotation alone, columns 7–8 (HEAD and DEPREL). Morphosyntactic annotation is considered compatible with UD (in version 1 or 2) if is follows the UD annotation guidelines (in the corresponding version).[13] It is further considered compatible with a certain release of UD, e.g. with UD 2.5 if, for the same sentences, it contains the same data as this release or if it is automatically generated using a parser trained on this release.

[8] http://zil.ipipan.waw.pl/SkładnicaMWE

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

[10] `https://gitlab.com/parseme/parseme_corpus_pl`

[11] `http://multiword.sourceforge.net/cupt-format`

[12] https://universaldependencies.org/ext-format.html

[13] See `https://universaldependencies.org/guidelines.html` for version 2, and `https://universaldependencies.org/docsv1/` for version 1.

The corpus in version 1.2 extends and enhances the one in version 1.1. Firstly, we annotated new texts and made the previous and the new annotations mutually consistent (Sec. 4.1). Secondly, we updated the morphosyntactic annotation to make it compatible with the UD version 2.5 (see Sec. 4.2). Finally, we provided a companion raw corpus, automatically annotated for morphosyntax (Sec. 4.3) and meant for automatic discovery of unseen VMWE.

4.1 Manual annotation

To increase the size of the manually annotated corpus, we selected new sentences from PDB. The manual annotation, based on the PARSEME guidelines v 1.2,[14] was performed by one native annotator with the PARSEME-customized online annotation platform FLAT.[15] No automatic pre-annotation had been performed, but all verbal tokens were underlined in the FLAT interface, so as to easily spot potential VMWEs. In hard cases, the decision process was supported by an NKJP concordancer,[16] Polish online dictionaries[17] and, sporadically, the valence e-dictionary Walenty (cf. Sec. 3). All the resulting manual annotations, both the new ones and those from version 1.1, were checked for consistency, by the same annotator, with a PARSEME tool (Savary et al., 2018), grouping annotated and non-annotated instances of the same lemma sets. At the same time, known errors from edition 1.1 were manually corrected. Finally, 900 sentences taken from the newly annotated texts were double-annotated by another native expert for the sake of inter-annotator agreement estimation (cf. Sec. 5). Some interesting phenomena, hard challenges and decisions taken during manual annotation are documented in Sec. 6.

4.2 Updating the morphosyntactic annotation

New VMWE annotations were performed on UD-2.5-compatible files, while the corpus in version 1.1 used an older UD tagset. Therefore, upgrades to UD 2.5 were performed for the sake of consistency.

We first split the entire dataset (excluding the part with new annotations) into three parts based on sentence origin: PDB, NKJP1M or PCC. Next, each of the three parts was processed separately, paying attention to their different characteristics. Sentences originating from PCC (which does not contain manual morphosyntactic annotations) were re-parsed with UDPipe using the latest Polish model.[18] For the NKJP1M part, with the manually annotated morphological layer, we first performed a morphological tagset conversion using conversion tables specifically (semi-automatically) compiled for the task. This was necessary because the morphological layer of NKJP1M uses a different tagset than the remaining, UD-compliant part of the dataset. After that, we used UDPipe to re-parse the NKJP1M part at the syntactic level only (dependencies are not manually annotated in NKJP1M). Finally, in PDB, all annotations result from the conversion of manual annotations in Składnica (see Sec. 3). Hence, for this part of the dataset it was only necessary to update the morphosyntactic layer of the corpus with respect to the latest version of PDB.

4.3 Companion "raw" corpus

Together with the main corpus, manually annotated for VMWEs, we prepared a large (159,115,022 sentences, 1,902,279,431 tokens) UD-2.5-compliant raw corpus automatically annotated for morphosyntax and dependencies with UDPipe.[19] The raw corpus is released in the CoNLL-U format and does not contain any VMWE annotations. It is meant to facilitate automatic discovery of unseen VMWEs, i.e. VMWEs with no occurrences in the (training) corpus. Unseen VMWEs are known to be hard to capture with purely supervised methods, due to their Zipfian distribution and the particular nature of their idiosyncrasies, which show at the level of types (sets of occurrences) rather then tokens (single occurrences) (Savary et al., 2019). Edition 1.2 of the PARSEME shared task brought unseen VMWEs into focus and raw corpora, accompanying manually annotated corpora, were released for all participating

[14] https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.2/

[15] https://github.com/proycon/flat

[16] http://www.nkjp.pl/poliqarp/

[17] Wikisłownik (https://pl.wiktionary.org/, Słownik PWN (https://sjp.pwn.pl/), and Wielki Slownik Języka Polskiego (https://wsjp.pl/index.php)

[18] polish-pdb-ud-2.5-191206

[19] Using the same model as for the automatically tagged parts of the manually annotated corpus.

languages. The Polish raw corpus is based on the CoNLL 2017 shared task raw corpus[20] (Zeman et al., 2017), which we upgraded to UD-2.5 for the sake of compatibility with the main corpus.

5 Results

The resulting UD-2.5-compatible corpus, manually annotated for VMWEs, comprises 23,547 sentences, 396,140 tokens, and 7,186 manually annotated VMWEs in total. 12,187 sentences originate from PDB-UD, 9,241 from NKJP1M and 2,119 from PCC. Morphological annotation is manually performed in the first two sources[21] and automatically in the third one. Syntactic annotation is manual only in PDB. 7,426 new sentences from PDB-UD were added in edition 1.2.

While the corpus covers a rather broad spectrum of different genres (cf. Sec.4), a large majority (over 68% sentences) are newspaper texts. Double annotation performed over 900 newspaper sentences, new in edition 1.2, resulted in inter-annotator agreement (IAA) scores of $F_{\text{span}} = 77.4\%$ (F-measure between annotators), $\kappa_{\text{span}} = 73.2\%$ (agreement on the annotation span) and $\kappa_{\text{cat}} = 90.7\%$ (agreement on the VMWE category).[22] See (Savary et al., 2017) for the definitions of these three IAA measures.

Table. 1 presents the statistics of the corpus concerning the different VMWE categories as well as the fine-grained VMWE phenomena – discontinuity, one-token length and overlapping – as defined by Savary et al. (2018) – in comparison with version 1.1 of the corpus. The number of overlapping VMWE tokens decreased since version 1.1 most likely due to the removal of IAVs (annotated experimentally in version 1.1), which often co-occur with other VMWEs. Figure 2 (a) illustrates the variability of the different categories of VMWEs in the Polish corpus. We follow the PARSEME-based definition of a variant: it is a sequence of words starting from the first VMWE component and ending on the last VMWE component, including the non-lexicalized words in between. The linear regression models fitted to the numbers of different variants of various categories suggest that LVC.cause and LVC.full VMWEs are the most variable, followed by VIDs, which in turn are more variable than IRVs. Figure 2 (b) on the other hand shows the variability of VMWEs in the Polish corpus in general, in contrast with several other PARSEME 1.2 corpora. It shows that, even though morphologically rich and with relatively free word order, VMWEs in Polish are not as variable as those in Chinese or Turkish, and have a similar level of variability as VMWEs in German or Basque. Interestingly, the variant-of-traindev F-scores[23] achieved by the two best systems, in both the open and the closed track of the PARSEME shared task 1.2, are higher for Polish than for any other language. However, it can be stipulated that the variability captured by the PARSEME definition is influenced by non-related factors such as the average length of the (non-lexicalized) gap,[24] which is in particular significantly higher in the German (average gap length 2.06) than in the Polish corpus (average gap length 0.55).[25]

6 Findings from the manual annotation

This section describes selected interesting phenomena, challenging cases, as well as findings and lessons learned from the manual annotation, across all 3 versions of the Polish PARSEME corpus.

6.1 Interactions with tokens, lemmas, morphology and syntax

The PARSEME definitions and annotation methodology heavily rely on the underlying morphosyntactic annotation (Sec. 2), inherited from the source corpora or from tools, most often trained on UD treebanks.

[20] http://hdl.handle.net/11234/1-1989

[21] PDB-UD has priority over NKJP regarding sentences which belong to the overlap of the two corpora.

[22] All three scores improved in comparison with edition 1.1 of the corpus, where a similar IAA estimation based on 2079 sentences resulted in $F_{\text{span}} = 61.9\%$, $\kappa_{\text{span}} = 56.8\%$ and $\kappa_{\text{cat}} = 88.2\%$.

[23] According to edition 1.2 of the PARSEME shared task, the variant-of-traindev evaluation metrics is the MWE-based F-measure calculated only on those VMWEs which occur in the test corpus and: (i) are seen, i.e. their multisets of lemmas occur, as annotated VMWEs, in the training or in the development corpus, (ii) are not identical to their training/development occurrences, when the strings between the first and the last lexicalized component (including the non-lexicalized elements in between) are compared.

[24] Gap length is defined as the number of non-lexicalized elements in a VMWE's variant (Savary et al., 2018).

[25] Note also that the ratio of discontinuous VMWEs (with a gap) is higher in DE (42.74%) than in PL (28.68%).

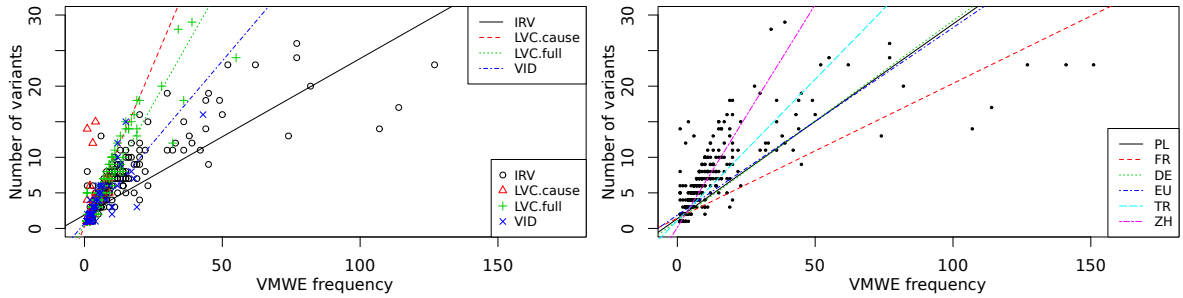

Figure 2: (a) Number of different variants per VMWE frequency for various VMWE categories in Polish, together with the corresponding linear regression fit. (b) Number of variants per VMWE frequency (only displayed for Polish) together with the corresponding linear regression fit for different languages.

The impact of these pre-existing choices on the VMWE annotation is seen in Polish in at least three cases.

Firstly, since a VMWE, by definition, contains at least two words, we have to conform to the definition of a word stemming from the pre-existing corpora. This imposes a careful annotation of some multiword tokens (MWTs). In Polish, contracting two words into one token is very productive in past tense verbal forms like *widziałem* 'I saw' or *widzieliśmy* 'we saw'.[26] According to the so-called flexemic tagset (Przepiórkowski and Woliński, 2003), such forms are regular combinations of a past participle form common for all persons of the same number and gender (*widział*.PRAET:SG:M1, *widzieli*.PRAET:PL:M1) and of a 'floating' form of the auxiliary 'to be' specific for the given person and number (*em*.SG:PRI, *śmy*.SG:PRI).[27] Therefore, while annotating a VMWE like ***na własne oczy widziałem*** (lit. 'on own eyes I-saw') 'I saw sth for myself', we should not include the auxiliary *em* since the same VMWE can appear without it, as in ***na własne oczy widział*** (lit. 'on own eyes he-saw') 'he saw sth with his own eyes'.[28]

The UD tagset does not fully standardize the annotation of some verb forms, like gerunds and participles, which share properties of nouns and adjectives. For instance, Polish gerunds stem from verbs by regular inflection but they behave like nouns (e.g. they inflect for number and case, and have gender). Therefore, in the Polish UD corpora, a gerund like *rzucanie* 'throwing.SG:NOM:N' is tagged as NOUN but receives a verbal lemma, here *rzucać* 'throw'.[29] This means that many Polish VMWEs contain no word tagged as VERB.[30] It should, therefore, be kept in mind that the guidelines apply to the canonical form instead of the actual occurrence of a VMWE candidate. Without the canonical form, examples such as ***rzucanie czarów*** 'casting spells' could not be considered headed by a verb.

	Nb. of categories				Fine-grained phenomena		
	VID	IRV	LVC full	LVC cause	Discontinuous	Single-token	Overlapping
1.2	826	3,629	2,420	311	28.68	0.0	0.87
1.1	487	2,275	1,837	246	29.76	0.0	2.92

Table 1: Statistics of the Polish corpus in version 1.2 in comparison with version 1.1.

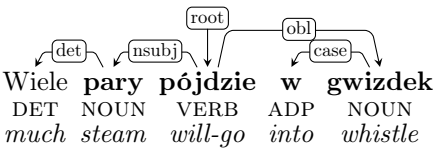

Figure 3: A VMWE with a numeral phrase.

Another phenomenon related to canonical forms shows the usefulness of the UD annotation scheme for the PARSEME methodology. A major UD principle is that dependencies hold between content words, and the latter head function words. This approach has received criticism (Osborne and Gerdes, 2019), and in Polish there is, indeed, strong evidence that many function words, such as numerals and determiners,

[26] A similar analysis concerns conditional forms of verbs.

[27] The Polish-specific morphological tags stemming from the NKJP corpus are documented at `http://nkjp.pl/poliqarp/help/ense2.html`.

[28] FLAT shows both the contracted and the split versions of MWTs in the annotation interface, and only the split version should be used.

[29] Similarly, all present and past participles, like *rzucająca* 'throwing.SG:NOM:F' and *rzucane* 'thrown.SG:NOM:N', receive the UPOS value of ADJ but their lemma is a verb, here *rzucać* 'throw'.

[30] The morphological features `VerbForm=Part` and `VerbForm=Vnoun` do indicate the verbal stem.

determine the grammatical forms of content verbs (Przepiórkowski and Patejuk, 2020). However, the UD assumption helps keep the PARSEME definition of a VMWE (cf. Sec. 2) relatively simple. Consider the example and its UD-style dependency tree in Fig. 3, meaning 'one's efforts will bring no result'. According to Savary et al. (2018), this form is canonical (the head verb occurs in a finite non-negated form and all its arguments are in singular and realized with no extraction).[31] Note that the numeral *wiele* 'a-lot-of' is not lexicalized. In Polish formal linguistics, e.g. in the HPSG framework (Przepiórkowski, 1999), *wiele pary* 'much steam' is seen as a numeral phrase headed by the indefinite numeral *wiele* 'much'. If this principle were not overridden by the content word primacy, the dependency arc between *wiele* 'much' and *pary* 'steam' would be inverted and the lexicalized components of this VMWE would be disconnected, conversely to PARSEME's definition of a VMWE.

6.2 IRV-specific phenomena

IRVs are, by far, the most frequent VMWE category in Polish (cf. Sec. 5). Hard cases include those verbs which are much more frequent with than without a RCLI. For instance, *delektować* 'to-delight' is found 563 times by a NKJP concordancer with a RCLI, as in ***delektować się*** *piwkiem* (lit. 'delight RCLI beer.INST') 'enjoy a beer', and only 3 times without it, as in *delektuje nas znakomitymi zdjęciami* 'delights us with great photos'. The latter use can easily be missed by the annotator, who then concludes that the verb never occurs without the RCLI (test IRV.1), i.e. it is an IRV, although the former use is simply a reflexive variant of the latter (IRV.6).

Another specificity of Polish (and Czech), is the so-called haplology of the RCLI (Kupść, 1999; Rosen, 2014): a single occurrence of RCLI can satisfy several requirements for this item. For instance, in the sentence from Fig. 1 two IRVs co-occur: ***postarać się*** (lit. 'try RCLI') 'try hard' and ***pogodzić się*** (lit. 'reconcile RCLI') 'reconcile' and share the RCLI.[32]

We also found that many Polish simple verbs can be simultaneously preceded by the prefix *na-* and accompanied by the RCLI, to express the fact that the given action has been performed frequently or for a long time, as in *czytał* 'he-read' → ***naczytał się*** 'he-has-read-a-lot', *siedziała* 'she-sat' → ***nasiedziała się*** 'she-has-sat-a-lot', *zamiatali* 'they-swept' → ***nazamiatali się*** 'they-have-swept-a-lot', etc. This phenomenon is productive, and should, intuitively, not be considered idiomatic. However, all the above examples have to be annotated as IRVs, according to the PARSEME guidelines (due to test IRV.3).

Let us also mention that the RCLI in Polish (and other Slavic languages) does not inflect for person and number, as in ***boję się*** 'I am afraid', ***boicie się*** 'you are afraid'.[33] However, it does inflect for case. Even if its accusative form *się* is predominant, the IRVs with its dative form *sobie* should not be omitted, e.g. ***wyobrazić sobie*** (lit. 'imagine RCLI') 'imagine', ***poradzić sobie*** (lit. 'advise RCLI') 'cope'.

6.3 LVC-specific phenomena

LVCs are the second most frequent VMWE type in Polish. A major challenge was to distinguish LVC.full and LVC.cause when the cause belongs to the semantic arguments of the noun. In example (1), *stwarzać* 'create' is a typical causative verb. It also occurs in several LVC.cause expressions, e.g. ***stwarzać okazję/szansę/warunki*** 'to create an occasion/chance/conditions'. Here, however, the predicative noun *zagrożenie* 'danger' requires an agent/cause, i.e. *produkty* 'products' belong to its semantic arguments. Since the test for being a semantic argument of the noun (LVC.2) is placed earlier in the decision flowchart then the one for being its cause (LVC.5), this expression has to be tagged as an LVC.full.

(1) Produkty te **stwarzają zagrożenie** dla zdrowia konsumenta.
products these create danger for health consumers.GEN
'These products constitute a danger for the health of the consumers.'

[31] One might argue that a form omitting the determiner *wiele* 'much' is canonical instead. Recall, however, that a canonical form is to be constructed in context, while keeping the meaning of the whole expression possibly unchanged. Omitting the determiner would contradict this principle.

[32] Repeating RCLI would be ungrammatical here: **postaraj się z tym pogodzić się*. The annotators have to be careful with such cases, so as not to miss the overlapping annotation.

[33] This is in contrast e.g. with Romance languages, where the RCLI agrees for person and number with the subject of the verb, as in (FR) *je* ***me trouve*** 'I find myself', *vous* ***vous trouvez*** 'you find yourself', etc.

(2) **umożliwili** mi **preprowadzenie badań**
they-allowed me carrying-out researches
'They allowed me to carry out research.'

Another interesting, even if quantitatively minor, question is how to annotate LVCs in which the direct object of the verb is itself a light verb. In example (2), ***przeprowadzenie badań*** 'carrying-out research' is clearly an LVC.full. The other verb *umożliwili* 'allowed', has a causative meaning but one may hesitate as to the choice of its predicative noun. One natural candidate is the syntactic object *przeprowadzenie* 'carrying-out'. Since, however, it is a nominalisation of a light verb *przeprowadzić* 'carry-out', it is dubious to establish its semantic arguments (needed in tests LVC.2 and LVC.5). Another choice would be to consider ***umożliwić badania*** 'allow research' as an LVC.full but the structural tests (S.1 to S.4) require the predicative noun to be a dependent of the verb. The problem lies, truly, in not knowing how to establish the canonical form of such nested LVCs. The nominalisation needs to be converted to a finite form, e.g. ***przeprowadziłam badania*** 'I-carried-out research'. But then, the finite verb *przeprowadziłam* 'I-carried-out' can no longer be the object of *umożliwili* 'allowed'. One solution is not to annotate *umożliwili* 'allowed' at all. Another one would consist in a more elaborated definition of a canonical form, so as to yield strong reformulations, e.g. ***przeprowadziłam badania****, oni* ***umożliwili*** *te* ***badania***. 'I-carried-out research, they allowed my research.'

6.4 VMWEs and peripheral phenomena

As discussed by Savary et al. (2018), the VMWE-ness has fuzzy borders with related phenomena, and we encountered them while annotating Polish texts. Firstly, VMWE are often hard to discriminate from collocations, defined by PARSEME as word combinations whose idiomaticity is of statistical nature only. Thus, word combinations like *stawiać stopnie* (lit. 'put grades') 'to-grade' or *zapuścić wąsy* 'grow a mustache', look idiomatic because the mutual lexical selection between both components is statistically strong (i.e. test VID.2 based on component replacement seems likely to be passed). Corpus searches often help to invalidate this hypothesis but doubts remain if: (i) the verb selects only a small class of nouns (*zapuścić wąsy, brodę, włosy, paznokcie* 'grow a mustache, beard, hair, nails'), (ii) it has several close senses[34] (iii) the variants stemming from lexical replacement are infrequent in corpora.

Metaphor is another challenging peripheral phenomenon, because most VMWEs are lexicalized metaphors. It seems, therefore, that the only difference between the two is the degree of lexicalization, which is however hard to establish, even with corpus studies, for the same reasons as with collocations. Particularly testing are those metaphors which are collocations at the same time. For instance, *pękać ze śmiechu* 'burst with laughter' is a frequent metaphor in NKJP. Luckily, some rare examples do reveal that *pękać* 'burst' can be used metaphorically with many emotions (*z dumy/bólu/przemęczenia/migreny/etc.* 'with pride/pain/fatigue/fatigue/etc.'). Other examples of metaphors judged as non-VMWE include: *nabrzmiewać ironią* (lit. 'swell with irony'), *omiatać (horyzont) spojrzeniem* (lit. 'sweet (the horizon) with a glance'), *znaleźć kij na prawicę* (lit. 'find a stick against the right wing'), etc.

Finally, MWEs are particular cases of grammatical constructions, i.e. conventional associations of lexical, syntactic and pragmatic features, such as *the-Adj-the-Adj* (*the more the merrier*, *the higher the better*, etc.). In the corpus we encountered examples of Polish constructions which are no VMWEs but contain non-lexicalized verbs, e.g. *mało nie V*, as in *mało nie zwariował* (lit. 'little not went-crazy') 'he almost went crazy', *V.*INF *V*, as in *rozumieć rozumiem* (lit. 'understand.INF I-understand') 'I do understand', or *nie sposób V.*INF, as in *nie sposób zapomnieć* (lit. 'not way to-forget') 'one cannot forget'.

Attending constructions led us to detecting a minor flaw in the IRV tests. Namely, examples like *bać się* (lit. 'fear RCLI') 'be-afraid' are tagged as IRVs because the verb can never appear without the RCLI (test IRV.1). There are, however, some constructions which contain a slot for any IRV, and a duplication of its verb alone, without the RCLI. Examples include: *V RCLI, oj V*, as in ***działo się****, oj działo* (lit. 'happened RCLI, oh happened') 'there was really a lot going on' and *V.*INF *RCLI nie V*, as in ***bać się*** *nie bał* (lit. 'to-fear RCLI not he-feared') 'as to being afraid, he was not'. These constructions

[34] *Zapuścić korzenie* 'take root' might be an instance of the same or a different sense than *zapuścić wąsy* 'grow a mustache', which is or is not an evidence of lexical flexibility, respectively.

are productive and omitting the RCLI is clearly licensed by the duplication. Therefore, they should not be considered counterexamples in the IRV decision process.

7 Towards enhanced PARSEME guidelines

Several enhancements in the PARSEME annotation guidelines can be proposed based on our experience.

Firstly, nesting of VMWEs should be more accurately accounted for. Currently, the verb in an LVC is allowed to only have one lexicalized dependent (test S.2), which excludes inherently reflexive light verbs, as in ***nosić się z zamiarem*** (lit. 'carry RCLI with intention') 'to have an intention'. Such examples can only be annotated as VIDs, although they function like LVCs. We might therefore allow for more than one lexicalized dependent of the verb in test S.2, provided that all but one of them belong to a previously annotated VMWE. This would also allow verb-particle constructions (VPCs)[35] to be nested in IRVs, as in (DE) *er [[**stellt**]$_{VPC}$ **sich** [**vor**]$_{VPC}$]$_{IRV}$* (lit. 'he puts RCLI forward') 'he imagines' (now such cases are formally VIDs).

Secondly, the reciprocal uses of the RCLI listed in test IRV.8 do not accurately cover Slavic languages. The test checks if a plural or coordinated subject can be distributed over two occurrences of the same verb. For instance *Jan i Ela się całują* (lit. 'Jan and Ela RCLI kiss') 'Jan and Ela kiss each other' can be transformed into *Jan całuje Elę, a Ela całuje Jana* 'Jan kisses Ela and Ela kisses Jan'. Therefore, *całować się* (lit. 'kiss RCLI') is a reciprocal use of *się* and not an IRV. But in Polish, there is another reciprocal form with a singular subject and an oblique: *Jan całuje się z Elą* 'Jan kisses RCLI with Ela'. Adding this case to IRV.8 is necessary, at least for language-specific variants of this test. But this is not sufficient since the verb alone does not admit the same subcategorization: **Jan całuje Elę z Eleną* (lit. 'Jan kisses Ela with Elena'). Thus, test IRV.3 is always passed, and such cases have to be annotated as IRVs, although they are productive. A possible solution would be to change the order of the IRV tests so that those checking non-idiomatic uses of the RCLI (currently IRV.4 to IRV.8) are placed first.

Thirdly, specific constructions with duplicated verbs invalidate some genuine IRVs (cf. Sec. 6.4). Language-specific lists of such constructions, to be neglected by test IRV.1, could be proposed.

Finally, an open problem is how to ensure that the decision diagrams always yield the same outcome for the same sense of a verb, whatever its non-lexicalized arguments. In ***stawiam*** *sobie/komuś* ***cel*** (lit. 'I-put myself/someone a goal') 'set a goal to myself/someone', the outcome of test LVC.2 depends on the indirect object. With a reflexive object *sobie* 'myself', the subject of the verb (I) is the agent/beneficiary of the noun *cel* 'aim', which suggests the LVC.full status. But with another object, the verb's subject does not fill any semantic role of the noun, which leads to LVC.cause. We would of course like both of these uses to be annotated in the same way, here as LVC.cause. But this would imply applying the test to all possible instances of the (non lexicalized) object, rather than to the precise example being annotated. With such a major difference in the annotation strategy, the decision replicability might be jeopardized.[36]

8 Conclusions

We described the construction of the Polish corpus of VMWEs, which is the 4th biggest dataset in the PARSEME suite. We presented some details of the annotation process and its outcomes. We also discussed some Polish-specific phenomena, interpreted in the light of the PARSEME annotation guidelines. We displayed several drawbacks of these guidelines and put forward suggestion for their enhancements. We believe that these observations can help continuous enhancement of the PARSEME methodology, and can be useful to annotators of other languages, linguists studying the MWE phenomenon, as well as authors of VMWE identification tools.

Acknowledgements

This work was partly funded by the French PARSEME-FR grant (ANR-14-CERA-0001). We are grateful to the anonymous reviewers for their useful comments.

[35] VPC is a PARSEME VMWE category, pervasive notably in Germanic languages but non-existent in Polish.

[36] A similar issue, recently raised on the PARSEME discussion forum, concerns IRV tests for supposedly middle passive uses of the RCLI, like *ograniczać się* (lit. 'limit itself') 'be limited to'. See guidelines Gitlab issue #98 for details.

References

Monika Czerepowicka and Agata Savary. 2018. SEJF - A Grammatical Lexicon of Polish Multiword Expressions. In Zygmunt Vetulani, Joseph Mariani, and Marek Kubis, editors, *Human Language Technology. Challenges for Computer Science and Linguistics*, pages 59–73, Cham. Springer International Publishing.

Katarzyna Głowińska and Adam Przepiórkowski. 2010. The design of syntactic annotation levels in the National Corpus of Polish. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC 2010*, Valletta, Malta. European Language Resources Association (ELRA).

Katarzyna Głowińska. 2012. Anotacja składniowa. In Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk, editors, *Narodowy Korpus Języka Polskiego*, pages 107–127. Wydawnictwo Naukowe PWN, Warsaw.

Elżbieta Hajnicz, Anna Andrzejczuk, and Tomasz Bartosiak. 2016. Semantic layer of the valence dictionary of Polish *Walenty*. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation, LREC 2016*, pages 2625–2632, Portorož, Slovenia. European Language Resources Association (ELRA), European Language Resources Association (ELRA).

Anna Kupść. 1999. Hapology of the polish reflexive marker. In Robert D. Borsley and Adam Przepiórkowski, editors, *Slavic in Head-Driven Phrase Structure Grammar*. CSLI Publications, Stanford, CA.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajic, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France, May. European Language Resources Association.

Maciej Ogrodniczuk, Katarzyna Głowińska, Mateusz Kopeć, Agata Savary, and Magdalena Zawisławska. 2015. *Coreference in Polish: Annotation, Resolution and Evaluation*. Walter De Gruyter.

Timothy Osborne and Kim Gerdes. 2019. The status of function words in dependency grammar: A critique of Universal Dependencies (UD). *Glossa: a journal of general linguistics*, January.

Agnieszka Patejuk and Adam Przepiórkowski. 2014. Synergistic development of grammatical resources: A valence dictionary, an LFG grammar, and an LFG structure bank for Polish. In Verena Henrich, Erhard Hinrichs, Daniël de Kok, Petya Osenova, and Adam Przepiórkowski, editors, *Proceedings of the Thirteenth International Workshop on Treebanks and Linguistic Theories (TLT 13)*, pages 113–126, Tübingen. Department of Linguistics (SfS), University of Tübingen.

Adam Przepiórkowski and Agnieszka Patejuk. 2020. From Lexical Functional Grammar to enhanced Universal Dependencies: The UD-LFG treebank of Polish. *Language Resources and Evaluation*, 54:185–221.

Adam Przepiórkowski and Marcin Woliński. 2003. A flexemic tagset for Polish. In *Proceedings of* Morphological Processing of Slavic Languages*, EACL 2003*.

Adam Przepiórkowski, Elżbieta Hajnicz, Agnieszka Patejuk, and Marcin Woliński. 2014. 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)*, pages 83–91, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Adam Przepiórkowski. 1999. *Case Assignment and the Complement-Adjunct Dichotomy: A Non-Configurational Constraint-Based Approach*. Ph.D. dissertation, Universität Tübingen.

Carlos Ramisch, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Behrang QasemiZadeh, Renata Ramisch, Nathan Schneider, Ivelina Stoyanova, Ashwini Vaidya, and Abigail Walsh. 2018. Edition 1.1 of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222–240. Association for Computational Linguistics.

Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archna Bhatia, Uxoa Iñurrieta, Voula Giouli, Tunga Güngör, Menghan Jiang, Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh, and Hongzhi Xu. 2020. Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons (MWE-LEX 2020)*. Association for Computational Linguistics.

Alexandr Rosen. 2014. Haplology of reflexive clitics in Czech. In Elżbieta Kaczmarska and Motoki Nomachi, editors, *Slavic and German in Contact: Studies from Areal and Contrastive Linguistics*, pages 97–116. Slavic Research Center, Hokkaido University, Sapporo, Japan.

Agata Savary and Jakub Waszczuk. 2017. Projecting multiword expression resources on a Polish treebank. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, pages 20–26, Valencia, Spain, April. Association for Computational Linguistics.

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and Antoine Doucet. 2017. The PARSEME Shared Task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions* , MWE '17, pages 31–47. Association for Computational Linguistics.

Agata Savary, Marie Candito, Verginica Barbu Mititelu, Eduard Bejček, Fabienne Cap, Slavomír Čéplö, Silvio Ricardo Cordeiro, Gülşen Eryiğit, Voula Giouli, Maarten van Gompel, Yaakov HaCohen-Kerner, Jolanta Kovalevskaitė, Simon Krek, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Lonneke van der Plas, Behrang QasemiZadeh, Carlos Ramisch, Federico Sangati, Ivelina Stoyanova, and Veronika Vincze. 2018. PARSEME multilingual corpus of verbal multiword expressions. In Stella Markantonatou, Carlos Ramisch, Agata Savary, and Veronika Vincze, editors, *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, pages 87–147. Language Science Press., Berlin.

Agata Savary, Silvio Cordeiro, and Carlos Ramisch. 2019. Without lexicons, multiword expression identification will never fly: A position statement. In *MWE-WN 2019*, pages 79–91, Florence, Italy. ACL.

Marcin Woliński, Elżbieta Hajnicz, and Tomasz Bartosiak. 2018. A new version of the Składnica treebank of Polish harmonised with the Walenty valency dictionary. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1839–1844, Paris, France. European Language Resources Association (ELRA).

Alina Wróblewska. 2012. Polish Dependency Bank. *Linguistic Issues in Language Technology*, 7(1).

Alina Wróblewska. 2020. Towards the Conversion of National Corpus of Polish to Universal Dependencies. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5308–5315, Marseille, France. European Language Resources Association (ELRA).

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada, August. Association for Computational Linguistics.

Marek Świdziński and Marcin Woliński. 2010. Towards a bank of constituent parse trees for Polish. In Petr Sojka, Aleš Horák, Ivan Kopeček, and Karel Pala, editors, *Text, Speech and Dialogue: 13th International Conference, TSD 2010, Brno, Czech Republic*, number 6231 in Lecture Notes in Artificial Intelligence, pages 197–204, Heidelberg. Springer-Verlag.