AGNIESZKA PATEJUK
ADAM PRZEPIÓRKOWSKI
Instytut Podstaw Informatyki
Polskiej Akademii Nauk
ul. Jana Kazimierza 5
01-248 Warszawa
tel.: +48 223 800 500
e-mail: aep@ipipan.waw.pl, adamp@ipipan.waw.pl

PARALLEL DEVELOPMENT OF LINGUISTIC RESOURCES: TOWARDS A STRUCTURE BANK OF POLISH

SŁOWA KLUCZOWE: słownik walencyjny, gramatyka formalna, korpus składniowy, Lexical Functional Grammar

KEYWORDS: valence dictionary, formal grammar, syntactic corpus, LFG

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 improvement of the quality and completeness of two other resources. The new resource is a corpus of sentences annotated with two kinds of linguistic structures: the usual constituency trees and so-called functional structures. The latter are nested feature structures (also called attribute-value matrices or AVMs) 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.

[1] This paper is a revised and extended version of Patejuk and Przepiórkowski 2014. The work described here was partially financed by the CLARIN-PL project (http://clip.ipipan.waw.pl/CLARIN-PL). We gratefully acknowledge the comments of two anonymous reviewers of *Prace Filologiczne*, which led to some improvements of this article.

Figures 1a and 1b (on the next page) schematically illustrate these two kinds of structures for the sentence *Lingwiści tworzą zasoby językowe* 'Linguists create language resources': the top subfigure is the constituent structure (or c-structure) and the bottom one is the functional structure (or f-structure). The two structures are related as indicated by the boxed numbers: the whole f-structure – labelled with 0 – does not only correspond to the whole c-structure, but also to other projections of the main verb. Moreover, the f-substructure labelled as 1 corresponds to nominal c-structure nodes projecting from *Lingwiści* 'linguists', the f-substructure 2 corresponds to nominal nodes projecting from *zasoby* 'resources', and 3 corresponds to adjectival nodes projecting from *językowe* 'linguistic (referring to language, not the discipline of linguistics)'.

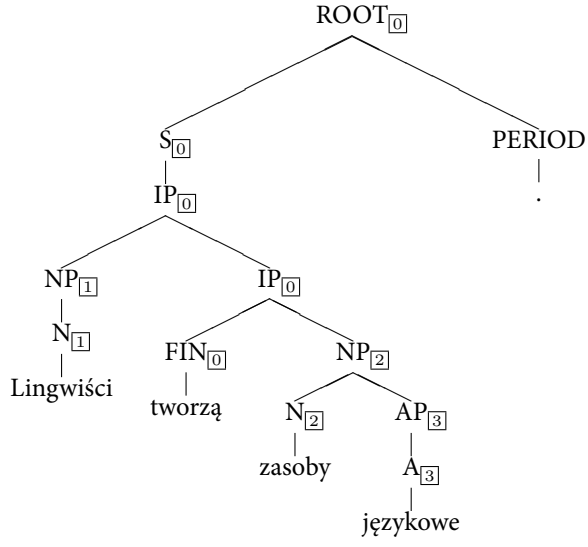
A corpus of Polish annotated with such trees and AVMs is presented in section *Structure bank* below. The term *structure bank* – used e.g. in Frank *et al.* 2003 – parallels the widely accepted term *treebank*, referring to collections of sentences annotated with syntactic trees. The main linguistic reason for developing this structure bank is the improvement of two more basic resources: the valence dictionary *Walenty* and the formal grammar *POLFIE*, both described in the following two sections. The details of the parallel development procedure are given in section *Improvements*, section *Related work* compares this procedure to similar projects for other languages, and *Conclusion* summarises the main points of this paper.

2. Valence dictionary

One of the two resources to be verified and improved is the valence dictionary *Walenty*, presented in detail elsewhere (Przepiórkowski *et al.* 2014b, c, Hajnicz *et al.* 2015), so we describe it here only briefly.

The dictionary has a number of interesting or unique features. First of all, 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* (Sag *et al.* 1985: 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, nonchromatic arguments, etc.



(a) c-structure

$$
\left[\begin{array}{ll}
\textsc{pred} & \text{`TWORZYĆ}\langle\boxed{1},\boxed{2}\rangle\text{'} \\
\textsc{subj}\ \boxed{1} & \left[\begin{array}{ll}
\textsc{pred} & \text{`LINGWISTA'} \\
\textsc{case} & \textsc{nom} \\
\textsc{gend} & \textsc{m1} \\
\textsc{num} & \textsc{pl}
\end{array}\right] \\
\boxed{0} & \left[\begin{array}{ll}
\textsc{pred} & \text{`ZASÓB'} \\
\textsc{case} & \textsc{acc} \\
\textsc{gend} & \textsc{m3} \\
\textsc{num} & \textsc{pl} \\
\textsc{obj}\ \boxed{2} & \\
\textsc{adj} & \left\{\left[\begin{array}{ll}
\textsc{pred} & \text{`JĘZYKOWY'} \\
\textsc{case} & \textsc{acc} \\
\boxed{3}\ \textsc{gend} & \textsc{m3} \\
\textsc{num} & \textsc{pl} \\
\textsc{degree} & \textsc{positive}
\end{array}\right]\right\}
\end{array}\right] \\
\textsc{tns-asp} & \left[\begin{array}{ll}
\textsc{tense} & \textsc{pres} \\
\textsc{aspect} & \textsc{imperf} \\
\textsc{mood} & \textsc{ind}
\end{array}\right]
\end{array}\right]
$$

(b) f-structure

Fig. 1. Two structures for "Lingwiści tworzą zasoby językowe."

Third, the dictionary contains a very rich phraseological component (Przepiórkowski *et al.* 2014a) 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 early June 2015, it contains over 67000 schemata for nearly 14000 lemmata. Moreover, by early 2016 *Walenty* is planned to cover 15000 lemmata, including at least 3000 non-verbal ones. The dictionary is publicly available at http://walenty.ipipan.waw.pl/. Snapshots of the dictionary are released on an open source licence roughly half-yearly; see http://zil.ipipan.waw.pl/Walenty.

3. Formal grammar

While *Walenty* does not assume any single linguistic theory, but rather takes from various structuralist and generative approaches, the formal grammar *POLFIE* follows the tradition of Lexical Functional Grammar (LFG; Bresnan 2001, Dalrymple 2001). LFG, originally developed in late 1970s and early 1980s (Bresnan 1982), belongs to the family of non-transformational constraint-based generative theories. Unlike transformational grammars, associated with the name of Noam Chomsky, LFG assumes a number of parallel representations of various linguistic levels of any utterance, of which c-structure and f-structure – exemplified above – are the main syntactic structures.

POLFIE is encoded in such a way that appropriate software – the XLE system (Crouch *et al.* 2011) – may read it and automatically produce representations of Polish sentences which are modelled by this grammar. As described in more detail in Patejuk and Przepiórkowski 2012, rules used in *POLFIE* were written on the basis of two previous formal grammars of Polish: the DCG (Warren and Pereira 1980) grammar *GFJP2* (Świdziński 1992, Świdziński and Woliński 2010) used by the parser *Świgra* (Woliński 2004) and the HPSG (Pollard and Sag 1994) gram-

mar described in Przepiórkowski *et al.* 2002. 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. the publications by Agnieszka Patejuk and Adam Przepiórkowski 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* (Woliński 2006, 2014), from the *National Corpus of Polish* (*NKJP*; Przepiórkowski *et al.* 2012) and from *Składnica*, a treebank of parses produced by the *Świgra* parser (Świdziński and Woliński 2010, Woliński *et al.* 2011). Valence information is taken from *Walenty* – its schemata are automatically converted into LFG representations. Finally, some information is manually added 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.

The XLE system with *POLFIE* currently parses around a third of sentences in the 1-million-word manually annotated balanced subcorpus (Degórski and Przepiórkowski 2012) 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. Structure bank

The structure bank of Polish sentences is the youngest of these resources, and it is presented here for the first time (but cf. fn. 1). 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 6500 sentences (over 58000 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 just a handful of 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.[2] 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 intensively supervised by the chief grammar writer, 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 (Rosén *et al.* 2007, 2012). Figure 2 (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. So, instead of relying on c-structure discriminants in the table at the bottom left corner of this figure, annotators may rely on f-structure discriminants in the table above it, 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 3.

[2] As in case of the manual annotation of *NKJP* (Przepiórkowski and Murzynowski 2011), pairs of annotators were not constant; instead annotators were shuffled so as to avoid co-learning the same mistakes.

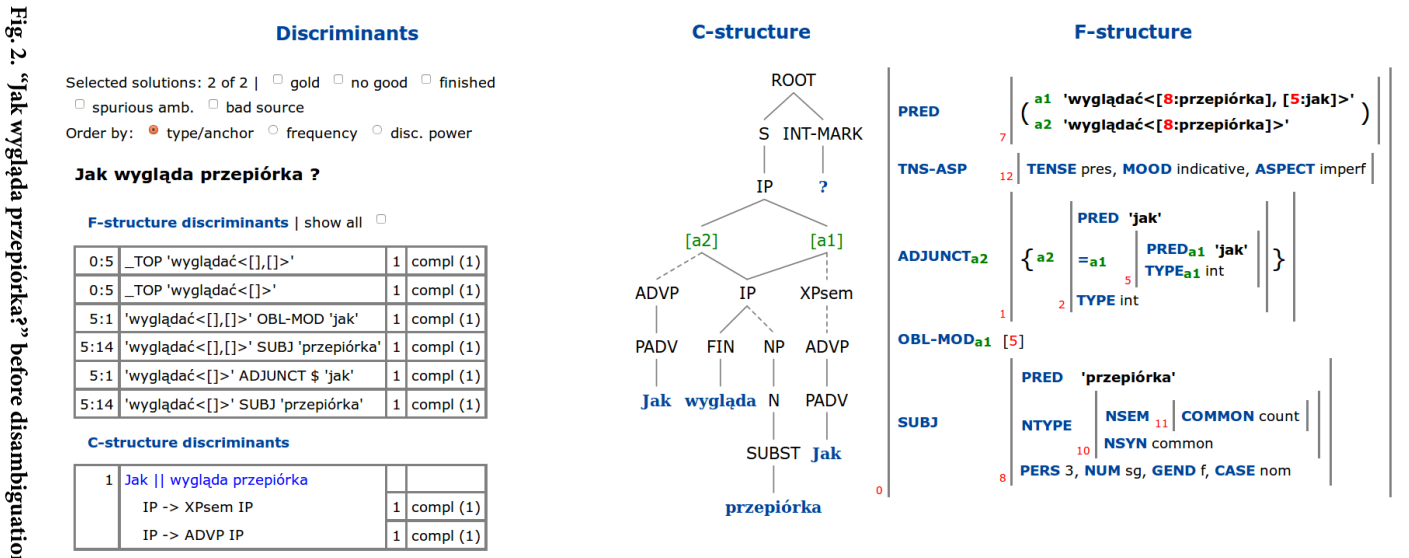


Fig. 2. "Jak wygląda przepiórka?" before disambiguation

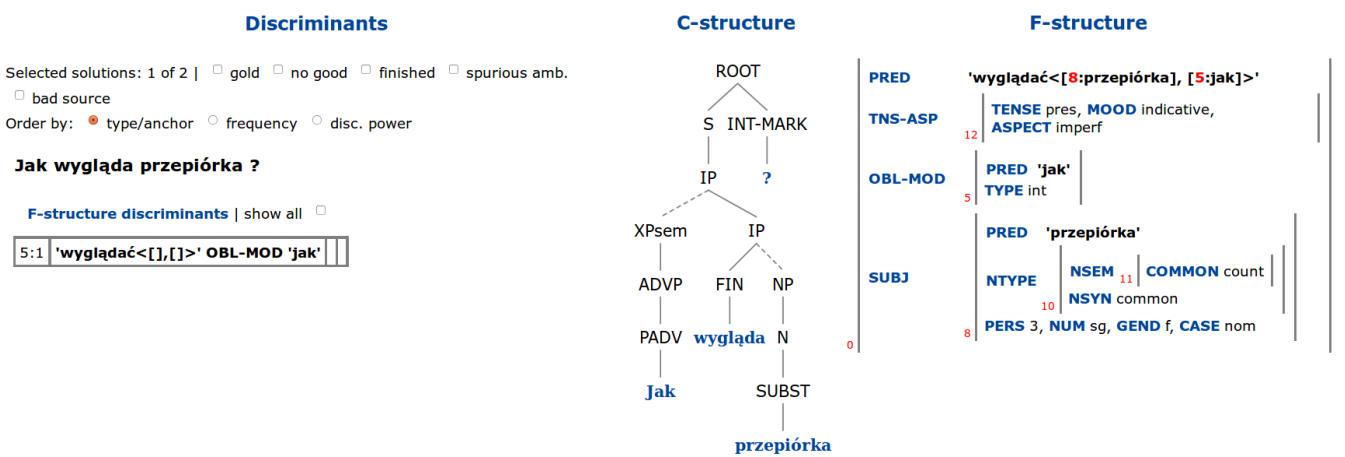


Fig. 3. "Jak wygląda przepiórka?" after disambiguation

5. Improvements

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, as schematically shown in Figure 4:

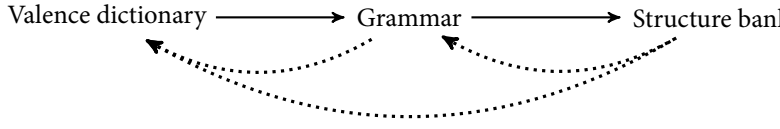


Fig. 4. Flow of information to downstream resources (straight solid arrows) and feedback to upstream resources (curved dotted arrows)

First, *Walenty* is automatically converted to LFG representations to be used in the grammar, and many inconsistencies in the dictionary are 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 problems such 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, those sentences which have syntactic analyses are fed into INESS for disambiguation. 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 it 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 and improves 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 3000 comments in the system.

The whole annotation process is divided into rounds, each involving around 1000 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 earlier work on the LinGO Redwood HPSG treebank (Oepen *et al.* 2002a, b, 2004): choices made for one version of the grammar mostly 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 4 above.

6. Related work

One line of research closely related to ours has already been alluded to above and consists in the parallel development of a grammar and a treebank containing disambiguated trees produced by a parser based on this grammar (hence the term *parsebank*, sometimes used to denote such treebanks; Rosén *et al.* 2009). Well-known examples of this approach include the *LinGO English Resource Grammar* of English and the related *LinGO Redwoods Treebank* (Oepen *et al.* 2002a, b, 2004) developed at Stanford and based on the Head-driven Phrase Structure Grammar formalism of Pollard and Sag 1994, as well as the Norwegian LFG grammar and treebank (Rosén *et al.* 2005, 2007, 2009, 2012, 2014) developed at Bergen. A similar approach has been followed during the development of the *Składnica* treebank and a new version of the DCG grammar of Polish mentioned in section *Formal grammar* above. Obviously, all these grammars contain some valence information, but it is often hard-wired into the grammatical formalism and empirically limited.[3]

Similarly, a tightly coupled development of a treebank and a valence dictionary is certainly not a new idea. The prime example of such an approach is the construction of the *Prague Dependency Treebank* (Böhmová *et al.* 2003) together with the Czech valence dictionary *PDT-Vallex* (Hajič *et al.* 2003, Urešová 2009), where valence frames were added to the dictionary as they were encountered during the development of the treebank. Just as in the enterprise described in this paper, the flow of information was also bidirectional there: the valence dictionary was used to control the quality of subsequent annotations in the treebank. A similarly tightly-coupled approach was followed during the construction of the *TüBa-D/Z Treebank* and the related valence lexicon of German (Hinrichs and Telljohann 2009). Note that such a concurrent development of linguistic resources should be distinguished from the more usual serial (or pipeline) approach, as witnessed for example in the PropBank project (Palmer *et al.* 2005), where an existing treebank (Marcus *et al.* 1993) is extended with semantic roles and a semantic valence lexicon is derived as a by-product.

7. Conclusion

Evaluation of the quality and completeness of any linguistic resources, and especially of valence dictionaries, is difficult (cf. e.g. Przepiórkowski 2009). By the concurrent development of a relatively theory-independent dictionary and a com-

[3] However, let us note that the Polish DCG parser *Świgra* is currently being coupled with *Walenty* (Marcin Woliński, p.c.).

prehensive LFG grammar taking advantage of almost all types of information in this dictionary, the quality of the dictionary is partially verified and further improved. 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 and improved. This setup goes beyond the development of a valence dictionary on the basis of – or in parallel with – a treebank, not only because it also involves the development of a grammar, but also because of the richness of linguistic information in the LFG structure bank. While this approach to the parallel development of linguistic resources is often difficult – due to the scarcity of non-linguistic resources (both budgetary and human) – we maintain that such a holistic approach should always be strived for.

Our future plans extend this approach even further. The addition of formalised semantic information to the valence dictionary is planned. This information will be subsequently verified and further improved via the use of semantic representations – produced by the grammar based on this extended dictionary – in the task of recognising textual entailment (Dagan *et al.* 2013), on the basis of a new textual entailment corpus of Polish (Przepiórkowski 2015).

References

Böhmová, A., Hajič, J., Hajičová, E., and Hladká, B. (2003). The Prague Dependency Treebank: Three-level annotation scenario. In A. Abeillé, editor, *Treebanks: Building and Using Parsed Corpora*, volume 20 of *Text, Speech and Language Technology*, pages 103–127. Kluwer, Dordrecht.

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. (2001). *Lexical-Functional Syntax*. Blackwell Textbooks in Linguistics. Blackwell, Malden, MA.

Crouch, D., Dalrymple, M., Kaplan, R., King, T., Maxwell, J., and Newman, P. (2011). XLE documentation. http://www2.parc.com/isl/groups/nltt/xle/doc/xle_toc.html.

Dagan, I., Roth, D., Sammons, M., and Zanzotto, F. M. (2013). *Recognizing Textual Entailment: Models and Applications*. Morgan & Claypool.

Dalrymple, M. (2001). *Lexical Functional Grammar*. Academic Press, San Diego, CA.

Degórski, Ł. and Przepiórkowski, A. (2012). Ręcznie znakowany milionowy podkorpus NKJP. In Przepiórkowski *et al.* (2012), pages 51–58.

Frank, A., Sadler, L., van Genabith, J., and Way, A. (2003). From treebank resources to LFG f-structures. In A. Abeillé, editor, *Treebanks: Building and Using Parsed Corpora*, volume 20 of *Text, Speech and Language Technology*, pages 367–389. Kluwer, Dordrecht.

Hajič, J., Panevová, J., Urešová, Z., Bémová, A., Kolářová, V., and Pajas, P. (2003). PDT-VALLEX: Creating a large-coverage valency lexicon for treebank annotation.

In J. Nivre and E. Hinrichs, editors, *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT 2003)*, Växjö, Norway.

Hajnicz, E., Nitoń, B., Patejuk, A., Przepiórkowski, A., and Woliński, M. (2015). Internetowy słownik walencyjny języka polskiego oparty na danych korpusowych. *Prace Filologiczne*, LXV. To appear.

Hinrichs, E. W. and Telljohann, H. (2009). Constructing a valence lexicon for a treebank of German. In F. van Eynde, A. Frank, K. De Smedt, and G. van Noord, editors, *Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories (TLT 7)*, pages 41–52, Groningen, The Netherlands.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19, 313–330.

Oepen, S., Flickinger, D., Toutanova, K., and Manning, C. D. (2002a). LinGO Redwoods: A rich and dynamic treebank for HPSG. In E. Hinrichs and K. Simov, editors, *Proceedings of the First Workshop on Treebanks and Linguistic Theories (TLT2002)*, pages 139–149, Sozopol.

Oepen, S., Toutanova, K., Shieber, S., Manning, C., Flickinger, D., and Brants, T. (2002b). The LinGO Redwoods Treebank: Motivation and preliminary applications. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, Taipei.

Oepen, S., Flickinger, D., Toutanova, K., and Manning, C. D. (2004). LinGO Redwoods: A rich and dynamic treebank for HPSG. *Research on Language and Computation*, 4(2), 575–596.

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1), 71–105.

Patejuk, A. and Przepiórkowski, A. (2012). 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. ELRA.

Patejuk, A. and Przepiórkowski, A. (2014). Synergistic development of grammatical resources: A valence dictionary, an LFG grammar, and an LFG structure bank for Polish. In V. Henrich, E. Hinrichs, D. de Kok, P. Osenova, and A. Przepiórkowski, editors, *Proceedings of the Thirteenth International Workshop on Treebanks and Linguistic Theories (TLT 13)*, pages 113–126, Tübingen, Germany. Department of Linguistics (SfS), University of Tübingen.

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

Przepiórkowski, A. (2009). Towards the automatic acquisition of a valence dictionary for Polish. In M. Marciniak and A. Mykowiecka, editors, *Aspects of Natural Language Processing. Essays dedicated to Leonard Bolc on the Occasion of His 75th Birthday*, volume 5070 of *Lecture Notes in Computer Science*, pages 191–210. Springer-Verlag, Berlin.

Przepiórkowski, A. (2015). Towards a linguistically-oriented textual entailment test-suite for Polish based on the semantic syntax approach. *Cognitive Studies / Études Cognitives*, 15. To appear.

Przepiórkowski, A. and Murzynowski, G. (2011). Manual annotation of the National Corpus of Polish with Anotatornia. In S. Goźdź-Roszkowski, editor, *Explorations across Languages and Corpora: PALC 2009*, pages 95–103, Frankfurt am Main. Peter Lang.

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

Przepiórkowski, A., Bańko, M., Górski, R. L., and Lewandowska-Tomaszczyk, B., editors (2012). *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN, Warsaw.

Przepiórkowski, A., Hajnicz, E., Patejuk, A., and Woliński, M. (2014a). 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. Association for Computational Linguistics and Dublin City University.

Przepiórkowski, A., Skwarski, F., Hajnicz, E., Patejuk, A., Świdziński, M., and Woliński, M. (2014b). Modelowanie własności składniowych czasowników w nowym słowniku walencyjnym języka polskiego. *Polonica*, XXXIII, 159–178.

Przepiórkowski, A., Hajnicz, E., Patejuk, A., Woliński, M., Skwarski, F., and Świdziński, M. (2014c). Walenty: Towards a comprehensive valence dictionary of Polish. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, pages 2785–2792, Reykjavík, Iceland. ELRA.

Rosén, V., Meurer, P., and Smedt, K. D. (2005). Constructing a parsed corpus with a large LFG grammar. In M. Butt and T. H. King, editors, *The Proceedings of the LFG'05 Conference*, pages 371–387, University of Bergen, Norway. CSLI Publications.

Rosén, V., Meurer, P., and Smedt, K. D. (2007). Designing and implementing discriminants for LFG grammars. In M. Butt and T. H. King, editors, *The Proceedings of the LFG'07 Conference*, pages 397–417, University of Stanford, California, USA. CSLI Publications.

Rosén, V., Meurer, P., and Smedt, K. D. (2009). LFG Parsebanker: A toolkit for building and searching a treebank as a parsed corpus. In F. van Eynde, A. Frank, K. De Smedt, and G. van Noord, editors, *Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories (TLT 7)*, pages 127–133, Groningen, The Netherlands.

Rosén, V., Meurer, P., Losnegaard, G. S., Lyse, G. I., De Smedt, K., Thunes, M., and Dyvik, H. (2012). An integrated web-based treebank annotation system. In I. Hendrickx, S. Kübler, and K. Simov, editors, *Proceedings of the Eleventh International Workshop on Treebanks and Linguistic Theories (TLT 11)*, pages 157–168, Lisbon, Portugal.

Rosén, V., Haugereid, P., Thunes, M., Losnegaard, G. S., and Dyvik, H. (2014). The interplay between lexical and syntactic resources in incremental parsebanking. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, pages 1617–1624, Reykjavík, Iceland. ELRA.

Sag, I. A., Gazdar, G., Wasow, T., and Weisler, S. (1985). Coordination and how to distinguish categories. *Natural Language and Linguistic Theory*, 3, 117–171.

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 P. Sojka, A. Horák, I. Kopeček, and K. Pala, editors, *Text, Speech and Dialogue: 13th International Conference, TSD 2010, Brno, Czech Republic*, volume 6231 of *Lecture Notes in Artificial Intelligence*, pages 197–204, Heidelberg. Springer-Verlag.

Urešová, Z. (2009). Building the PDT-Vallex valency lexicon. In *On-line Proceedings of the fifth Corpus Linguistics Conference*. University of Liverpool.

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, Institute of Computer Science, Polish Academy of Sciences, Warsaw.

Woliński, M. (2006). Morfeusz — a practical tool for the morphological analysis of Polish. In M. A. Kłopotek, S. T. Wierzchoń, and K. Trojanowski, editors, *Intelligent Information Processing and Web Mining*, Advances in Soft Computing, pages 503–512. Springer-Verlag, Berlin.

Woliński, M. (2014). Morfeusz reloaded. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, pages 1106–1111, Reykjavík, Iceland. ELRA.

Woliński, M., Głowińska, K., and Świdziński, M. (2011). A preliminary version of Składnica—a treebank of Polish. In Z. 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.

Skoordynowany rozwój zasobów językowych: bank struktur języka polskiego

s t r e s z c z e n i e

Celem niniejszego artykułu jest prezentacja nowego zasobu językowego – korpusu struktur składniowych polszczyzny zgodnych z teorią Lexical Functional Grammar – i wpływu procesu opracowywania tego zasobu na jakość i pełność dwóch innych zasobów: słownika walencyjnego *Walenty* i gramatyki formalnej *POLFIE*.