A Preliminary Version of Składnica — a Treebank of Polish

Marcin Woliński*, Katarzyna Głowińska*, and Marek Świdziński†

* Institute of Computer Science PAS
† Warsaw University

Abstract

We present a bank of constituent parse trees for Polish sentences taken from the balanced hand-annotated subcorpus of the National Corpus of Polish (NKJP). The treebank has been built by automatic parsing and manual disambiguation of the resulting trees. The feedback from this process has been used to improve the grammar.

In this paper, we briefly describe our selection of texts analysed, the grammar and the parser we use, the process of validating trees, and the resulting treebank. We also analyse the types of errors in the treebank annotation.

1. Introduction

This paper reports on the results of a project aimed at building a constituency treebank of Polish. The three-year project, completed in October, was partially funded by the research grant N N104 224735 from the Polish Ministry of Science and Higher Education. It was carried out by a small group of researchers from the Institute of Computer Science PAS and Warsaw University accompanied by about 15 annotators.

To attain consistency of the treebank we decided to apply a semi-automatic method: trees were generated by an automatic parser and then selected and validated by humans. This way annotators were guided through a formally defined set of possibilities.

We decided to build on our experience in implementing Świdziński's constituency grammar (Świdziński, 1992), which provided us with a formal description of a large subset of Polish. That, however, means we could not draw much experience from projects for other Slavic languages, as they concentrate mostly on dependency formalisms (most notably the PDT: Böhmová et al., 2003). It is worth noting that the possibilities for parsing Polish are far more limited than for English, where one can choose from several readily available parsers. While formal descriptions of many interesting Polish syntactic phenomena exist (Obrębski, 2002, Przepiórkowski et al., 2002), the parsers are rather limited, in particular, none of them have been tested against a large corpus.

It should be emphasized that the Polish language provides the researcher with challenges not known to those who deal with English. Polish is a free word order and highly inflected language, very complicated morphologically (no less than 500 conjugation and declension patterns) and dramatically homonymous (42 in 100 words have more than one grammatical or semantic interpretation).

Since no one before us has really attempted deep parsing of a corpus of Polish, we view our project mainly as a pilot work undertaken to gain experience. At this stage, we accept the fact that the resulting treebank will be biased by the grammar, but we already plan further work to alleviate this effect.

2. The selection of texts

We worked on the one-million-word balanced subcorpus of the National Corpus of Polish (NKJP, `http://nkjp.pl`, Przepiórkowski et al. (2010, 2008)). The subcorpus was manually annotated with morphological features within the NKJP project. The output of a morphological analyser was disambiguated and grammatical features of words unknown to the analyser (mainly proper names) were added. Consequently, every word in the subcorpus has exactly one morphological interpretation.

We have randomly selected 20,000 sentences from this corpus. However, we excluded speech transcripts (due to absence of punctuation characters, which play an important role in the Polish syntax) as well as Internet texts (which are 'dirty', i.e., use non-standard grammar and awkward punctuation). This is why we concentrated on edited written texts in the present project. We think that even such a limited task is sufficiently ambitious, since this is the first attempt at building a treebank of Polish.

One of the annotators' duties in the project was to classify the sentences with respect to their suitability for parsing. The classification is presented in Table 1. Overall, about 67% of sentences were admitted for further processing in this project, while the rest will be taken care of in follow-up works.

	of all	of rejected	of 'too difficult'
accepted for processing	67%		
rejected	33%		
grammatical errors	2%	6%	
problem in NKJP	2%	5%	
no finite form	11%	35%	
'too difficult'	18%	55%	
direct speech	7%	21%	39%
quotes	5%	15%	27%
dashes	3%	10%	19%
brackets	3%	8%	14%
colon	1%	2%	3%
discontinuity	1%	3%	5%
other	0%	1%	2%

Table 1: Classes of sentences in the working set

We did not attempt any robust parsing techniques in this project, so sentences involving errors got rejected at this stage. It is worth noting that most typos in texts were corrected in NKJP (in fact, the typos remain in the text but correct base forms and tags have been inserted, which is enough for parsing). Other errors were of syntactic nature, but the class is rather small (about 2% of all sentences).

We decided not to change morphological descriptions obtained from NKJP in any way. This means that errors in the NKJP description itself are yet another instance in which sentences are rejected. This also includes some isolated cases where NKJP annotation rules are incompatible with ours.

We limited the scope of the project to sentences consisting of (possibly coordinated) finite clauses (i.e., those that are based upon the finite verb). As it turns out, constructions without verbal predicates amount to 11% of our corpus, resulting in 35% of rejections.

We also excluded from our analysis some finite sentences too cumbersome to account for (labelled 'too difficult' in the table[1]). The largest group in this category corresponds to the use of direct speech. Originally, we assumed that direct speech can be embedded in sentences in rather unpredictable ways. Inspection shows, however, that almost all of these sentences follow one pattern: optional dash, 'quoted' clause, dash, 'narrator's' clause. Thus, inclusion of most instances of this type of structures in the next versions of the grammar will be rather simple.

The next group of problems concerns punctuation characters which are used in an idiosyncratic manner: quotation marks, dashes, brackets, and colons. These marks often introduce bracketing, e.g., signalling ancillary elements in the sentence, but can also act as conjunctions, or just signal an arbitrary pause. We think that these phenomena require a separate analysis.

Since we use a constituent-based formalism, discontinuous structures also posed a problem. It is worth noting, however, that we did analyse some structures which might be considered discontinuous. For example, the auxiliary verb form in the analytical future tense can be separated from the main verb by other phrases. For that reason we introduced a separate 'phrase' for that auxiliary element that connects with the main verb at the level of the clause, avoiding discontinuity (cf. Fig. 1). It turns out that due to such interpretations, only about 1% of sentences (3% of rejected sentences) remained discontinuous in accordance with our view. It is a promisingly low number for a language involving free word order (cf. Derwojedowa, 2000).

To sum up, the analysis shows that to enlarge the percentage of analysed sentences we should first of all describe sentences without a finite verb (35% of all rejected); sentences with quotes, brackets, dashes and colons (34%); and sentences containing direct speech (21%). Fortunately, discontinuous structures do not seem to cause much of a problem.

[1]The division of this category was calculated based on a random sample of 100 sentences since annotators used free descriptions at this point. The numbers do not amount up to 100% since some sentences had more than one problem.

3. The Grammar

The grammar used in the project is a new version of Marek Świdziński's grammar of Polish (Świdziński, 1992), expressed in the Definite Clause Grammar formalism (Pereira and Warren, 1980). The original version of this grammar was implemented as the Świgra parser (Woliński, 2004). For the current project, the grammar has undergone a deep reconstruction (Świdziński and Woliński, 2010, 2009).

The trees we work with are constituency trees, whose nodes can be classified into four types (or layers), cf. Fig. 1 and Świdziński and Woliński, 2010:

- Syntactic words represent word forms including analytical forms (e.g., analytical future forms of verbs *będzie czytać* and other cases where one form, from the syntactic viewpoint, corresponds to several tokens in the IPI PAN tagset, cf. Przepiórkowski and Woliński, 2003) and other multiword units like two-word prepositions *wraz z* 'together with' and adverbs *po ciemku* 'in the dark'.

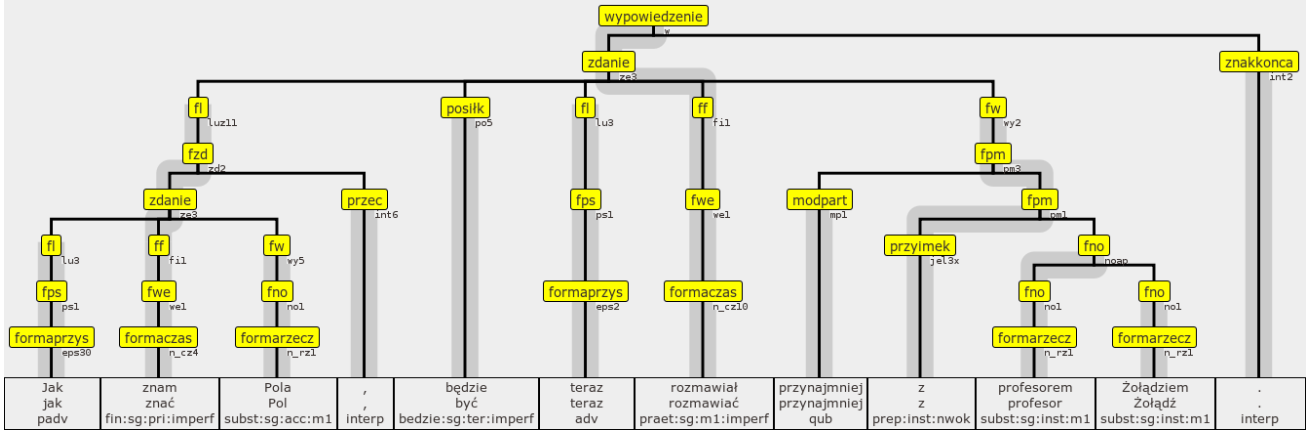
- Constituent phrases are used to describe the attachment of various modifiers to verbal, nominal, adjectival, and adverbial centres. Also at this level, prepositional-nominal phrases and subordinate clauses are formed. Constituent phrases can also be coordinate structures.

- Clause structure or functions played by constituent phrases in the clause are represented by the nodes of the third level comprising the finite phrase ff, which is the clause centre, and its dependents: required phrases (arguments) fw and optional phrases (adjuncts) fl.

- The fourth layer comprises clauses. Simple clauses consist of phrases of the third level. Coordinate clauses, based upon the conjunction as their centre, have clauses as their constituents.

Punctuation characters are treated as constituents in the tree. In particular, we are trying to capture a system of constraints on the use of commas in the sentence. This is somewhat complex, since the comma in Polish sometimes acts as a coordinate conjunction, and sometimes appears just as an orthographic separator. Both functions can be fulfilled simultaneously by one 'real' comma, which, moreover, disappears in the context of final punctuation.

It should be noted that in the project we did not pay attention to the problem of over-generating trees by the grammar. Since the output of the parser was to be manually disambiguated, it was more important to have the right tree among those generated than not to generate spurious ones. Actually, the grammar often generated very many trees, which is hardly avoidable in a non-lexicalised grammar since applicability of some interpretations depends on a variety of conditions, be them syntactic, semantic, or pragmatic. For example, to assess whether a given phrase is a complement or adjunct, a human annotator has to resort to all these criteria, and sometimes the problem remains unclear and the decision becomes arbitrary. Therefore, the grammar has to generate interpretations with both optional and required phrases (taking valence frames for the given verb into account). Given that many arguments may accompany a verb, the number of possible optional/required patterns can easily go into hundreds.



Jak *znam* *Pola,* *będzie* *teraz* *rozmawiał* *przynajmniej* *z* *profesorem* *Żołądziem.*
how know [surname] will now talk at least with professor [surname]

'As far as I know Pol, he will now be talking at least with professor Żołądź.'

Figure 1: An example constituency tree from the treebank

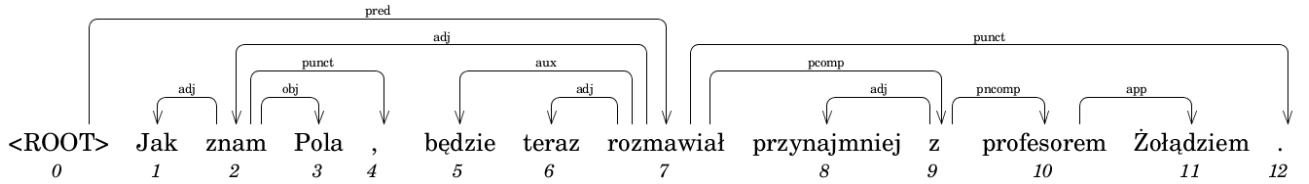


Figure 2: A dependency tree resulting from automatic conversion of the tree in Fig. 1

If our grammar is to be used in parsing without human disambiguation, it will perhaps need some changes limiting over-generation. In fact, we hope to study such perspectives using the treebank at hand.

4. Disambiguation of Parse Trees

Disambiguation and validation of parse trees was carried out in a web-based system Dendrarium. As noted in section 2, the first decision made by annotators was whether the sentence qualifies for processing. If so, the annotator had to choose the right tree from the forest by selecting interpretations for ambiguous nodes as prompted by the system (Woliński, 2010). If no tree was correct, the sentence was marked as calling for changes in the grammar. The annotator was expected to provide a comment on the flaws of the trees present. Thus, the process was iterative: the grammar and the treebank were developed in parallel, as advocated for, e.g., by Branco (2009), Rosén et al. (2006).

As is commonly done, each sentence was considered by two annotators. If their opinions differed, the sentence was passed to the adjudicator. The inter-annotator agreement (percentage of cases where two annotators agreed on an answer) is 88% in our project. In 71% of collisions, the adjudicator selected one of the answers of annotators as correct.

Table 2 shows numbers of collisions for various types of answers. The second column shows partition of collisions by type of final answer as decided by the adjudicator. The third column contains percentage of answers of the given type where the adjudicator selected one of the answers of annotators as the correct one. As we can see, in the case of

type of final answer	collisions	one of the answers accepted
full	40%	79%
no tree	37%	65%
rejected	23%	66%

Table 2: Collisions by type of answer given by the adjudicator

full answers (where the correct tree was found) one of the annotators used to be right in 79% of cases. The annotators used to have more trouble with cases where no correct tree was found: only in about 65% of sentences the adjudicator accepted one of the answers. We think that the reason for the problematic 'no tree' answers was that annotators were reluctant to reject existing trees (thinking 'the grammar is right'), while for 'rejected' ones, the instructions given to annotators were probably too vague.

As an additional check, we have gone through a random sample of 100 sentences with full answers on which both annotators agreed. We have found 18 erroneous trees, some containing more than one problem. Six times, annotators attached a subordinate phrase in the wrong place (these were evident errors, not ambiguous attachments). Eight problems stem from the complement/adjunct distinction, or, in our terminology: the distinction between an optional and a required phrase; four of those cases concerned arguments of non-finite forms: gerunds and participles. The type of a required phrase was selected wrong once. Twice, annotators had a problem with the genitive of negation (in

case of the verb *nie ma*, which is a bit tricky, as it means *there isn't* while looking like the negative form of the verb *mieć* 'to have'). Four times, annotators selected a gravely wrong structure of the tree, in two of those cases, that allowed them to abstain from assessing that no correct tree is present. In one sentence, a wrong subject was selected for the predicative 'to' (which is an idiosyncratic variant of *to be*).

We have to admit that the error ratio is high. Unfortunately, this may be a result of the nature of the problem at hand. In fact, the majority of problems arise at places we have expected them. In particular, the required/optional phrase distinction is difficult, with many border cases.

Some of the spotted problems are connected with the more advanced syntactic constructs (like apposition and required non-agreement nominal phrases). It seems that annotators used to have trouble remembering specific rules concerning such cases. Fortunately, problems of this kind can be easily picked in the results and systematically corrected, which we intend to do.

5. The Treebank

From the working set of 20,000 sentences our grammar accepts 11,535, that is 57.7%. Annotators decided that no correct tree is present for 24.6% of the successfully generated trees. On the other hand, 34.0% of sentences were rejected by the annotators as described in section 2. If we take these two factors into account, we will find that the parser generates correct parse trees for 62% of the sentences it was expected to parse. Currently our treebank contains 7841 manually validated trees.

Ogrodniczuk (2006) reports being able to enhance Świgra's coverage to the level of 84% of sentences in a certain corpus (although he did not perform a thorough validation of the resulting trees). We expected to be able to achieve similar results on a larger corpus. Unfortunately, our results are worse, even though the current version of the grammar seems to cover most typical structures of Polish, including those added to Świgra by Ogrodniczuk. It turns out that many of the problems are of lexical nature. For example, a detailed subclassification of particles is needed. This holds true for many multi-word expressions (syntactic words) appearing in the texts. These problems are easy to correct individually, but it takes much work. Unfortunately, every failed sentence has to be inspected, and the correction of one sentence may help us parse at best a few others. This process has turned out to take much more time than we originally expected.

In the course of the project, it turned out that some level of lexicalisation of the grammar would be helpful. In our current structures the required phrases do not directly show the lexical element of their centre. This feature, however, would be very convenient in describing idiomatic expressions involving verbs which are somewhat similar to English phrasal verbs. There are verbs that allow for certain required phrases only in the presence of specific lexically bound elements. For example, the Polish expression *mieć do czynienia z X* 'to deal with X' involves the verb *mieć* 'to have' which does not subcategorize for a prepositional phrase *z/np(gen)*. However, in the presence of the phrase *do czynienia*, such a prepositional phrase becomes mandatory. Access to the lexical heads of phrases would allow us to adequately describe such constructions. We will probably introduce this information in phrases of all levels since this could also help fight over-generation by limiting some constructions to lists of lexical realisations. However, this change would have too much impact on the trees currently in the system, so we will perform appropriate transformation in the following work.

The trees currently in the treebank are not completely uniform with respect to the version of the grammar. They were generated with various versions of the parser. Most changes in the grammar involved only adding new rules, which means the previously selected trees should still be valid with respect to the newer grammar, but this claim should be explicitly verified. On the other hand, since we plan some restructuring of the grammar in the near future, we think it will be better to bring the trees to a uniform state after those changes.

An interesting feature of our treebank is that it can be viewed as a hybrid constituency/dependency treebank. Although we work with constituency structures, we mark one of constituents of each node as its centre. This way, the constituency trees can be mapped into dependency trees in an obvious way. Moreover, the information provided is in most cases sufficient to derive labels for dependency arcs (cf. Fig. 2). E.g., the required/optional distinction of level 3 nodes provides information on complements and adjuncts, which is refined by types of required phrases carried in an attribute. In fact, promising experiments on training data-driven dependency parsers on such converted trees have already been carried out (Wróblewska and Woliński, 2011).

6. Summary and Prospects

During the project we have built the first relatively large treebank of Polish (though much smaller than currently existing treebanks for Czech and Russian).

Although the current project has ended, we will continue our work in the field, since a follow-up project is already scheduled. The main directions will be: correction of errors signalled in section 4, improving the grammar so that its bias is reduced, and obviously extending the treebank with new sentences. One of our goals will be to eliminate the class of 'too difficult' sentences by extending the grammar.

Although there is plenty of room for improvement, we hope that we have already built an interesting resource. It may seem that the present version of the treebank is not suitable for machine learning, but the abovementioned experiments in dependency parsing suggest otherwise. This is not completely unexpected, since learning algorithms have some ability to generalise, and the treebank already contains a rich collection of Polish syntactic structures. For the first time, such a set of structures built over real sentences is available to Polish linguists. We think it can be reasonably considered as a representative corpus of Polish finite sentences, which in itself should prove interesting and useful for those who study Polish syntax.

References

Böhmová, A., Hajič, J., Hajičová, E., and Hladká, B. (2003) The Prague Dependency Treebank: A 3-level annotation scenario. In Abeillé, A., editor, *Treebanks. Building*

and Using Parsed Corpora, chapter 7, pages 103–127. Kluwer Academic Publishers.

Branco, A. (2009) LogicalFormBanks, the next generation of semantically annotated corpora: Key issues in construction methodology. In Kłopotek, M.A., Przepiórkowski, A., Wierzchoń, S., and Trojanowski, K., editors, *Recent Advances in Intelligent Information Systems*, pages 3–11. Akademicka Oficyna Wydawnicza Exit, Warsaw.

Derwojedowa, M. (2000) *Porządek linearny składników zdania elementarnego w języku polskim*. Elipsa, Warszawa.

Obrębski, T. (2002) *Automatyczna analiza składniowa języka polskiego z wykorzystaniem gramatyki zależnościowej*. PhD thesis, Instytut Podstaw Informatyki PAN, Warszawa.

Ogrodniczuk, M. (2006) *Weryfikacja korpusu wypowiedników polskich (z wykorzystaniem gramatyki formalnej Świdzińskiego*. PhD thesis, Uniwersytet Warszawski, Wydział Neofilologii.

Pereira, F. and Warren, D. (1980) Definite clause grammars for language analysis–a survey of the formalism and a comparison with augmented transition networks. *Artificial Intelligence*, 13:231–278.

Przepiórkowski, A. and Woliński, M. (2003) The unbearable lightness of tagging: A case study in morphosyntactic tagging of Polish. In *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-03), EACL 2003*, pages 109–116.

Przepiórkowski, A., Kupść, A., Marciniak, M., and Mykowiecka, A. (2002) *Formalny opis języka polskiego: Teoria i implementacja*. EXIT, Warszawa.

Przepiórkowski, A., Górski, R., Lewandowska-Tomaszczyk, B., and Łaziński, M. (2008) Towards the National Corpus of Polish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation, LREC 2008*, Marrakech, ELRA.

Przepiórkowski, A., Górski, R., Łaziński, M., and Pęzik, P. (2010) Recent developments in the National Corpus of Polish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation, LREC 2010*, Valetta, Malta, ELRA.

Rosén, V., de Smedt, K., and Meurer, P. (2006) Towards a toolkit linking treebanking to grammar development. In Hajič, J. and Nivre, J., editors, *Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories*, pages 55–66.

Świdziński, M. (1992) *Gramatyka formalna języka polskiego*. Rozprawy Uniwersytetu Warszawskiego. Wydawnictwa Uniwersytetu Warszawskiego, Warszawa.

Świdziński, M. and Woliński, M. (2009) A new formal definition of Polish nominal phrases. In *Aspects of Natural Language Processing*, LNCS 5070, pages 143–162. Springer.

Świdziński, M. and Woliński, M. (2010) Towards a bank of constituent parse trees for Polish. In Sojka, P. et al., editors, *Text, Speech and Dialogue, 13th International Conference, TSD 2010, Brno, September 2010, Proceedings*, volume 6231 of *LNAI*, pages 197–204, Heidelberg, Springer.

Woliński, M. (2004) *Komputerowa weryfikacja gramatyki Świdzińskiego*. PhD thesis, Instytut Podstaw Informatyki PAN, Warszawa.

Woliński, M. (2010) Dendrarium—an open source tool for treebank building. In Kłopotek, M., Marciniak, M., Mykowiecka, A., Penczek, W., and Wierzchoń, S., editors, *Intelligent Information Systems*, pages 193–204. Siedlce, Poland. ISBN 978-83-7051-580-5.

Wróblewska, A. and Woliński, M. (2011) Preliminary experiments in Polish dependency parsing. In Bouvry, P., Kłopotek, M.A., Leprevost, F.; Marciniak, M., Mykowiecka, A., Rybiński, H., editors, *Security and Intelligent Information Systems*, volume 7053 of *LNCS*, Springer.