Synergistic development of grammatical resources: a valence dictionary, an LFG grammar, and an LFG structure bank for Polish

Agnieszka Patejuk and Adam Przepiórkowski

Institute of Computer Science, Polish Academy of Sciences
E-mail: {aep,adamp}@ipipan.waw.pl

Abstract

The aim of this paper is to present the simultaneous development of three interrelated linguistic resources for Polish: a valence dictionary, a formal grammar, and a syntactic corpus. This parallel development creates a strong synergistic effect: the valence dictionary is constantly verified by its use in the formal grammar, and both are verified in the process of constructing the structure bank.

1 Introduction

The aim of this paper[1] is to introduce a new linguistic resource of Polish and discuss the role it plays in the verification of the quality and completeness of two other resources. The new resource is a corpus of sentences annotated with LFG syntactic structures: the usual constituency trees and functional structures bearing information about grammatical functions of various constituents in the tree, about their morphosyntactic features, and about the predicates introduced by the heads of these constituents.

The other two resources – the valence dictionary and the LFG grammar – were originally developed relatively independently. However, once the process of converting the valence dictionary to an LFG lexicon started, many inconsistencies and gaps in the dictionary were discovered; since then, the employment of the dictionary in the grammar has been the source of additional quality and completeness control of the dictionary, with a constant flow of bug reports and data requests.

The third resource, an LFG structure bank (presented here for the first time), is empirically based on an earlier constituency treebank, but the LFG structures are constructed independently of those in that treebank: sentences are parsed with

[1]The work described here was partially financed by the CLARIN-PL project (http://clip.ipipan.waw.pl/CLARIN-PL).

an XLE parser implementing the LFG grammar and then manually disambiguated using the INESS tool. In the process of manual disambiguation, problems in the LFG grammar and in the valence dictionary are discovered and corrected, leading to new versions of both resources.

The main thesis of this paper is that the parallel development of such resources is preferable to the usual – and often unavoidable for practical reasons – procedure of, say, developing a valence dictionary on the basis of a closed treebank, as the flow of information to upstream resources is considerable and leads to massive improvements.

The paper is structured as follows: §2 presents the three resources, §3 discusses the synergy effect during their parallel development and §4 concludes the paper.

2 Resources

2.1 *Walenty*, a valence dictionary of Polish[2]

One of the two initial resources is the valence dictionary *Walenty*, presented in detail elsewhere [22, 19], so we will only illustrate it with a couple of valence schemata.

A simplified example of an entry for the verb ADMINISTROWAĆ 'administrate' is given below:

(1) `administrować: imperf: subj{np(str)} + obj{np(inst)}`

This is an imperfective verb and the schema specifies two arguments: the subject and a passivisable[3] object (*somebody administrates something*). They are both nominal phrases (NPs), and while the object bears the fixed instrumental case, the subject's case is specified as structural, as its morphological realisation depends on the category of the head assigning case (gerunds uniformly assign genitive case to their subject) and – at least on some theories [16] – on the category of the subject, namely whether it is a certain kind of numeral phrase (in which case it is accusative) or not (in which case it is nominative).

A slightly more complex schema is needed for the verb DEDYKOWAĆ 'dedicate', as used in the following example from the National Corpus of Polish:

(2) Gola dedykuję [dla rodziców] i [sympatii Iwonie].
goal.ACC dedicate for parents.GEN and girlfriend.DAT Iwona.DAT
'I dedicate this goal to my parents and my girlfriend Iwona.' (NKJP)

In the above example, the first person subject is *pro*-dropped, the pre-verbal object occurs in the accusative case, but it would occur in the genitive if the verb were

[2]This section provides information about the version of *Walenty* of 20 November 2014. However, entries from plain text export, (1) and (3), were simplified by reducing them to the following fields: lemma, aspect, valence schema. The remaining ones are not directly relevant to the discussion.

[3]The label `obj` is used to mark arguments which can become the subject under passive voice.

negated, so its case is marked as structural in the valence schema, and there is one more argument, whose grammatical function is not marked explicitly in the schema below:

(3) `dedykować: imperf:`
`subj{np(str)} + obj{np(str)} + {np(dat); prepnp(dla,gen)}`

This argument must be specified as being realisable by two kinds of phrases: a dative NP or a prepositional phrase (PP; `prepnp` above) headed by the preposition DLA 'for' and including a genitive NP. The fact that these two kinds of phrases occupy the same argument position follows from the possibility to coordinate them, as in (2) above.

These two valence schemata illustrate only a couple of a number of interesting or unique features of *Walenty*. First of all, as already pointed out above, it explicitly defines an argument position via the coordination test, so one position in one valence schema may be filled by categorially diverse constituents, as in the famous English example *Pat became a republican and quite conservative* [27, p. 142], where the noun phrase *a republican* is coordinated with the adjectival phrase *quite conservative* within an argument position of *became*. It turns out that such coordination of unlike categories is relatively common in Polish.

Second, *Walenty* – while remaining relatively theory-neutral – is informed by contemporary linguistic theories and encodes linguistic facts often ignored in other valence dictionaries, e.g., control and raising, structural case, passivisation, non-chromatic arguments [15], etc.

Third, the dictionary contains a very rich phraseological component [18] which makes it possible to precisely describe lexicalised arguments and idiomatic constructions, e.g., the fact that one may welcome (Pol.: *witać*) somebody "with open arms" (Pol.: *z otwartymi ramionami*) or "with arms wide open" (Pol.: *z szeroko otwartymi ramionami*), but not just "with arms" (Pol.: **z ramionami*) or "with unusually wide open arms" (Pol.: **z niezwykle szeroko otwartymi ramionami*).

Fourth, while the process of adding deep semantic information to *Walenty* has begun only recently, some arguments are already defined semantically, e.g., the manner arguments as occurring with the verbs ZACHOWYWAĆ SIĘ 'behave (in some way)' or TRAKTOWAĆ 'treat (somebody in some way)' – such arguments may be realised via adverbial phrases of a certain kind, but also via appropriate prepositional or sentential phrases.

Finally, the dictionary, continually developed within various projects, is already the biggest and most detailed valence dictionary of Polish: as of 20 November 2014, it contains 54 328 schemata for 11 913 lemmata. Moreover, by the end of 2015 *Walenty* is planned to cover 15 000 lemmata, including at least 3000 non-verbal ones. Snapshots of the dictionary are released on an open source licence roughly half-yearly; see `http://zil.ipipan.waw.pl/Walenty`.

2.2 *POLFIE*, an LFG grammar of Polish

POLFIE [13] is an LFG [1, 6] grammar of Polish implemented in XLE [4]. As described in more detail in [13], rules used in *POLFIE* were written on the basis of two previous formal grammars of Polish: the DCG [30] grammar *GFJP2* used by the parser *Świgra* [31] and the HPSG [14] grammar described in [20]. While the former provided the basis for constituent structure rules, the latter was used as the basis of f-descriptions. The basis provided by these previous grammars was the starting point for extensions which were introduced in areas such as coordination and agreement (see, e.g., various publications by the current authors in proceedings of LFG conferences 2012–2014; `http://cslipublications.stanford.edu/LFG`).

Also the lexicon of *POLFIE* is heavily based on other resources. Morphosyntactic information is drawn from a state-of-the-art morphological analyser of Polish, *Morfeusz* [32, 33], from the *National Corpus of Polish* (*NKJP*; [17]) and from *Składnica*, a treebank of parses produced by the *Świgra* parser [29, 34]. While some (very limited) syntactic information is added manually to selected lexical entries – e.g., those of *wh*-words (such as *kto* 'who' or *dlaczego* 'why'), *n*-words (such as *nikt* 'nobody', *nigdy* 'never' or *żaden* 'none'), etc. – valence information is automatically converted from *Walenty*. For example, the schema for ADMINISTROWAĆ 'administrate' in (1) is converted to an XLE entry whose simplified version is given below:

(4)
```
(^ PRED)='administrować<(^ SUBJ)(^ OBJ)>'
(^ SUBJ PRED:
        {(<- CASE)=c nom | (<- CASE)=c acc (<- ACM)=c rec})
(^ OBJ CASE)=c inst
(^ TNS-ASP ASP)=c imperf
```

The first line of this lexical entry specifies the so-called semantic form of the verb, i.e., that the predicate is `administrować` and that it takes two arguments: SUBJ and OBJ. The last line says that it is an imperfective verb, and the penultimate line – that the case of its object is instrumental. The subject specification, split into two lines for typographic reasons, is more complex: it says that the subject is either in the nominative case or it is a governing numeral (see the slightly cryptic `(<- ACM)=c rec`) in the accusative.[4,5]

The XLE system with *POLFIE* parses around a third of sentences in the 1-million-word manually annotated balanced subcorpus [7] of *NKJP*. This may sound like a poor result, but it is typical of deep parsers not propped with any fall-back pre-processing or post-processing strategies. (Such supporting strategies are currently being developed for *POLFIE*.)

[4]Such numerals are assumed to be defective and have no nominative form, so no ambiguity follows from the fact that the first disjunct is not specified as not being a governing numeral.

[5]This information is encoded via the mechanism of off-path constraints [6, p. 148] for reasons explained in detail in [12].

2.3 An LFG structure bank of Polish sentences

The structure bank of Polish sentences is the youngest of these resources, and it is presented here for the first time. It is based on the aforementioned *Składnica* treebank, but only in a weak sense: the same morphosyntactically annotated Polish sentences – originally drawn from the 1-million-word subcorpus of *NKJP* – are assigned syntactic structures here, but these structures are not based on those in *Składnica*. This way interesting cross-theoretical comparisons should be possible in the future between the DCG representations contained in *Składnica* and the LFG representations in the structure bank described in this section.

The resource currently contains almost 6 500 sentences (over 58 000 segments, in the *NKJP* sense of this term). It has been created semi-automatically. First, the sentences were parsed using the *POLFIE* grammar and the XLE system mentioned above. In effect, often multiple analyses were produced for many sentences, since any grammar of a reasonable size must be ambiguous; in case of *POLFIE*, the average number on parses is 717 and the median is 10. (This means that there are a few sentences with a very large number of parses and many with very few analyses.) After this automatic process, analyses were manually disambiguated by a group of linguists – each sentence independently by two linguists, to ensure the high quality of the resulting structure bank.[6] 4 linguists spent 4 half-time months each (i.e., 2 person-months) on the task, 1 spent 1 half-time month, and all of them spent some 2–3 half-time months on learning LFG and the disambiguation system used for this task. During annotation, the linguists were not allowed to individually communicate or to see each other's comments. On the other hand, they could communicate via a mailing list accessible to all of them and to the developers of the grammar. The process was supervised by the main grammar writer (the first author), who responded to all questions and many comments.

This high speed of annotation could be attained thanks to the use of the INESS infrastructure for building structure banks [23, 25]. Figure 1 (on the next page) presents a screenshot of the system for the sentence *Jak wygląda przepiórka* 'What does a quail look like?', lit. 'How looks quail?', before it is disambiguated. Both the c-structure and the f-structure are shown in a compact format encompassing a number of analyses (here, two) at the same time. For example, in the c-structure in the middle of the screenshot, the choice is at the level of the highest IP node: should it be rewritten to ADVP IP (the analysis marked as [a2]) or to IP XPsem (analysis [a1], with the order of nodes reversed, as the lower IP is shared between these two analyses)? The correct parse may be selected by the annotator by clicking on one of the two rules in the bottom left corner of the screenshot: IP –> XPsem IP or IP –> ADVP IP.

This choice at the level of c-structure is correlated with a choice at the level of f-structure. For example, the f-structure will contain the feature ADJUNCT only if a2 is selected. Otherwise, if a1 is chosen, it will contain the feature OBL-MOD.

[6] As in case of the manual annotation of *NKJP* [21], pairs of annotators were not constant; instead annotators were shuffled so as to avoid co-learning the same mistakes.

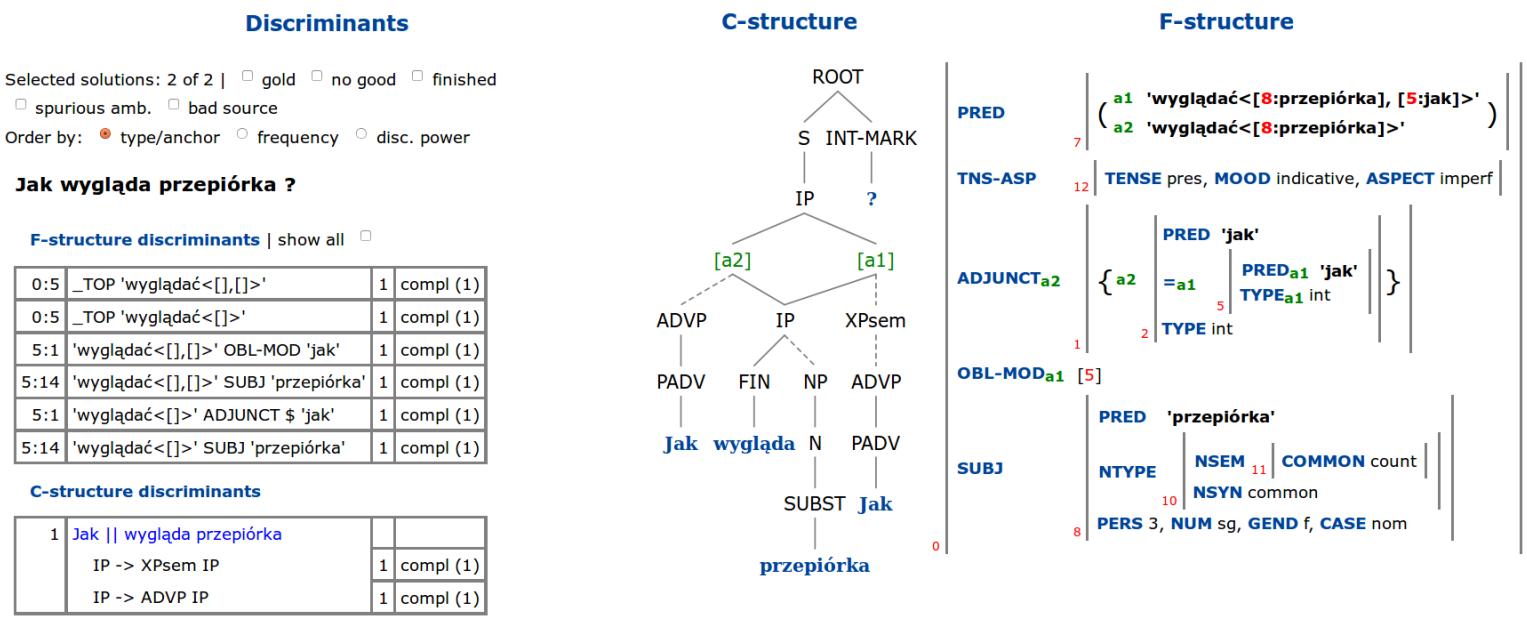

Figure 1: *Jak wygląda przepiórka* before disambiguation

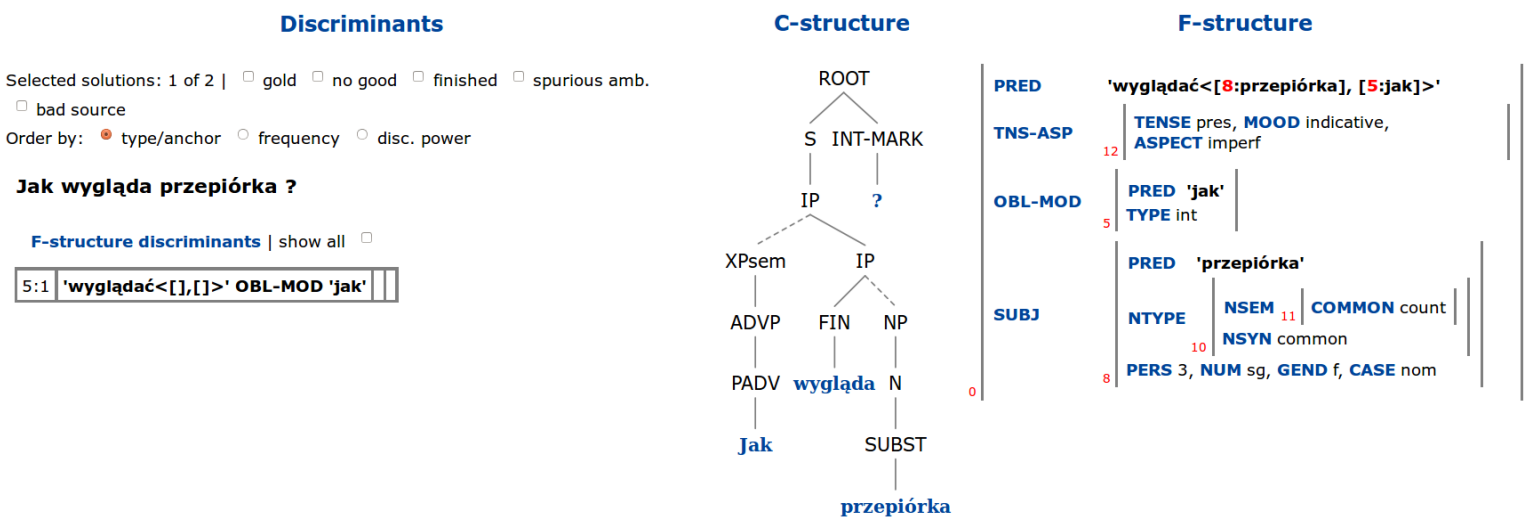

Figure 2: *Jak wygląda przepiórka* after disambiguation

So, instead of relying on c-structure discriminants in the table at the bottom left of Figure 1, annotators may rely on f-structure discriminants in the table above, and select either the third row of the table, mentioning OBL-MOD 'jak', or the fifth row, mentioning ADJUNCT $ 'jak'. In fact, the choice boils down to whether the verb WYGLĄDAĆ 'look like' is a two-argument verb (see the first row in this table) or a one-argument verb (see the second row). As the first of these options seems correct, the annotator may disambiguate this sentence by clicking on the first row or – equivalently – on the third row. The result of choosing the latter discriminant is shown in Figure 2 (on the previous page).

3 Synergy effect with parallel development

As should be clear from the above descriptions of the three Polish resources, the valence dictionary feeds the formal grammar, which is in turn used to build the structure bank. Work on each of these resources also results in the verification and significant improvements of the upstream resources.

First, *Walenty* is automatically converted to LFG constraints to be used in the grammar, and many inconsistencies in the valence dictionary can be identified already at this stage. During this process, morphosyntactic information stored in *Walenty* is compared with information provided by the morphological analyser *Morfeusz*, which makes it possible to discover such problems as wrong aspect of the predicate, wrong case required by the preposition, etc. More importantly, potentially problematic schemata are also discovered, e.g., ones containing no subject when a passivisable object is present or ones with mismatched control relations.

Second, omissions in the valence dictionary are identified when the resulting grammar is used for parsing a corpus of Polish sentences. Analysed sentences are inspected and, if the lack of correct parses results from the incompleteness of *Walenty*, new schemata are added to the dictionary.

Third, sentences parsed with XLE are fed into INESS (including sentences for which XLE returned no good parse: there were over 9 000 sentences). Those sentences which have syntactic analyses (there are over 6 500 such sentences, see §2.3) are disambiguated. The annotators are encouraged to look at f-structure discriminants rather than c-structure discriminants and, especially, at values of PRED, which contain information about the number and type of arguments of particular predicates. This way wrong valence in f-structures is discovered, which may be caused by errors in the *Walenty*-to-LFG conversion procedure, but is more often caused by problems in the valence dictionary itself. Of course, other errors in f-structures are also spotted, relating directly to the grammar. This way, the construction of the structure bank verifies both the formal grammar and the valence dictionary.

Error reports during the construction of the structure bank are facilitated by the rich system of comments offered by INESS. For this task, there are three main types of comments: issue, todo and bad_interp. The last is reserved for reports on wrong morphosyntactic annotation of some words, i.e., it is concerned with the

Składnica treebank and the *NKJP* subcorpus from which the morphosyntactically annotated sentences are taken. This way, the development of the LFG structure bank also influences resources other than the valence dictionary and the grammar. Problems with valence constitute one subtype of todo comments, other subtypes are concerned with the grammar. Finally, comments of type issue signal more subtle problems, e.g., doubts about the proper attachment place of a constituent, doubts about the choice of a grammatical function for an argument, a multi-word expression which should probably have a separate entry in the dictionary, etc. It should be noted that annotators are encouraged to leave comments to suboptimal analyses even when one of the analyses of the sentence is fully correct. Currently, there are almost 3 000 comments in the system.

The whole annotation process is divided into rounds, each involving around 1 000 sentences and lasting 2–3 weeks. After a round of annotation is completed, comments created by annotators are inspected by the grammar writer, who responds to each of them (after they have been anonymised) using the mailing list. The purpose of this review is to give feedback to annotators: explain some analyses, improve their skills by making them aware of certain linguistic issues, encourage them to contribute comments.

Subsequently, relevant comments containing confirmed issues are passed together with responses (and additional comments, if needed) to the developers of relevant resources. Developers of *Walenty* are asked to inspect relevant entries and introduce appropriate changes, if the suggestion is right. Issues related to the conversion are handled by the grammar writer. Finally, comments related to problems in the grammar are collected and passed to the grammar writer to introduce appropriate modifications to improve the treatment of relevant phenomena.

After relevant changes have been introduced in *Walenty* and the grammar, a new lexicon is created, sentences are reparsed and a new version of analyses is fed into INESS so that discriminants can be reapplied from the previous disambiguated version of the structure bank. This takes advantage of an ingenious feature of INESS, based on an idea of [3] and on earlier work on the LinGO Redwoods HPSG treebank [10, 11]: choices made for one version of the grammar remain valid for the next version of the grammar. After discriminants have been reapplied, annotators are asked to return to those sentences which did not have a complete good solution in the previous version, consult their comments and check if the relevant problem is solved in the current version.

The entire procedure described above is repeated until a good solution is obtained for all the sentences. As a result, all three resources, the valence dictionary, the formal grammar and the structure bank, are improved incrementally in parallel, as illustrated in Figure 3.

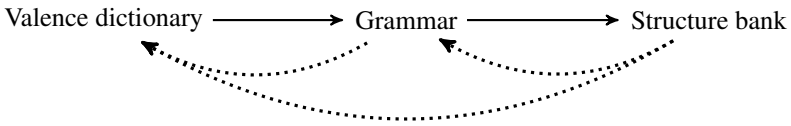

Figure 3: Flow of information to downstream resources (straight solid arrows) and feedback to upstream resources (curved dotted arrows)

4 Conclusion

Evaluation of the quality and completeness of valence dictionaries is difficult. By the concurrent development of a relatively theory-independent dictionary and a comprehensive LFG grammar taking advantage of almost all types of information in this dictionary, the quality of the dictionary is partially verified. By applying the grammar to a relatively balanced corpus of Polish, both the quality and the completeness of the dictionary – as well as the quality of the grammar – are verified.

Obviously, this is not the first attempt at the parallel development of language resources. Grammars have been developed together with treebanks, e.g., the HPSG grammar of English [10, 11] or the LFG grammar of Norwegian [26], as well as – more recently – the DCG grammar of Polish [29, 34]. Similarly, for Czech, valence dictionaries have been extracted from treebanks automatically [28] or developed manually in sync with treebank construction [8]; see also [9] for similar work on German and [24] for a discussion of various improvements of a Norwegian lexicon when constructing a Norwegian parsebank (treebank based on automatic parsing and manual disambiguation). The current setup extends such work by showing the benefits of the simultaneous developments of three different resources, including an independent valence dictionary, not based on the linguistic theory underlying the grammar and the structure bank. While this approach to the parallel development of multiple linguistic resources is often difficult – due to the scarcity of non-linguistic resources (budgetary and human) – we maintain that such a holistic approach should always be strived for.

Our future plans extend this approach even further and involve the addition of semantic information to the valence dictionary, and subsequent verification of this information via the use of semantic representations produced by the grammar based on this extended dictionary in the task of recognising textual entailment ([5]).

References

[1] Joan Bresnan. *Lexical-Functional Syntax*. Blackwell Textbooks in Linguistics. Blackwell, Malden, MA, 2001.

[2] Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios

Piperidis, editors. *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, Reykjavík, Iceland, 2014. ELRA.

[3] David Carter. The TreeBanker: A tool for supervised training of parsed corpora. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence*, pages 598–603, Providence, RI, 1997.

[4] Dick Crouch, Mary Dalrymple, Ron Kaplan, Tracy King, John Maxwell, and Paula Newman. XLE documentation. `http://www2.parc.com/isl/groups/nltt/xle/doc/xle_toc.html`, 2011.

[5] Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. *Recognizing Textual Entailment: Models and Applications*. Morgan & Claypool, 2013.

[6] Mary Dalrymple. *Lexical Functional Grammar*. Academic Press, 2001.

[7] Łukasz Degórski and Adam Przepiórkowski. Ręcznie znakowany milionowy podkorpus NKJP. In Przepiórkowski et al. [17], pages 51–58.

[8] Jan Hajič, Jarmila Panevová, Zdeňka Urešová, Alevtina Bémová, Veronika Kolářová, and Petr Pajas. PDT-VALLEX: Creating a large-coverage valency lexicon for treebank annotation. In Joakim Nivre and Erhard Hinrichs, editors, *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT 2003)*, Växjö, Norway, 2003.

[9] Erhard W. Hinrichs and Heike Telljohann. Constructing a valence lexicon for a treebank of German. In Frank van Eynde, Anette Frank, Koenraad De Smedt, and Gertjan van Noord, editors, *Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories (TLT 7)*, pages 41–52, Groningen, The Netherlands, 2009.

[10] Stephan Oepen, Dan Flickinger, Kristina Toutanova, and Christoper D. Manning. LinGO Redwoods: A rich and dynamic treebank for HPSG. In Erhard Hinrichs and Kiril Simov, editors, *Proceedings of the First Workshop on Treebanks and Linguistic Theories (TLT2002)*, pages 139–149, Sozopol, 2002.

[11] Stephan Oepen, Dan Flickinger, Kristina Toutanova, and Christoper D. Manning. LinGO Redwoods: A rich and dynamic treebank for HPSG. *Research on Language and Computation*, 4(2):575–596, 2004.

[12] Agnieszka Patejuk and Adam Przepiórkowski. A comprehensive analysis of constituent coordination for grammar engineering. In *Proceedings of the 24rd International Conference on Computational Linguistics (COLING 2012)*, 2012.

[13] Agnieszka Patejuk and Adam Przepiórkowski. Towards an LFG parser for Polish: An exercise in parasitic grammar development. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012*, pages 3849–3852, Istanbul, Turkey, 2012. ELRA.

[14] Carl Pollard and Ivan A. Sag. *Head-driven Phrase Structure Grammar*. Chicago University Press / CSLI Publications, Chicago, IL, 1994.

[15] Paul M. Postal. *Skepical Linguistic Essays*, chapter Chromaticity: An overlooked English grammatical category distinction, pages 138–158. Oxford University Press, 2004.

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

[17] Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk, editors. *Narodowy Korpus Języka Polskiego [Eng.: National Corpus of Polish]*. Wydawnictwo Naukowe PWN, Warsaw, 2012.

[18] Adam Przepiórkowski, Elżbieta Hajnicz, Agnieszka Patejuk, and Marcin Woliński. Extended phraseological information in a valence dictionary for NLP applications. In *Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing (LG-LP 2014)*, pages 83–91, Dublin, Ireland, 2014. Association for Computational Linguistics and Dublin City University.

[19] Adam Przepiórkowski, Elżbieta Hajnicz, Agnieszka Patejuk, Marcin Woliński, Filip Skwarski, and Marek Świdziński. Walenty: Towards a comprehensive valence dictionary of Polish. In Calzolari et al. [2], pages 2785–2792.

[20] Adam Przepiórkowski, Anna Kupść, Małgorzata Marciniak, and Agnieszka Mykowiecka. *Formalny opis języka polskiego: Teoria i implementacja [Eng.: Formal description of Polish: Theory and implementation]*. Akademicka Oficyna Wydawnicza EXIT, Warsaw, 2002.

[21] Adam Przepiórkowski and Grzegorz Murzynowski. Manual annotation of the National Corpus of Polish with Anotatornia. In Stanisław Goźdź-Roszkowski, editor, *Explorations across Languages and Corpora: PALC 2009*, pages 95–103, Frankfurt am Main, 2011. Peter Lang.

[22] Adam Przepiórkowski, Filip Skwarski, Elżbieta Hajnicz, Agnieszka Patejuk, Marek Świdziński, and Marcin Woliński. Modelowanie własności składniowych czasowników w nowym słowniku walencyjnym języka polskiego. *Polonica*, XXXIII:159–178, 2014.

[23] Victoria Rosén, Paul Meurer, and Koenraad De Smedt. Designing and implementing discriminants for LFG grammars. In Miriam Butt and Tracy Holloway King, editors, *The Proceedings of the LFG'07 Conference*, pages 397–417, University of Stanford, California, USA, 2007. CSLI Publications.

[24] Victoria Rosén, Petter Haugereid, Martha Thunes, Gyri S. Losnegaard, and Helge Dyvik. The interplay between lexical and syntactic resources in incremental parsebanking. In Calzolari et al. [2], pages 1617–1624.

[25] Victoria Rosén, Paul Meurer, Gyri Smørdal Losnegaard, Gunn Inger Lyse, Koenraad De Smedt, Martha Thunes, and Helge Dyvik. An integrated web-based treebank annotation system. In Iris Hendrickx, Sandra Kübler, and Kiril Simov, editors, *Proceedings of the Eleventh International Workshop on Treebanks and Linguistic Theories (TLT 11)*, pages 157–168, Lisbon, Portugal, 2012.

[26] Victoria Rosén, Paul Meurer, and Koenraad De Smedt. Constructing a parsed corpus with a large LFG grammar. In Miriam Butt and Tracy Holloway King, editors, *The Proceedings of the LFG'05 Conference*, pages 371–387, University of Bergen, Norway, 2005. CSLI Publications.

[27] Ivan A. Sag, Gerald Gazdar, Thomas Wasow, and Steven Weisler. Coordination and how to distinguish categories. *Natural Language and Linguistic Theory*, 3:117–171, 1985.

[28] Anoop Sarkar and Daniel Zeman. Automatic extraction of subcategorization frames for Czech. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, pages 691–697, Saarbrücken, 2000.

[29] Marek Świdziński and Marcin Woliński. 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, 2010. Springer-Verlag.

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

[31] Marcin Woliński. *Komputerowa weryfikacja gramatyki Świdzińskiego [Eng.: A computational verification of Świdziński's grammar]*. Ph.D. dissertation, Institute of Computer Science, Polish Academy of Sciences, Warsaw, 2004.

[32] Marcin Woliński. Morfeusz — a Practical Tool for the Morphological Analysis of Polish. In Mieczysław Kłopotek, Sławomir T. Wierzchoń, and Krzysztof Trojanowski, editors, *Intelligent information processing and web mining*, pages 503–512. Springer-Verlag, 2006.

[33] Marcin Woliński. Morfeusz reloaded. In Calzolari et al. [2], pages 1106–1111.

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