Towards an LFG Parser for Polish
An Exercise in Parasitic Grammar Development

Agnieszka Patejuk, Adam Przepiórkowski

Institute of Computer Science, Polish Academy of Sciences
ul. Jana Kazimierza 5, 01-248 Warszawa, Poland
agnieszka.patejuk@gmail.com, adamp@ipipan.waw.pl

Abstract

While it is possible to build a formal grammar manually from scratch or, going to another extreme, to derive it automatically from a treebank, the development of the LFG grammar of Polish presented in this paper is different from both of these methods as it relies on extensive reuse of existing language resources for Polish. LFG grammars minimally provide two levels of representation: constituent structure (c-structure) produced by context-free phrase structure rules and functional structure (f-structure) created by functional descriptions. The c-structure was based on a DCG grammar of Polish, while the f-structure level was mainly inspired by the available HPSG analyses of Polish. The morphosyntactic information needed to create a lexicon may be taken from one of the following resources: a morphological analyser, a treebank or a corpus. Valence information from the dictionary which accompanies the DCG grammar was converted so that subcategorisation is stated in terms of grammatical functions rather than categories; additionally, missing valence frames may be extracted from the treebank. The obtained grammar is evaluated using constructed testsuites (half of which were provided by previous grammars) and the treebank.

Keywords: parsing, LFG, Polish

1. Introduction

The aim of this paper is to present a parasitic approach to grammar development, where a new LFG[1] grammar is created on the basis of a variety of resources, including a DCG[2]-like grammar for Polish and a currently developed treebank based on this grammar. The new grammar extends the original grammar "vertically", by adding the level of f-structure to the c-structure offered by the DCG grammar, and "horizontally", by attempting to cover a wider range of phenomena. Moreover, the coverage of the LFG grammar is regularly evaluated on the basis of the constantly extended treebank for Polish.

Section 2. briefly describes the resources the present exercise builds upon. Then section 3. presents the process of grammar development in more detail, while section 4. outlines the adopted method of ensuring a reasonable quality of the grammar during its development. Finally, section 5. concludes the paper.

2. Resources for Grammar Development

The effort of creating an LFG grammar implemented in the XLE platform (`http://www2.parc.com/isl/groups/nltt/xle/`) consists of two major tasks: creating annotated rules and building the lexicon. Since manual development of large-scale grammars is a rather costly and time-consuming task, the adopted strategy is to reuse as many available resources as possible instead of developing another grammar from scratch. As there is a wide range of language resources for Polish at hand, it is possible to draw on the results of many projects, completed and ongoing, and minimise the workload, concentrating on further improvements.

2.1. Previous Grammars

The context-free grammar rules based on Świdziński's 1992 grammar which were first implemented for use by another parser for Polish, Świgra (Woliński, 2004), constitute the basis of the current implementation. These rules were annotated with instructions on how to build an additional level of structure on top of this, namely the f-structure. This provides a representation employing grammatical functions which is considered more universal across languages than the constituent structure which is subject to much variation. The f-structure annotation was inspired by two resources: the original metamorphosis grammar used by Świgra and a small-scale but linguistically sophisticated HPSG[3] grammar of Polish (Przepiórkowski *et al.*, 2002).

2.2. Morfeusz

While most large-scale grammars implemented in XLE use XFST morphology combined with an additional set of rules, namely sublexical rules, the current grammar relies on Morfeusz, a state-of-the-art morphological analyser for Polish (Woliński, 2006). Therefore, rather than trying to build an FST morphology for Polish from scratch – a very demanding task in itself – the output provided by Morfeusz is converted into ready-made XLE lexical entries which correspond to full, inflected forms. Though the current solution gives satisfactory results, it would be possible to

[1] *Lexical Functional Grammar* (Bresnan, 1982, 2000; Dalrymple, 2001).

[2] *Definite Clause Grammar* (Warren and Pereira, 1980).

[3] *Head-driven Phrase Structure Grammar* (Pollard and Sag, 1987, 1994).

convert the output of Morfeusz into an FST using a text-specified transducer or to use a dedicated grammar library transducer. The latter solution is particularly worth considering as it could improve the efficiency of the grammar.

2.3. The National Corpus of Polish

The National Corpus of Polish (Przepiórkowski *et al.* 2010; `http://nkjp.pl/`), the largest currently available corpus of Polish which contains around 1.5 billion words out of which 1 million were manually annotated, is used in a twofold way. First, it may be used as one of alternative sources of information about morphosyntax and segmentation which is necessary to create a lexicon. Morphosyntactic information is specified according to the NCP tagset (Przepiórkowski and Woliński, 2003; Przepiórkowski, 2009) which additionally provided the names of many attributes and values in the f-structures created by Polish LFG, especially the non-standard ones. It is worth mentioning that morphosyntactic intepretations available for every segment in the NCP were disambiguated (automatically or, in case of the manual subcorpus, by human annotators), which results in far fewer possibilities than provided for the same segment by Morfeusz, for instance. Secondly, the NCP provides a rich body of interesting examples, which makes it possible to ensure that further extensions of the grammar have firm empirical grounding.

2.4. Składnica

The last main resource actively used in the development of Polish LFG is Składnica (Woliński, 2010; Świdziński and Woliński, 2010), a treebank containing parse trees selected by human annotators from the rich output generated by Świgra for selected sentences from the manually annotated subcorpus of the NCP. Składnica serves as the main testbed for the current grammar, ensuring backwards compatibility with the original grammar and checking grammar coverage on authentic texts. The information about morphosyntax and segmentation from manually disambiguated trees is converted into XLE lexical entries, which considerably reduces the amount of interpretations in comparison to Morfeusz. Additionally, it is possible to extract missing valence frames which were implicitly chosen by human annotators when selecting the correct parse.

3. Towards an LFG Grammar for Polish

As already mentioned, grammar development in XLE can be roughly divided into the creation of rules and the lexicon – this section presents this process in some more detail.

3.1. Annotating c-structure with f-descriptions

The original c-structure rules provided by GFJP2, the grammar currently used by Świgra (and constantly developed as new trees are added to Składnica), were manually rewritten so as to comply with XLE notational conventions. Even though this conversion could probably have been done automatically, there are some gains stemming from adopting this approach. Some linguistic generalisations expressed in rules, at the level of syntax, were transferred to the lexicon or gathered from various places in the grammar and stored in new syntactic templates.[4] The grammar writer had also the chance to better understand the mechanisms employed by the original grammar, which in some cases led to a decision to adopt a different analysis, either better motivated linguistically or more suitable from the perspective of the LFG formalism.

Adding f-structure annotation to the obtained c-structure required in the first place the identification of grammatical functions appropriate for Polish. This choice was made on the basis of rich LFG literature, as well as the solutions adopted within the ParGram project (`http://pargram.b.uib.no/`).

Analyses of many linguistic phenomena offered by the original DCG grammar could often be translated into the f-structure representation almost unchanged. There are, however, some significant differences, especially in the area of agreement, case assignment and negation where the LFG analysis draws broadly from the available HPSG analyses of these phenomena (Przepiórkowski, 1999; Przepiórkowski *et al.*, 2002).

Currently the LFG grammar is in the process of undergoing major c-structure changes. These changes have framework-independent motivation and are aimed at providing a better model of the interaction of some phenomena (such as coordination, negation and case assignment) and obtaining better f-structures. Therefore, rather than using original flat structure analyses in the relevant areas, the LFG grammar adopts a hierarchical structure which better accommodates relevant phenomena. It is perhaps worth noting that the choice of these analyses over the original ones was largely based on data from the NCP.

Last but not least, by contrast to GFJP2, which treats punctuation as a syntactic issue and consequently models it in its rules, the LFG grammar leaves the phenomenon of punctuation haplology at the discretion of the tokenizer, following the practice adopted in ParGram grammars.

3.2. Lexicon Creation

The morphosyntactic information necessary for the construction of a lexicon may be provided by Morfeusz, but it may also be extracted from manually disambiguated parse trees taken from the treebank (Składnica) or from the NCP (from the manually annotated subcorpus, for instance). Data obtained from any of these sources is passed on to part-of-speech templates which are bundles of calls to simple templates which set the values of appropriate features, etc.

There is a wide range of lexicalised information, mainly in the form of valence information which accompanies morphosyntactic data in the lexicon. The information in valence dictionaries supplied with the original grammar (GFJP2) was provided in the form of slots filled with categories having appropriate parameters (argument case, preposition form, complementiser type, etc). Since LFG

[4] Templates are bundles of f-descriptions; they provide a convenient means of expressing linguistic generalisations at various levels (in the lexicon, syntax).

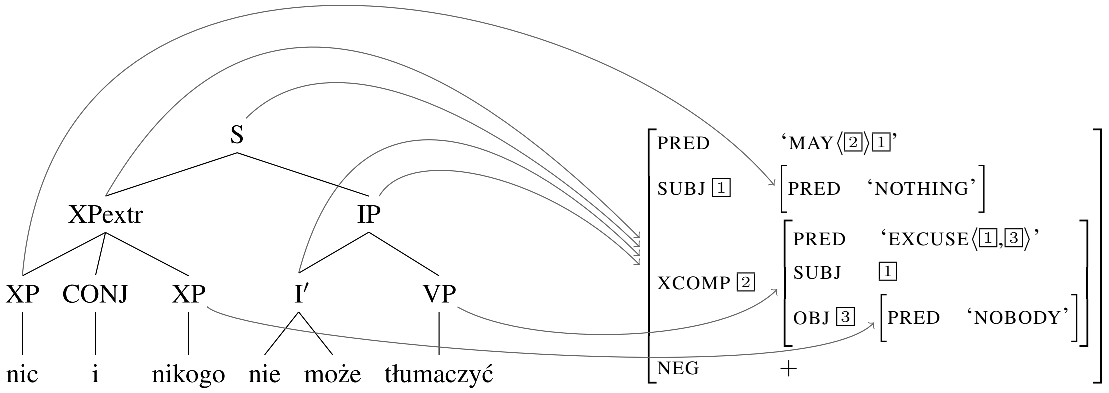

Figure 1: c-structure and f-structure representation of a sentence extracted from the NCP

defines valence requirements in terms of grammatical functions rather than c-structure categories, the original valence dictionaries for verbs, nouns and adjectives were automatically converted to an appropriate format.

Even though many verbs appearing in sentences in Składnica were not included in the original valence dictionary, entries for such predicates may be automatically extracted from their parse trees. This is due to the fact that the original grammar provided a wide range of parses using default frames from which human annotators selected the correct parse, choosing at the same time an appropriate valence frame.

3.3. Resulting Structures: an Example

Figure 1 provides sample structures (constituent and functional) representing the following attested sentence:

(1) nikogo i nic nie może tłumaczyć.
nobody.GEN and nothing.NOM NEG may excuse
(‘[. . .] nothing may excuse anybody.’)

It demonstrates that c- to f-structure mapping (arrows link particular positions in the c-structure with relevant fragments of the f-structure; boxed numbers indicate structure-sharing) may be very indirect in Polish. Under a special (yet at the same time very productive, as ample evidence available in the NCP suggests) variety of coordination, namely lexico-semantic coordination (Kallas, 1993), particular conjuncts may not only map to different grammatical functions but also belong to various levels of the f-structure. While the first conjunct, *nic*, is the subject (SUBJ) of the main clause, the second conjunct, *nikogo*, serves as the object (OBJ) of the infinitival complement (XCOMP).

4. Quality Control

The evaluation of the currently developed LFG grammar of Polish is performed against two independent measures: constructed testsuites and authentic sentences from the treebank. While the aim of the former is to ensure that the grammar correctly models particular linguistic phenomena, the latter checks its robustness, real-life coverage, as well as compatibility with the grammar which provided the original c-structure.

4.1. Constructed Testsuites

There are approximately 1 200 constructed testsuite sentences. More than 700 were designed specifically for the purposes of the present implementation while the remainder was provided by testsuites which were used in the development of earlier grammars. These include constructed sentences extracted from the source code of GFJP2 and elicited sentences which were used for testing the HPSG grammar of Polish (Marciniak *et al.*, 2003).

It is worth noting that, by contrast with treebank sentences, constructed testsuites are not limited to positive examples – almost half of these sentences are negative examples which are not supposed to be accepted by the grammar. While treebank testing is the main method of ensuring a reasonable overall coverage of the grammar, constructed testsuites provide an indispensable measure of ensuring the high quality of the linguistic analysis, making it possible to detect minute changes and identify potential problems as early and precisely as possible.

4.2. Treebank Testing

The other method of evaluation is treebank testing which takes the form of reparsing all the sentences currently available in Składnica for which human annotators identified a correct parse among the trees provided by the output of Świgra. The most recent results amount to 85% out of approximately 8 220 sentences (Składnica is under development and the number of trees available is growing steadily.)

The remaining 15% are mainly sentences which were not parsed due to the fact that the limit of available resources, time, in most cases, or memory, was exceeded. Such problematic sentences were subsequently parsed manually in fragments and the obtained c- and f-structures were inspected carefully. Fragments were chosen so as to constitute a representative subset of the original sentence. The results of this experiment suggest that the grammar would accept such sentences if it were not for issues related to resources. It seems that introducing further changes in the c-structure and limiting the reliance on f-structure constraints at the same time (through the use of more parameterised rules, for instance) could be a viable solution to this problem.

4.3. ParGram

Finally, Polish LFG structures created using the present grammar have recently taken part in the biannual ParGram structure comparison, an initiative to ensure cross-linguistic compatibility of a number of LFG grammars. The goal of the ParGram project is to develop parallel grammars by means such as sharing a common set of features which are used in the f-structures and attempting to use similar analyses of particular linguistic phenomena across various languages. Suggested modifications are currently being implemented in the grammar.

5. Conclusion and Future Outlook

New formal grammars are usually either created from scratch or read off a treebank. This paper presents a different approach to grammar construction, dubbed here "parasitic grammar development", where a grammar is based on an already existing grammar and extends it both "horizontally" and "vertically". Additionally, both a treebank and a testsuite are used to constantly control the quality of the developed grammar.

While at present the LFG grammar of Polish is not a probabilistic grammar, the underlying XLE formalism makes it possible to turn it into a probabilistic grammar by adding weights read off a treebank. Additionally, Optimality Theory mechanisms implemented in XLE make it possible to state constraints used to rank and categorise analyses. These are the most natural extensions of the work presented here.

Acknowledgements

This research is supported by the POIG.01.01.02-14-013/09 project which is co-financed by the European Union under the European Regional Development Fund.

References

Bresnan, J., editor (1982). *The Mental Representation of Grammatical Relations*. MIT Press Series on Cognitive Theory and Mental Representation. The MIT Press, Cambridge, MA.

Bresnan, J. (2000). *Lexical-Functional Syntax*. Blackwell Textbooks in Linguistics. Blackwell.

Dalrymple, M. (2001). *Lexical-Functional Grammar*. Academic Press.

Kallas, K. (1993). *Składnia współczesnych polskich konstrukcji współrzędnych*. Wydawnictwo Uniwersytetu Mikołaja Kopernika, Toruń.

Marciniak, M., Mykowiecka, A., Przepiórkowski, A., and Kupść, A. (2003). An HPSG-annotated test suite for Polish. In A. Abeillé, editor, *Treebanks: Building and Using Parsed Corpora*, volume 20 of *Text, Speech and Language Technology*, pages 129–146. Kluwer, Dordrecht.

Pollard, C. and Sag, I. A. (1987). *Information-Based Syntax and Semantics, Volume 1: Fundamentals*. Number 13 in CSLI Lecture Notes. CSLI Publications, Stanford, CA.

Pollard, C. and Sag, I. A. (1994). *Head-driven Phrase Structure Grammar*. Chicago University Press / CSLI Publications, Chicago, IL.

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

Przepiórkowski, A. (2009). A comparison of two morphosyntactic tagsets of Polish. In V. Koseska-Toszewa, L. Dimitrova, and R. Roszko, editors, *Representing Semantics in Digital Lexicography: Proceedings of MONDILEX Fourth Open Workshop*, pages 138–144, Warsaw.

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

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

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.

Świdziński, M. (1992). *Gramatyka formalna języka polskiego*, volume 349 of *Rozprawy Uniwersytetu Warszawskiego*. Wydawnictwa Uniwersytetu Warszawskiego, Warsaw.

Świdziński, M. and Woliński, M. (2010). Towards a bank of constituent parse trees for Polish. In *Text, Speech and Dialogue: 13th International Conference, TSD 2010, Brno, Czech Republic*, Lecture Notes in Artificial Intelligence, pages 197–204, Berlin. Springer-Verlag.

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

Woliński, M. (2004). *Komputerowa weryfikacja gramatyki Świdzińskiego*. Ph.D. dissertation, Instytut Podstaw Informatyki, Polska Akademia Nauk, Warsaw.

Woliński, M. (2006). Morfeusz — a Practical Tool for the Morphological Analysis of Polish. In M. Kłopotek, S. T. Wierzchoń, and K. Trojanowski, editors, *Intelligent information processing and web mining*, pages 503–512. Springer-Verlag.

Woliński, M. (2010). Dendrarium — an Open Source Tool for Treebank Building. In M. Kłopotek, M. Marciniak, A. Mykowiecka, W. Penczek, and S. T. Wierzchoń, editors, *Intelligent Information Systems*, pages 193–204. Wydawnictwo Akademii Podlaskiej.