Extended and Enhanced Polish Dependency Bank in Universal Dependencies Format

Alina Wróblewska
Institute of Computer Science
Polish Academy of Sciences
ul. Jana Kazimierza 5
01-248 Warsaw, Poland
alina@ipipan.waw.pl

Abstract

The paper presents the largest Polish Dependency Bank in Universal Dependencies format – PDBUD – with 22K trees and 352K tokens. PDBUD builds on its previous version, i.e. the Polish UD treebank (PL-SZ), and contains all 8K PL-SZ trees. The PL-SZ trees are checked and possibly corrected in the current edition of PDBUD. Further 14K trees are automatically converted from a new version of Polish Dependency Bank. The PDBUD trees are expanded with the enhanced edges encoding the shared dependents and the shared governors of the coordinated conjuncts and with the semantic roles of some dependents. The conducted evaluation experiments show that PDBUD is large enough for training a high-quality graph-based dependency parser for Polish.

1 Introduction

Natural language processing (NLP) is nowadays dominated by machine learning methods, especially deep learning methods. Data-driven NLP tools not only perform more accurately than rule-based tools, but are also easier to develop. The shift towards machine learning methods is also visible in syntactic parsing, especially dependency parsing. The vast majority of the contemporary dependency parsing systems (e.g. Nivre et al., 2006; Bohnet, 2010; Dozat et al., 2017; Straka and Straková, 2017) take advantage of machine learning methods. Based on training data, parsers learn to analyse sentences and to predict the most appropriate dependency structures of these sentences. Even if various learning methods were applied to data-driven dependency parsing (e.g. Jiang et al., 2016), the best results so far are given by the supervised methods (cf. Zeman et al., 2017). Supervised dependency parsers trained on correctly annotated data achieve high parsing performance even for languages with rich morphology and relatively free word order, such as Polish.

The supervised learning methods require gold-standard training data, whose creation is a time-consuming and expensive process. Nevertheless, dependency treebanks have been created for many languages, in particular within the Universal Dependencies initiative (UD, Nivre et al., 2016). The UD leaders aim at developing a cross-linguistically consistent tree annotation schema and at building a large multilingual collection of dependency treebanks annotated according to this schema.

Polish is also represented in the Universal Dependencies collection. There are two Polish treebanks in UD: the Polish UD treebank (PL-SZ) converted from *Składnica zależnościowa*[1] and the LFG enhanced UD treebank (PL-LFG) converted from a corpus of the Polish LFG structures.[2] PL-SZ contains more than 8K sentences with 10.1 tokens per sentence on average. PL-LFG is larger and contains more than 17K sentences, but the average number of tokens per sentence is only 7.6.[3]

This paper presents the largest Polish Dependency Bank in Universal Dependencies format – PDBUD[4] – with 22K trees and 352K tokens (hence 15.8 tokens per sentence on average). PDBUD builds on its previous version, i.e. the Polish UD treebank (PL-SZ), and contains all 8K PL-SZ trees. The PL-SZ trees are checked and possibly corrected in the current edition of

[1] *Składnica zależnościowa* was converted to the UD format by Zeman et al. (2014).

[2] LFG structures were converted by A. Przepiórkowski and A. Patejuk.

[3] A detailed comparison of PL-SZ and PL-LFG is presented on http://universaldependencies.org/treebanks/pl-comparison.html.

[4] PDBUD is publicly available on http://zil.ipipan.waw.pl/PDB.

PDBUD. Further 14K trees are automatically converted from a new version of Polish Dependency Bank (PDB, see Section 2). Polish sentences underlying the additional PDB trees contain problematic linguistic phenomena whose conversion requires some modifications of the UD annotation schema (see Section 3). Furthermore, the PDBUD trees are expanded with the enhanced edges encoding the shared dependents and the shared governors of the coordinated conjuncts (see Section 4) and with the semantic roles of some dependents (see Section 5). Finally, we conduct some evaluation experiments. The evaluation results show that PDBUD is large enough for training a high-quality graph-based dependency parser for Polish (see Section 6).

2 Polish Dependency Bank

2.1 PDB

The first Polish dependency treebank – *Składnica zależnościowa* (Wróblewska, 2012) – was a collection of about 8K trees which were automatically converted from Polish constituent trees of *Składnica frazowa* (Woliński et al., 2011). All sentences of *Składnica* were derived from Polish National Corpus (Przepiórkowski et al., 2012). The annotated sentences are rather short with 10.2 tokens per sentence on average and corresponding trees are relatively simple (there is only 289 non-projective trees,[5] i.e. 3.5% of all trees).

This first version of Polish dependency treebank was enlarged with 4K trees (Wróblewska, 2014). The additional trees resulted from the projection of English dependency structures on Polish parallel sentences from *Europarl* (Koehn, 2005), *DGT-Translation Memory* (Steinberger et al., 2012), OPUS (Tiedemann, 2012) and *Pelcra Parallel Corpus* (Pęzik et al., 2011). The additional sentences with the average length of 15.9 tokens per sentence were longer than the sentences from *Składnica*. The projection-based trees were also more complex and 235 of them are non-projective (i.e. 5.9% of all added trees). The entire set of *Składnica* trees and the projection-based trees is called Polish Dependency Bank (PDB).

PDB is still being developed at the Institute of Computer Science PAS. The current version of PDB is enlarged with a suite of 10K sentences annotated with the dependency trees. The additional sentences are relatively complex (20.5 tokens per sentence on average) and come from Polish National Corpus (Przepiórkowski et al., 2012), Polish CDSCorpus[6] (Wróblewska and Krasnowska-Kieraś, 2017), and literature. There are 1388 non-projective trees in this set (i.e. 13.9% of 10K trees). Besides enlarging PDB, the development consists in correcting the previous PDB trees. The *Składnica* trees and the projection-based trees are manually checked and corrected if necessary.

The current version of PDB consists of more than 22K trees with 15.8 tokens per sentence on average (see Table 1). There are 1912 non-projective trees in PDB (i.e. 8.61% of all trees).

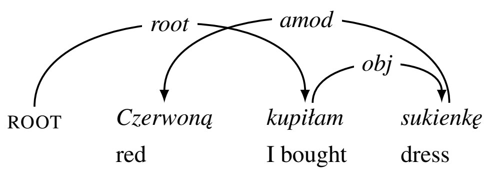
	PDB	PDBUD
# sentences	22,208	
# tokens	351,715	
# tokens per sentence	15.84	
# dependency types	28	31 (48)*
% non-projective edges	1.76	1.75
% non-projective trees	8.61	8.03
% enhanced edges	n/a	4.96
% enhanced graphs	n/a	41.58

Table 1: Statistics of Polish Dependency Bank (PDB) and its UD conversion (PDBUD). *There are 31 universal dependency types in PDBUD and 48 universal types with the Polish-specific subtypes.

2.2 PDBUD

The PDB trees are automatically converted to the UD trees according to the guidelines of Universal Dependencies v2[7] and the resulting set is called PDBUD (i.e. Polish Dependency Bank in Universal Dependencies format). PDBUD contains all trees of the Polish UD treebank (PL-

[5]Non-projective trees contain long distance dependencies resulting in crossing edges. See the topicalisation example *Czerwoną kupiłam sukienkę* 'I bought a red dress' (lit. 'Red I bought a dress') with the following non-projective dependency tree:



[6]http://zil.ipipan.waw.pl/Scwad/CDSCorpus

[7]http://universaldependencies.org/guidelines.html

SZ), which are possibly corrected. The size of PDBUD is exactly the same as the size of PDB, i.e. 22K trees and 351K tokens (see Table 1). 1783 of the PDBUD trees are non-projective, i.e. 8.03% of all trees. There are 17K enhanced edges (4.96% of all edges) in PDBUD and 41.6% of the PDBUD graphs have at least one enhanced edge.

The converted PDBUD trees are largely consistent with the PL-SZ trees. While converting, we try to preserve the universality principle of UD, but some necessary modifications are essential. The PL-SZ trees are rather simple and the sentences underlying this data set do not contain some linguistic phenomena, e.g. ellipsis, comparative constructions, directed speech, interpolations and comments, nominative noun phrases used in the vocative function, and many others. Therefore, the repertoire of the UD relation subtypes and language-specific features is slightly extended in PDBUD to cover these phenomena (see Section 3). Furthermore, in contrast to the PL-SZ trees, the PDBUD graphs contain enhanced edges encoding shared dependents or shared governors of coordinated elements (see Section 4). Finally, some semantic labels are added that goes beyond the standard annotation scheme of Universal Dependencies (see Section 5).

3 Corrections and extensions

Plenty of errors are corrected in the original *Składnica* trees (and the projection-based trees) and thus they are not transferred to these PDBUD trees, which correspond to the PL-SZ trees. The errors in the *Składnica* trees were predominantly caused by the inadequate automatic conversion of the phrase-structure trees into the dependency trees, particularly by the erroneous labelling. Defective part-of-speech tags, morphological features, lemmas, dependency relations and their labels are manually corrected by highly qualified linguists. The correction issues do not fall within the scope of this paper. The conversion issues and extension suggestions are described in the following sections.

3.1 Comparative constructions

Comparative constructions are distinguished in the PDB trees and thus they are also marked in PDBUD. According to Bondaruk (1998), there are two types of comparative constructions in Polish: *comparatives of equality* marked with e.g. *tak ... jak* ('as ... as'), *taki ... jaki* ('just like'), and *comparatives of inequality* marked with NIŻ ('than').[8] All markers introducing comparative constructions, e.g. JAK, NIŻ, JAKBY, NICZYM, are converted as the subordinate conjunctions SCONJ with the feature ConjType=Cmpr.[9] Comparative constructions are annotated with the following dependencies (see Figure 1): the comparative marker is labelled mark and it depends on the main element of the comparative construction labelled obl:cmpr (a new UD subtype).

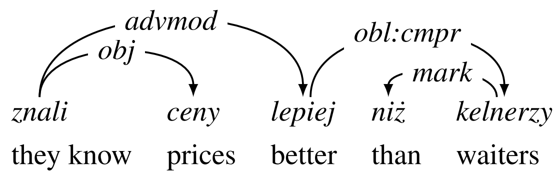


Figure 1: The PDBUD tree of [...] *znali ceny potraw lepiej niż kelnerzy* ('they know the prices of dishes better than the waiters') with the comparative construction.

3.2 Constructions with JAKO

The lexeme JAKO is one of the uninflectable Polish parts of speech. It causes considerable difficulties and is heterogeneously analysed as a preposition, a coordinating conjunction, a subordinating conjunction, or an adverb in the traditional Polish linguistics. According to the concept of the bi-functional subordinating conjunction JAKO (Wróblewska and Wieczorek, 2018), we convert all examples of JAKO as SCONJ with the feature ConjType=Pred (i.e. a predicative conjunction – a new Polish-specific feature). The subordinating conjunction JAKO, which is labelled mark, can be governed by the head of any constituent phrase (e.g. a nominal, prepositional, or verbal phrase) which is, in turn, governed by the sentence predicate subcategorising another phrase of the same type (see Figure 2). There is an identification relation between the sub-

[8]Comparatives of inequality are sometimes introduced by the comparative forms of adjectives or adverbs (marked in PDBUD with the feature Degree=Cmp). However, comparatives of inequality can also be introduced by non-comparative adjectives (e.g. *inny* 'other'), adverbs (e.g. *inaczej* 'in another way', *przeciwnie* 'on the contrary'), or even the verb *woleć* 'to prefer'.

[9]Cmp is the value of Degree in UD and comp stands either for the oblique complement obl:comp in French or for the object of comparison nmod:comp in Uyghur. We therefore decide to introduce a new value Cmpr/cmpr to indicate comparative constructions.

categorised argument and the phrase introduced by JAKO (hence the bi-functional subordinating conjunction) which could be marked with an enhanced edge.

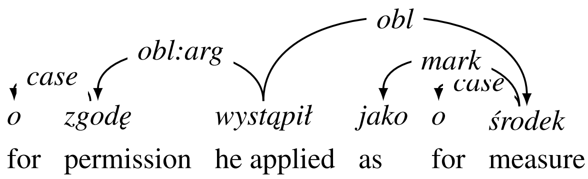


Figure 2: The PDBUD tree of the sentence *O zgodę taką wystąpił jako o środek zapobiegawczy* ('He applied for such permission as a precautionary measure') with JAKO.

3.3 Mobile inflection

The mobile inflections (marked as aglt in the Polish tagset, e.g. *-em* in *odwołałem* 'I$_{\text{Mask}}$ recalled' or *-ś* in *zrobiłabyś* 'you$_{\text{Fem}}$ would do') are the enclitics which substitute auxiliary verbs in the past perfect constructions. We convert them as AUX with Aspect, Number, and Person features, similar to PL-SZ. The repertoire of the morphological features of the mobile inflections is enriched with Clitic=Yes and its Variant – either Long (e.g. *-em* in *odwołałem* 'I$_{\text{Mask}}$ recalled') or Short (e.g. *-m* in *odwołałam* 'I$_{\text{Fem}}$ recalled'). The mobile inflections are marked with the further features VerbForm=Fin and Mood=Ind in the PL-SZ trees, but as they are not the proper finite verbs, these features seem to be incorrect and are not included in PDBUD. A mobile inflection is the special case of an auxiliary verb. Therefore, the relation between the mobile inflection and its governing participle is labelled with a special subtype aux:clitic (a new UD subtype).

3.4 Conditional particle

The conditional particle BY, e.g. *-by-* in *zrobiłabyś* ('you$_{\text{Fem}}$ would do'), is annotated in PL-SZ as an auxiliary AUX with the features Aspect=Imp, Mood=Cnd and VerbForm=Fin, and with the lemma BYĆ ('to be'). It is a particle which doesn't bear any grammatical features in Polish (cf. Przepiórkowski et al., 2012). Since it is not any verb form, it cannot be annotated with Aspect, Mood and VerbForm features which are reserved for verbs. Furthermore, its lemma form is BY and not BYĆ. The conditional particle BY is converted as PART in PDBUD. The relation between this particle and its governor is labelled with aux:cnd (a new UD subtype).

3.5 Other morphosyntactic extensions

We propose some morphosyntactic extensions of the schema which was used to annotate the PL-SZ trees. Some of these extensions are already defined in the UD guidelines, but they were not applied in PL-SZ. Other extensions are newly defined.

ADP There is only one postposition in Polish – TEMU ('ago'), which is converted in PDBUD as the adposition ADP with the feature AdpType=Post. In PL-SZ, the postposition TEMU was wrongly assigned the feature AdpType=Prep, which is reserved for prepositions.

CCONJ We convert the conjunctions PLUS and MINUS as the coordinating conjunction CCONJ with the feature ConjType=Oper (a mathematical operator). There was not any conjunction of this kind in PL-SZ.

Digits Digits (NumForm=Digit) and roman numbers (NumForm=Roman), which are distinguished in PDB, are converted as follows:

- ordinal numbers: the adjectives ADJ with the feature NumType=Ord and other standard features of the adjectives,
- cardinal numbers: the numerals NUM with the feature NumType=Card and other standard features of the numerals,
- other numbers: the tag X.

PUNCT Some features of the punctuation marks are specified:

- PunctSide with the values Initial or Final,
- PunctType with one of the following values: Brck (bracket), Colo (colon), Comm (comma), Dash, Elip (elipsis), Slsh (slash), Blsh (backslash), etc.

Note that Elip, Slsh and Blsh are the newly defined PunctType values.

SYM There are some symbols, e.g. %, §, $, +, ≤, and emojis, e.g. :-), :), in the PDB trees which are converted as the symbols SYM in PDBUD. Emojis are always labelled with the function discourse:emo in PDBUD (a new UD subtype).

VERB The impersonal verb forms[10] are converted as the adjectives ADJ with the feature Case in PL-SZ. In the Polish linguistics however, the impersonals are considered verb forms which cannot be conjugated by the grammatical case. Therefore, we convert them as the verbs VERB with the following features: Aspect (Perfective or Imperfective), Mood=Ind, Person=0, Tense=Past, VerbForm=Fin, and Voice=Act.

X The foreign words are converted as X tags with the feature Foreign=Yes. Abbreviations are also annotated as X tags with the features Abbr=Yes and Pun=Yes if the abbreviation requires a full stop (e.g. *art.* 'article'), or Pun=No if it doesn't (e.g. *cm* 'centimetre').

3.6 Additional relation subtypes

We also propose to extend the inventory of the UD relation subtypes with some additional subtypes listed in the alphabetical order below.[11]

acl:attrib A Polish clause can modify a noun phrase, even if it is not a proper relative clause, e.g. [...] *jest jedynie przejawem* [...] *prawa przyciągania seksualnego: owad nieomylnie trafia do pragnącej zapylenia rośliny.* ('[it] is just a sign of the law of sexual attraction: an insect infallibly goes to a plant that wants to be pollinated.') – the clause *owad nieomylnie trafia* [...] modifies the noun *prawa* ('of the law'). The relation subtype acl:attrib (adverbial clause modifier of a noun)[12] is therefore introduced to cover constructions of this type.

advmod:arg It is possible in Polish that an adverbial is subcategorised by the verb, e.g. *lepiej* ('better') is subcategorised by the infinitive *mieć* ('to have') in *Wiem, że możemy mieć lepiej* ('I know that our situation/conditions will improve', lit. 'I know that we can have better'). The relations between adverbials with the argument status and governing verbs are labelled with the subtype advmod:arg (an adverbial with the argument status) in PDBUD.

advmod:neg The relation between the negation particle NIE ('not') and its governor is labelled with advmod:neg.

aux:imp The relation between the imperative particle NIECH ('let's') and its governor is labelled with aux:imp.

ccomp:obj The PDB direct objects are these verb arguments which are shifted into the grammatical subjects in the passive sentences. Not only noun objects but also clausal objects undergo this shift, e.g. *Przewidział, że inflacja będzie spadać* ('He predicted that inflation would go down') and its passive version *Że inflacja będzie spadać zostało przewidziane* ('It was foreseen that inflation would go down', lit. 'That inflation would go down was foreseen'). In order to convert the clausal objects, the subtype ccomp:obj is proposed. It is worth considering whether it is not a better solution to introduce a new UD type cobj in analogy to csubj.

discourse:intj Interjections, e.g. *cześć* ('hello'), *Och* ('Oh'), *Okay*, are labelled with the function discourse:intj.

nmod:arg Noun complements of various parts of speech, except for verbs, are labelled with the function nmod:arg (noun complement), e.g. *środowiska* in *ochrona*$_{\mathsf{NOUN}}$ *środowiska*$_{\mathsf{NOUN}}$[13] ('environmental protection'), *dzieci* in *korytarz pełen*$_{ADJ}$ *dzieci*$_{\mathsf{NOUN}}$ ('a corridor full of children').

nmod:subj Polish allows the grammatical subject realised as a prepositional phrase, e.g. *do*$_{\mathsf{ADP}}$ *2 lat więzienia* in *Grozi mu do 2 lat więzienia* ('He faces up to two years in prison', lit. 'Up to two years in prison threatens him') or an adverbial phrase, e.g. *Rzadko*$_{\mathsf{ADV}}$ in *Rzadko nie znaczy*

[10]Impersonal verb forms are annotated with the tag imps in PDB.

[11]The list of all dependency labels used in PDBUD is as follows (the new dependency labels are underlined): <u>acl:attrib</u>, acl:relcl, advcl, advmod, <u>advmod:arg</u>, <u>advmod:neg</u>, amod, appos, aux, <u>aux:clitic</u> (see Section 3.3), <u>aux:cnd</u> (see Section 3.4), <u>aux:imp</u>, aux:pass, case, cc, cc:preconj, ccomp, <u>ccomp:obj</u>, conj, cop, csubj, det, <u>discourse:emo</u> (see Section 3.5), <u>discourse:intj</u>, expl:impers, fixed, flat, iobj, list, mark, nmod, <u>nmod:arg</u>, <u>nmod:subj</u>, nsubj, nsubj:pass, nummod, obj, obl, <u>obl:agent</u>, <u>obl:arg</u>, <u>obl:cmpr</u> (see Section 3.1), orphan, parataxis, <u>parataxis:insert</u>, <u>parataxis:obj</u>, punct, root, vocative, xcomp.

[12]We considered labelling this relation with the function advcl. However, "an adverbial clause modifier is a clause which modifies a verb or other predicate" (see the UD guidelines `http://universaldependencies.org/u/dep/advcl.html`). Therefore, we decided not to use the label advcl for an adverbial clause modifier of a noun. Alternatively, this relation could be labelled with parataxis.

[13]*Ochrona* ('a protection') is a deverbal noun that is derived from the verb *chronić* ('to protect') subcategorising an object.

wcale (‘It’s rare, nevertheless still occurs’, lit. ‘Rarely does not mean at all’). The relation between a prepositional or adverbial subject and its governing verb is labelled with the subtype nmod:subj. We realise that this subtype is not the best solution. Alternatively, an adverbial subject could be labelled advmod:arg and a prepositional subject could be labelled obl:arg, but then we lose information about their subject function. We also consider introducing two additional subtypes – advmod:subj and obl:subj, but they are extremely confusing.[14]

4 Enhanced graphs

The PDBUD graphs contain the enhanced edges encoding the dependents shared by the conjuncts in coordinate structures (see Figure 3) and the shared governors of the coordinated elements (see Figure 4).

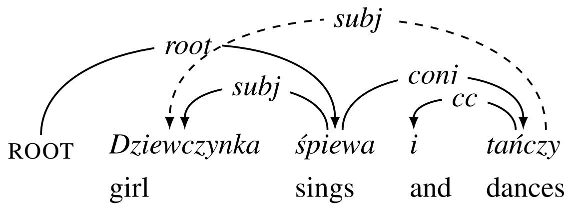


Figure 3: The PDBUD graph of the sentence *Dziewczynka śpiewa i tańczy* (‘A girl sings and dances’) with the shared subject. The enhanced edge is marked with the dashed arrow.

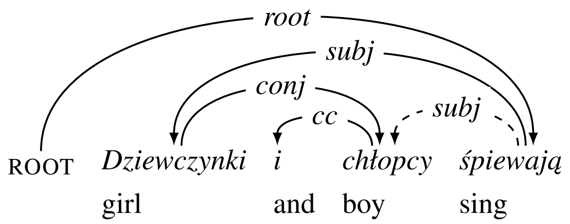


Figure 4: The PDBUD graph of *Dziewczynki i chłopcy śpiewają* (‘Girls and boys are singing’) with the shared governor of the coordinated subject. The enhanced edge is marked with the dashed arrow.

In the PDB trees, all coordinated elements depend on a conjunction and the relations between the conjunction and these elements are labelled with a technical dependency type – conjunct. A dependent shared by all conjuncts also depends on the conjunction, but this relation is labelled with the grammatical function of the shared dependent, e.g. subj, obj. The conversion of the PDB trees into the enhanced PDBUD graphs is thus a straightforward process. There are only enhanced edges involved in the coordination constructions in PDBUD, but they are numerous, i.e. more than 41% of all PDBUD trees contain at least one enhanced edge (see Table 1).

5 Semantic labels

The UD format is extended by adding some semantic labels in the 11th column. There are 28 semantic labels corresponding to some selected *frame elements* of FrameNet (Fillmore and Baker, 2009; Ruppenhofer et al., 2010). In addition to the common semantic roles: THEME, RECIPIENT/BENEFICIARY, RESULT, there are roles related to

- place: SOURCE, GOAL, PLACE, PATH,
- time: TIME, DURATION, STARTING_POINT, END_POINT, FREQUENCY/ITERATION,
- some other roles: ATTITUDE, CAUSE/EXPLANATION/REASON, CIRCUMSTANCES/OTHER, CONCESSIVE, CONDITION, CO-PARTICIPANT, DEGREE, EVENT_DESCRIPTION, INSTRUMENT, MANNER, PURPOSE, REPLACEE, ROLE, STIMULUS, SUPERSET, and TITLE.

The additional semantic labels extend the semantic meaning of indirect objects (iobj), oblique nominals (obl)[15], adverbial clause modifiers (advcl), some adverbial modifiers (advmod), some noun modifiers (nmod), etc.

6 Evaluation

6.1 Dependency parsing systems

Various contemporary dependency parsing systems are tested in our evaluation experiments. All of the tested systems allow dependency parsing, but only some of them allow part-of-speech tagging, morphological analysis and lemmatisation. We test transition-based parsers (i.e. MaltParser, UDPipe, and the transition-based version of BIST

[14] One of the reviewers of the paper suggests to use the label subj. It would be an ideal solution. However, the function subj does not belong to the repertoire of the UD functions.

[15] obl:arg is not semantically specified in PDBUD.

system	architecture	classifier	parsing	tagging	lemmatisation
MaltParser (Nivre et al., 2006)	trans	LR	yes	no	no
MATE parser (Bohnet, 2010)	graph	perceptron	yes	no	no
BIST parser (Kiperwasser and Goldberg, 2016)	trans/graph	biLSTM	yes	no	no
Stanford parser (Dozat et al., 2017)	graph	biLSTM	yes	yes	no
UDPipe (Straka and Straková, 2017)	trans	1-layer NN	yes	yes	yes

Table 2: Properties of the dependency parsing systems tested in our experiments. Explanation: trans – a transition-based parser, graph – a graph-based parser, LR – a linear classifier based on logistic regression, 1-layer NN – a non-linear classifier based on 1-layer neural network, biLSTM – Bidirectional Long-Short Term Memory network.

parser) as well as graph-based parsers (i.e. MATE parser, Stanford parser, and the graph-based version of BIST parser). The properties of the tested dependency parsing systems are summarised in Table 2.

6.2 Data split

PDBUD is divided into three parts – training, test and development data sets. The procedure of assigning dependency trees to particular data sets is generally random, but there is one constraint on the dividing procedure – the *Składnica* trees, and thus also the PL-SZ trees, are not included in the test set.[16] Since sentences underlying the *Składnica* trees are generally shorter than the remaining sentences, the average number of tokens per sentence is significantly higher in the test set than in two other sets. The statistics of the particular data sets is given in Table 3.

	PDBUD		
	train	test	dev
# sentences	17770	2219	2219
# tokens per sentence	15.4	20.2	15.1
# non-projective trees	1310	302	172
% non-projective trees	7.4	13.6	7.7
# enhanced graphs	7147	1181	855
% enhanced graphs	40.2	53.2	38.5

Table 3: Statistics of the training (train), test (test), and development (dev) data sets of PDBUD.

6.3 Evaluation methodology

We apply the evaluation measures defined for the purpose of CoNLL 2018 shared task on Multilingual Parsing from Raw Text to Universal Dependencies.[17] The proposed metrics, i.e. LAS, UAS, CLAS, MLAS, BLEX, evaluate the different prediction aspects.

Two evaluation scenarios are proposed: 1) testing the quality of dependency parsing of Polish, and 2) testing the quality of morphosyntactic prediction of dependency trees, i.e. part-of-speech tagging, lemmatisation, and dependency parsing of Polish. For the purpose of our evaluation, we use the script[18] of CoNLL 2018 shared task.

6.4 Results

6.4.1 Evaluation of dependency parsing

Stanford parser is the best performing parser on Polish data (see Table 4). The second best parser – MATE parser – performs surprisingly well. Even if it doesn't have any neural component, it outperforms not only the graph-based neural parser BIST (87.06 LAS vs. 84.88 LAS), but also all transition-based parsers. It is also worth mentioning that the worst graph-based parser – BIST parser – performs slightly better than its transition-based version, which achieves LAS of 84.79% and is the best of all transition-based parsers. It follows that the graph-based parsers are generally better suited for parsing Polish than the transition-based parsers.

[16] PDBUD is used in the shared task on dependency parsing of Polish – PolEval 2018 (http://poleval.pl). The organisers of this shared task decided not to use the PL-SZ trees, which have been publicly available for some time, for validation of the participating systems. Therefore, the PL-SZ trees are not part of the PDBUD test set.

[17] http://universaldependencies.org/conll18/evaluation.html

[18] http://universaldependencies.org/conll18/conll18_ud_eval.py In order to evaluate the dependency parsers in the first evaluation scenario, the script conll18_ud_eval.py is slightly modified, i.e. some conditions (e.g. single-root property) are disregarded.

System	UAS	LAS
MaltParser	79.73	74.57
BIST transition-based	87.91	84.79
UDPipe	86.23	83.41
MATE parser	89.49	87.06
BIST graph-based	87.97	84.88
Stanford parser	**92.41**	**90.03**

Table 4: Parsers are tested on the sentences with the gold-standard tokens, lemmas, and part-of-speech tags.

System	UAS	LAS	CLAS	MLAS	BLEX
Stanford	**91.33**	**88.04**	**85.48**	n/a	n/a
UDPipe	83.32	78.93	75.22	**64.33**	**71.17**

Table 6: The quality (F1 scores) of predicting unlabelled dependency trees (UAS), labelled dependency trees (LAS), governors and dependency relation types of content words (CLAS), governors, dependency relation types, universal part-of-speech tags and morphological features of content words (MLAS), governors, dependency relation types and lemmas of content words (BLEX).

6.4.2 Evaluation of morphosyntactic prediction of dependency trees

Two systems – Stanford system and UDPipe – are tested in the task of morphosyntactic prediction of dependency trees. These systems predict universal part-of-speech tags (UPOS) as well as language-specific tags (XPOS). Stanford system outperforms UDPipe in part-of-speech tagging (see Table 5). Only UDPipe predicts morphological features (UFEATS) and lemmas (LEMMA).

System	UPOS	XPOS	UFEATS	LEMMA
Stanford	**97.87**	**92.45**	n/a	n/a
UDPipe	96.81	86.05	**88.02**	**95.61**

Table 5: The quality (F1 scores) of predicting universal part-of-speech tags (UPOS), Polish-specific tags (XPOS), morphological features (UFEATS), and lemmas (LEMMA).

Stanford parser significantly outperforms UDPipe in predicting labelled dependency trees (LAS) and in predicting governors and dependency relation types of content words (CLAS), see Table 6. Since Stanford system doesn't predict morphological features and lemmas, we cannot compare MLAS and BLEX scores.

6.4.3 Summary

We carried out two evaluation experiments on PDBUD data. The results of these experiments show that the graph-based parsers, even the parsers without any neural component, are better suited for parsing Polish than the transition-based parsing systems. The best results in parsing Polish data without preceding morphosyntactic analysis are achieved with Stanford parser, i.e. 88.04 LAS. These results are slightly lower than those reported in Dozat et al. (2017), i.e. 90.32 LAS. The possible reason for this is that our test data contains the dependency trees of the longer sentences and thus there is more room for making mistakes. If Stanford parser operates on the PDBUD sentences with the gold-standard part-of-speech tags, it performs better, i.e. 90.03 LAS.

7 Conclusions and future work

We presented PDBUD – the largest Polish dependency bank with 22K dependency trees in Universal Dependencies format. PDBUD contains the corrected trees of the Polish UD treebank (PL-SZ) and 14K dependency trees automatically converted from Polish Dependency Bank. The PDBUD trees are expanded with the enhanced edges encoding the shared dependents and the shared governors of the coordinated conjuncts and with the semantic roles of some dependents. Our evaluation experiments showed that PDBUD is large enough for training a high-quality graph-based dependency parser for Polish.

We did our best to maintain consistency with the UD guidelines while building PDBUD. However, some of our annotation decisions could be arguable and should be discussed again in the context of the universality assumptions of Universal Dependencies.

There is plenty of elliptical constructions in Polish. Some of them are labelled with the function orphan in PDBUD. In our future works, we plan to add empty nodes representing the elided elements to the PDBUD trees. Furthermore, we are going to create a Polish version of Parallel Universal Dependency treebank.

PDBUD data were already used in the shared task on automatic identification of verbal multiword expressions (LAW-MWE-CxG-2018)[19] and are currently used in the shared task on dependency parsing of Polish (PolEval 2018).[20] This is a confirmation of the fact that PDBUD is of very high quality. Therefore, in the future we would like to replace the Polish UD treebank PL-SZ with its corrected, extended and enhanced version – PDBUD.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable comments that we will undoubtedly take into consideration before publishing the final version of our data set.
The research presented in this paper was founded by SONATA 8 grant no 2014/15/D/HS2/03486 from the National Science Centre Poland and by the Polish Ministry of Science and Higher Education as part of the investment in the CLARIN-PL research infrastructure.

References

Bernd Bohnet. 2010. Very High Accuracy and Fast Dependency Parsing is not a Contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING 2010, pages 89–97.

Anna Bondaruk. 1998. *Comparison in English and Polish Adjectives: A Syntactic Study*, volume 6 of *PASE Studies and Monographs*. Folium, Lublin.

Timothy Dozat, Peng Qi, and Christopher D. Manning. 2017. Stanford's Graph-based Neural Dependency Parser at the CoNLL 2017 Shared Task. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 20–30. Association for Computational Linguistics.

Charles J. Fillmore and Collin Baker. 2009. A Frames Approach to Semantic Analysis. In Bernd Heine and Heiko Narrog, editors, *The Oxford Handbook of Linguistic Analysis*, pages 313–340. Oxford University Press.

Yong Jiang, Wenjuan Han, and Kewei Tu. 2016. Unsupervised neural dependency parsing. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 763–771, Austin, Texas. Association for Computational Linguistics.

[19] http://multiword.sourceforge.net/PHITE.php?sitesig=CONF&page=CONF_04_LAW-MWE-CxG_2018___lb__COLING__rb__&subpage=CONF_40_Shared_Task

[20] http://poleval.pl

Eliyahu Kiperwasser and Yoav Goldberg. 2016. Simple and Accurate Dependency Parsing Using Bidirectional LSTM Feature Representations. *Transactions of the Association for Computational Linguistics*, 4:313–327.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the 10th Machine Translation Summit Conference*, pages 79–86.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2006. MaltParser: A Data-Driven Parser-Generator for Dependency Parsing. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, LREC'06, pages 2216–2219.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan T. McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016*, pages 1659–1666.

Piotr Pęzik, Maciej Ogrodniczuk, and Adam Przepiórkowski. 2011. Parallel and spoken corpora in an open repository of Polish language resources. In *Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 511–515.

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

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2010. *FrameNet II: Extended Theory and Practice*. International Computer Science Institute, Berkeley, California.

Ralf Steinberger, Andreas Eisele, Szymon Klocek, Spyridon Pilos, and Patrick Schlüter. 2012. DGT-TM: A freely Available Translation Memory in 22 Languages. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 454–459.

Milan Straka and Jana Straková. 2017. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 2214–2218.

Marcin Woliński, Katarzyna Głowińska, and Marek Świdziński. 2011. A preliminary version of Składnica treebank of Polish. In *Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 299–303, Poznań, Poland.

Alina Wróblewska. 2012. Polish Dependency Bank. *Linguistic Issues in Language Technology*, 7(1):1–15.

Alina Wróblewska. 2014. *Polish Dependency Parser Trained on an Automatically Induced Dependency Bank*. Ph.D. dissertation, Institute of Computer Science, Polish Academy of Sciences, Warsaw.

Alina Wróblewska and Katarzyna Krasnowska-Kieraś. 2017. Polish evaluation dataset for compositional distributional semantics models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 784–792, Vancouver, Canada. Association for Computational Linguistics.

Alina Wróblewska and Aleksandra Wieczorek. 2018. Status składniowy *jako* we współczesnej polszczyźnie. *Język Polski*, to appear.

Daniel Zeman, Ondřej Dušek, David Mareček, Martin Popel, Loganathan Ramasamy, Jan Štěpánek, Zdeněk Žabokrtský, and Jan Hajič. 2014. HamleDT: Harmonized Multi-Language Dependency Treebank. *Language Resources and Evaluation*, 48(4):601–637.

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 Gökırmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdenka 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 dePaiva, 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. Association for Computational Linguistics.