Towards a Bank of Constituent Parse Trees for Polish

Marek Świdziński[1] and Marcin Woliński[2]

[1] Institute of Polish, Warsaw University
[2] Institute of Computer Science, Polish Academy of Sciences

Abstract. We present a project aimed at construction of a bank of constituent parse trees for 20,000 Polish sentences taken from the balanced hand-annotated subcorpus of the National Corpus of Polish (NKJP).

The treebank is to be obtained by automatic parsing and manual disambiguation of resulting trees. The grammar applied by the project is a new version of Świdziński's formal definition of Polish. Each sentence is disambiguated independently by two linguists and, if needed, adjudicated by a supervisor. The feedback from this process is used to iteratively improve the grammar.

In the paper, we describe linguistic but also technical decisions made in the project. We discuss the overall shape of the parse trees including the extent of encoded grammatical information. We also delve into the problem of syntactic disambiguation as a challenge for our job.

Keywords: treebank of Polish, constituent parse trees, DCG grammar.

1 The Project

This paper reports on a project aimed at building a treebank of Polish.[1] To the best of our knowledge, it will be the first large treebank of Polish. In the present project trees for about 20,000 sentences will be compiled. The project will finish in a year but a follow-up work is already planned.

The treebank is built in a semi-automatic process. Parse trees (or, rather, parse forests) are generated by a parser and then selected and validated by human annotators. A linguist tries to choose the right tree for each sentence. If there is no valid tree in the forest the linguist can judge the sentence ungrammatical; in this case the processing ends. However, if the sentence is judged correct it is passed to the authors of the grammar, who will improve/correct the grammar, and a new parse forest will be presented for reconsideration.

Thus, the process is iterative: the grammar and the treebank are developed in parallel, as advocated for e.g. by [1,2]. Note, however, that since we work with a constituent grammar we cannot draw much experience from other projects for Slavic languages, as they work mostly in dependency formalisms (most notably the PDT [3]).

The process is facilitated by a web-based environment Dendrarium [4] developed especially for our treebank. Every sentence is processed independently by two users.

[1] The project is partially funded by the research grant N N104 224735 from the Polish Ministry of Science and Higher Education.

P. Sojka et al. (Eds.): TSD 2010, LNAI 6231, pp. 197–204, 2010.

Then, if their answers differ a supervisor makes a choice. An important feature of the system is that it allows for avoiding repeated work caused by changes in the grammar.

We work on one-million word balanced subcorpus of the National Corpus of Polish (NKJP, `http://nkjp.pl`, [5,6]). The subcorpus has been manually annotated with morphological features, which means that the output of a morphological analyser is disambiguated and grammatical features of words unknown to the analyser (mainly proper names) are added. Consequently, every word in the subcorpus bears exactly one morphological interpretation.

In the current project we have decided not to correct the texts or their morphological interpretations we receive from the NKJP project. Such corrections may be applied in a follow-up work.

Moreover, we do not assume that we are able to process every correct Polish utterance. We decide to describe only finite sentences (i.e., those that are based upon the finite verb) and coordinate sentences, leaving other types of expressions untouched. As we use a constituent-based formalism discontinuous structures pose a problem. In the present project we will cope with but a very limited set of types of such phrases, which means that most of discontinuous sentences will not get parse trees.

What follows from all this is that our treebank will be biased by the grammar. However, due to limited resources we are not able to achieve a 100% coverage on the corpus. Therefore, we will classify utterances rejected by the parser, which will hopefully allow us to reduce the bias in the follow-up work.

2 The Grammar

The grammar used in the project is a new version of Marek Świdziński's grammar of Polish [7] expressed in the Definite Clause Grammar formalism [8] and implemented as the Świgra parser [9]. The grammar has undergone a deep reconstruction. The set of nonterminals has been limited (e.g., seven clause types have been reduced to one), resulting in trees that are much simpler, their height being significantly reduced. We are in the process of supplementing the grammar with rules that define constructions neglected in the previous version [10].

The new grammar redefines finite sentences. The whole definition is a set of three rules that account for all possible permutations of three types of components: finite phrase (verb) plus 0–3 required phrases (arguments) and 0–3 free phrases (adjuncts)[2]. The DCG apparatus has been extended to allow right hand sides of rules to include sequences of nonterminals of arbitrary length.

The grammar accounts for punctuation as a sophisticated syntactic phenomenon. Commas are strongly context-bounded in Polish. To cope with the problem special parameters are introduced in most rules that either block comma appearance in some contexts or force it in other.

Another interesting feature is that rules provide information which of the components of the construction defined by a given rule is its center (head). It will make it possible to generate dependency trees from constituent ones (cf. [11]).

[2] There are three other auxiliary compoments not worth mentioning here.

It should be emphasized that the Świdziński's grammar has an ambition to catch all possible structurizations (interpretations) for sentences examined. As the parser does not include any statistical component it is typical that we get large sets of trees even for short (i.e., simple) sentences. Therefore the problem of syntactic disambiguation is of crucial importance in the project.

3 The Structure of Parse Trees

Figure 1 presents an example of a parse tree, as displayed in our treebanking environment.

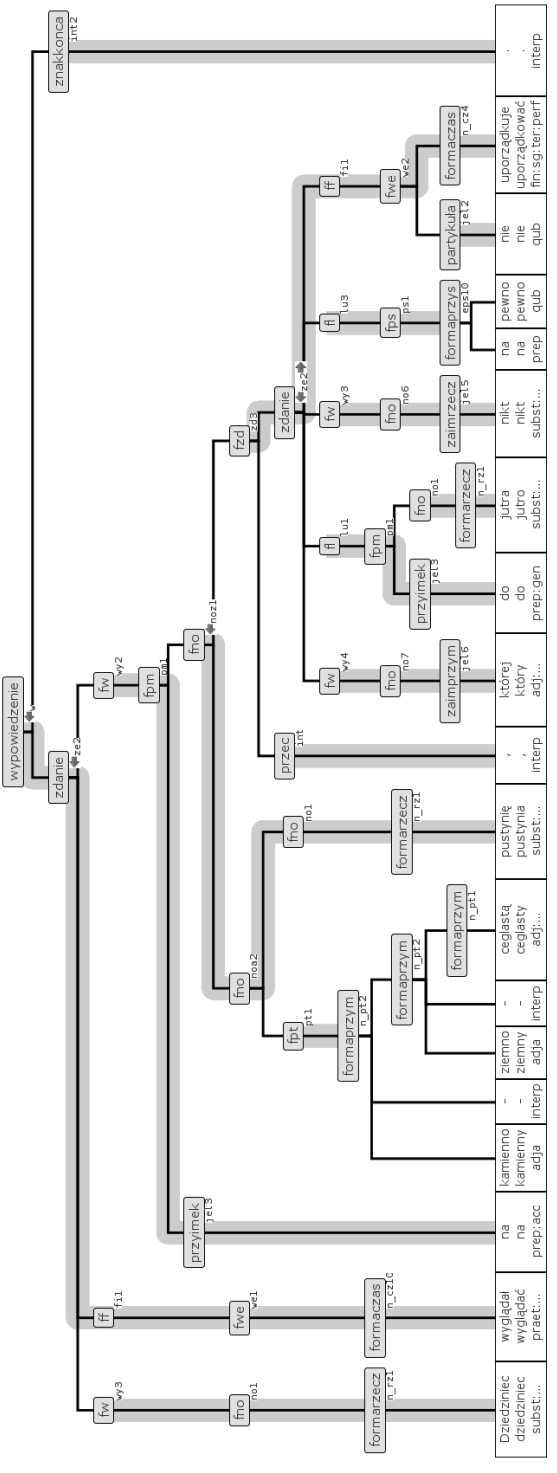
As mentioned before, the text analysed comes from the National Corpus of Polish, and both tokens and their morphological description originate directly therefrom. Tokens are displayed as a series of boxes at the bottom of the tree. Each box contains a word-form, a lemma, and morphological characteristics in the IPI PAN Tagset notation [12,13].

Nodes of the tree are assigned names of nonterminal units. Thick gray shadows emphasising some branches in the tree show distributional centres of phrases. They join each of the nonterminals with the token representing its centre. This way the utterance (wypowiedzenie) and the main clause (zdanie) in the example are connected with the verb *wyglądać* which is the centre of the whole sentence. For nominal phrases the subordinating noun is taken for their centre. Exocentric constructions get an arbitary interpretation. For example, prepositional-nominal phrases fpm have the preposition as their centre.

Nonterminal nodes in the tree can be seen as organised into several layers of annotation starting from the bottom.

The first layer represents what we call 'syntactic words' (the term coined by Przepiórkowski [14]). It comprises units like formarzecz (nominal form), zaimrzecz (nominal pronoun), formaczas (verbal form), przyimek (preposition), formaprzym (adjectival form), zaimprzym (adjectival pronoun), formaprzys (adverbial form), partykuła (particle), przec (comma). Typically, each corresponds to just one token. However, tokens (or segments) in NKJP are rather fine-grained, and sometimes a sequence of tokens (words) is taken for one unit on the 'presyntactic' level. For example, past forms of verbs are, following NKJP, formalised as composed of a 'past participle' and an agglutinated auxiliary form of the verb 'to be' (e.g., the form *zrobiłem* 'I-did' is a sequence of *zrobił-*, which carries tense and gender, and the auxiliary *-em* carrying person and number). The two segments need not be adjacent; the latter can precede the former, attached to some word before (sometimes obligatorily), as in *Chcesz, żebym to zrobił.* lit. you-want that-I-masc-sg it do 'You want me to do it.' (cf. [15]).

A similar approach applies to compounded adjectives—expressions locating at the boundary between lexicon and syntax. In Polish, we have got a special compound-building form of adjectives that ends in *-o* (marked adja in the IPI PAN Tagset) to be prefixed to the regular forms of adjectives with or without use of hyphen, e.g., *biało-czerwony* 'white and red' (as the Polish flag), *żółtozielony* 'greenish yelow', or *polsko-ukraińsko-rosyjski* 'Polish Ukrainian Russian' (e.g., a summit). This type of compounding is recursive. Similar nominal expressions exist in Polish as well though compounding is fairly more restricted.



Dziedziniec wyglądał na kamienno-ziemno-ceglastą pustynię, której do jutra na pewno nikt nie uporządkuje.
courtyard looked on stony earthy bricky desert which until tomorrow on sure nobody not will-tidy-up

'The courtyard looked like a stony, earthy, and bricky desert that no one will for sure tidy up until tomorrow.'

Fig. 1. An example of a parse tree. Morphological descriptions have been shortened. For explanation of particular labels see text.

Moreover, syntactic words are used to cater for analytical (multi-token) verb forms (e.g., *będzie robić* 'will do'), reflexive verbs (e.g., *boi się* lit. 'is-afraid-of self'), and some idiomatic multiword expressions (e.g., *na pewno* 'surely'). The last expression is treated as an adverb in the tree, while its tokens being the preposition *na* and an idiosyncratic adjectival word form *pewno* that appears only in this idiom (cf. [16]).

Although represented with regular nodes and edges, we treat this level as pre-syntactic since the units in question differ from standard syntactic units in some important ways. In particular, it is rather hard to identify their centres. We assume that the whole syntactic word is the centre for higher-level structures. It is also impossible to substitute higher level structures, like the adjectival phrase, for components thereof (e.g., **biało-ciemno czerwony* 'white and dark red' is not possible), which in itself justifies introducing special units for them.

The second layer of nonterminal units represents constituent phrases understood morphologically: verbal phrases fwe, nominal phrases fno, adjectival phrases fpt, prepositional-nominal phrases fpm, "sentential" phrases (i.e., subordinate clauses) fzd, and so on. Obviously, we allow for an arbitrary level of complication within constituents. They may be coordinate structures, include modifiers of various types, contain embedded clauses, etc. (cf. [10])

The third layer is needed to reveal clause structure, according to the definition of sentence (zdanie) in Świdziński´s grammar. The phrases of the second level are classified according to their function as constituents of a clause. On this level we locate the finite phrase ff, which is the clause centre, and its dependents: required phrases (arguments) fw and free phrases (adjuncts) fl. Subject of the clause is one of its required phrases, others being complements. This way valence frames are easily visible in the tree structure.

The fourth layer comprises clauses. Simple clauses consist of phrases of the third level. Subordinate clauses are regarded as phrases in our account. There are also coordinate clauses (not present in the example sentence).

Finally, the root of the tree is utterance (wypowiedzenie). It consists of a clause and a final punctuation (znakkonca). As can be seen, punctuation characters are treated as constituents in the tree.

The Świdziński's grammar assigns numerous attributes to nonterminals; we omit them in Fig. 1 to save space. They include morphological features but most attributes are, in fact, purely syntactic. They formalise various contextual co-occurrence restrictions. We have got, e.g., a parameter responsible for clasification of syntactic units (clauses and phrases) according to whether they contain an interrogative or relative terminal or whether they can serve as, or appear within, a subordinate clause of a given type. The general philosophy of the grammar is that everything is bounded (or controlled in a non-Chomskyan sense) in Polish expressions ("overagreement" philosophy).

4 Syntactic Ambiguities

Since we work on a manually disambiguated subcorpus of NKJP we are spared the issue of morphological ambiguities. It means that ambiguities we have to deal with are of structural or syntactic nature.

In the project disambiguation of parse forests is performed in terms of ambiguous nodes, i.e., nodes where different subtrees spanning the same tokens can be attached (cf.[4]). This happens when several rules can alternatively be used to obtain the same nonterminal with a fixed set of attributes.

According to our experiences so far, the node for clause (zdanie) seems most difficult. Variant realisations for a clause can be quite numerous (over 50) due to interaction of several problems:

1. Some constituents can be split into smaller units or aggregated further on. For example, in the relative clause taken from our example we find the free phrase *do jutra* and the required phrase *nikt* as components. They could be combined into one phrase. The phrase would not have any reasonable semantic interpretation but from the purely syntactic point of view it is perfectly admissible, since the phrase *do domu droga* 'way home' with the same structure is plausible. This type of variation leads to different number and extents of constituents of the clause.

2. The problem of distinguishing complements from adjuncts provides another source of ambiguity. In our formalism the distinction results in variants of zdanie differing only in labelling particular components as a required phrase fw or a free phrase fl. This ambiguity can be limited to some extent by appropriate entries in the valence dictionary. Unfortunately, it cannot be completely avoided. Most of possible realisations of required phrases are shared by free phrases. Even if we get a complement interpretation for a given component on the basis of valence data the parser will give an adjunct interpretation as well (we do not demand that valence frames be realised "non-elliptically").

A simple example can illustrate the point. Nominal phrases in accusative serve as a very common complement. Some accusative phrases can also be adjuncts describing the length of events (*godzinę* '[for] an hour', *chwilę* '[for] a moment', *dwie kolejki* '[for] two rounds'). Neither rejection of accusative nominal realization of fl by the grammar, nor allowing for it on a lexical basis (a list of possible centres of such fl's) seems easy to do. The decision has to be taken by annotators.

3. An additional level of complication pertains to verbs that require adverbial phrases. Required adverbial phrases can often be substituted with prepositional-nominal phrases with various prepositions. For example, verbs of location, movement or translocation, like *mieszkać* 'live', *pojechać* 'go (to)', or *zanieść* 'carry sth somewhere' require adverbial phrases. For each of them the adverbial phrase is substitutable by a prepositional-nominal one; cf., e.g., *Mieszkam tutaj (w Warszawie, nad morzem, pod Toruniem, ...)*. 'I live here (in Warsaw, at the seaside, close to Toruń, ...)'. The parser of course allows for such substitutions, and annotators have to decide whether a given required phrase is prepositional or adverbial. The problem is limited to verbs missing from the valence dictionary. In such cases the parser uses a default frame allowing both for advp and prepnp.

In other ambiguous nodes of our trees variants are generally much less numerous. The main problem here is that alternative modifier attachments lead to various chunkings of a phrase. So, the number of variants in a particular node is generally limited by the number of spanned tokens; usually it is 2 or 3. The worst case of ambiguity is represented by so called genitive clusters, i.e., series of nouns in genitive, each of which

can modify any other (respecting continuity). We have seen such sequences of the length of 11; theoretically, the length is unlimited.

Let us illustrate it with an example. The nominal phrase *ostatnie słowa Stalina* 'the last Stalin's words' can be structured twofold: either as *(ostatnie) (słowa Stalina)*, or as *(ostatnie słowa) (Stalina)*. Both interpretations are fully correct. No semantic, thematic-rhematic, or pragmatic difference takes place here.

All this shows that human disambiguators are unavoidable in the project. One can assume that each utterance, regarded as a part of a given speech act, meets exactly one interpretation. Therefore, our experts are provided with an (informal) instruction of how to evaluate outputs of the parser. Actually, the instruction gives room to semantic reflection and, moreover, access to the context.

There are some semi-formal (superficial) prompts to include in the instruction for our disambiguators. If a given prepositional-nominal phrase is a constituent of the idiomatic verbal expression it is a complement, not adjunct (*umrzeć ze śmiechu* 'die of laughter'). If such a phrase must be asked about with this preposition it is a prepositional, not adverbial complement (*pojechać do Piotra* 'go to Peter': *do kogo?* 'to whom?', and not *dokąd?* 'where. . . to?').

5 Conclusions

Probably the most important message of this paper is that finally, after years of Polish lagging behind, a relatively large treebank of Polish is under construction. Since there are no previous experiences with large treebanks of Polish we consider the present project a pilot work, in which we will recognise the main obstacles to be taken up later. In particular we assume to build a catalogue of types of discontinuous phrases in Polish, which will need some special treatment.

A follow-up project is already scheduled, which, we hope, will allow us to include sentences skipped in the pilot phase and to enlarge the treebank to 50,000 sentences in 3 years time.

Annotation of the treebank is in progress, so we are yet unable to provide any evaluation.

We assume that the treebank we wish to obtain will allow us to reformulate the grammar so that it could block some spurious interpretations. This pertains in particular to the problem of free phrases. We already see two possible directions of development for our grammar, both based on the treebank: lexicalisation and including some statistical component.

References

1. Branco, A.: LogicalFormBanks, the Next Generation of Semantically Annotated Corpora: Key Issues in Construction Methodology. In: Kłopotek, M.A., et al. (eds.) Recent Advances in Intelligent Information Systems, Exit, Warsaw, pp. 3–11 (2009)
2. Rosén, V., de Smedt, K., Meurer, P.: Towards a Toolkit Linking Treebanking to Grammar Development. In: Hajič, J., Nivre, J. (eds.) Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories, pp. 55–66 (2006)

3. Böhmová, A., Hajič, J., Hajičová, E., Hladká, B.: The Prague Dependency Treebank: A 3-level Annotation Scenario. In: Abeillé, A. (ed.) Treebanks. Building and Using Parsed Corpora, pp. 103–127. Kluwer Academic Publishers, Dordrecht (2003)
4. Woliński, M.: Dendrarium – an Open Source Tool for Treebank Building. In: Kłopotek, M.A., et al. (eds.) Intelligent Information Systems, Siedlce, pp. 193–204 (2010)
5. Przepiórkowski, A., Górski, R.L., Łaziński, M., Pęzik, P.: Recent Developments in the National Corpus of Polish. In: Proc. of LREC 2010, ELRA (2010)
6. Przepiórkowski, A., Górski, R.L., Lewandowska-Tomaszczyk, B., Łaziński, M.: Towards the National Corpus of Polish. In: Proc. of LREC, ELRA (2008)
7. Świdziński, M.: Gramatyka formalna języka polskiego. Rozprawy Uniwersytetu Warszawskiego. Wydawnictwa Uniwersytetu Warszawskiego, Warszawa (1992)
8. Pereira, F., Warren, D.H.D.: Definite Clause Grammars for Language Analysis – a Survey of the Formalism and a Comparison with Augmented Transition Networks. Artificial Intelligence 13, 231–278 (1980)
9. Woliński, M.: Komputerowa weryfikacja gramatyki Świdzińskiego. Ph.D. thesis, Instytut Podstaw Informatyki PAN, Warszawa (December 2004)
10. Świdziński, M., Woliński, M.: A New Formal Definition of Polish Nominal Phrases. In: Aspects of Natural Language Processing. LNCS, vol. 5070, pp. 143–162. Springer, Heidelberg (2009)
11. Nivre, J.: Theory-Supporting Treebanks. In: Proceedings of the Second Workshop on Treebanks and Linguistic Theories (2003)
12. Przepiórkowski, A.: A Comparison of Two Morphosyntactic Tagsets of Polish. In: Koseska-Toszewa, V., Dimitrova, L., Roszko, R. (eds.) Representing Semantics in Digital Lexicography, Warsaw, pp. 138–144 (2009)
13. Przepiórkowski, A., Woliński, M.: A Flexemic Tagset for Polish. In: Proceedings of the Workshop on Morphological Processing of Slavic Languages, EACL 2003, pp. 33–40 (2003)
14. Przepiórkowski, A.: Powierzchniowe przetwarzanie języka polskiego. Exit, Warsaw (2008)
15. Przepiórkowski, A., Woliński, M.: The Unbearable Lightness of Tagging: A Case Study in Morphosyntactic Tagging of Polish. In: Proc. of the 4th Workshop on Linguistically Interpreted Corpora (LINC 2003), EACL 2003, pp. 109–116 (2003)
16. Derwojedowa, M., Rudolf, M.: Czy burkina to dziewczyna i co o tym sądzą ich królewskie mości, czyli o jednostkach leksykalnych pewnego typu. Poradnik Językowy 3 (2003)