Intelligent Information Systems 2008
ISBN 978-83-60434-44-4, pages 491–499

Linguistic resources and tools at ICS PAS: Towards interoperability

Adam Przepiórkowski

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

Abstract

This is an overview paper, presenting linguistic resources and tools developed within projects carried out by the Linguistic Engineering Group (LEG) at the Institute of Computer Science of the Polish Academy of Sciences. It briefly describes corpora (esp., the *IPI PAN Corpus* of Polish) and corpus tools (*Poliqarp*), grammars and parsers (incl. *Spejd* and *Świgra*), and syntactic and semantic lexica, and it discusses the extent to which they may be claimed to satisfy the requirement of interoperability.

Keywords: language resources, language tools, Polish, Linguistic Engineering Group at ICS PAS

1 Introduction

The aim of this paper is to present language resources and tools developed at the Linguistic Engineering Group (LEG) of the Institute of Computer Science, Polish Academy of Sciences (ICS PAS) in Warsaw.[1] A brief description of the past and current work on corpora and corpus tools (§2), grammars and parsers (§3), and syntactic and semantic lexica (§4), is followed by a discussion of the interoperability of these resources (§5).

The resources and tools presented in the following sections do not exhaust the results of various projects carried out by LEG at ICS PAS: this overview concentrates on those products which are of general interest and which are currently publicly available. Other resources that may be mentioned here are:

- a test-suite of Polish sentences with grammaticality judgements and syntactic parses (Marciniak *et al.*, 2003), created with the European project *Co-operative research in information technology* (CRIT-2) and used, e.g., for the evaluation of the dependency parser of Obrębski (2002);
- various information extraction resources for Polish described in Mykowiecka *et al.* 2007a, developed within a national project (their publication is planned at the following URL: `http://ie.ipipan.waw.pl/`);

[1] LEG was created and headed by Prof. Leonard Bolc since early 1990ies. I have had the honour and pleasure to be the Head the group since January 2005.

- spoken dialogues (Mykowiecka *et al.*, 2007b) collected and annotated within the European project *Spoken Language Understanding in Multilingual Communication Systems* (LUNA);
- various resources, including shallow definition extraction grammars (Przepiórkowski *et al.*, 2007a,b) and classifiers (Degórski *et al.*, 2008; Przepiórkowski *et al.*, 2008b; Kobyliński and Przepiórkowski, 2008) created within the European project *Language Technology for eLearning* (LT4eL).

2 Corpora

One of the first publicly available linguistic resources developed at ICS PAS was the IPI PAN Corpus of Polish (Przepiórkowski *et al.*, 2003; Przepiórkowski, 2004), still the largest corpus of Polish (with 250 million segments) and the only one that is fully morphosyntactically annotated. The tagset used for the annotation, the IPI PAN Tagset, is based on the linguistic ideas of Janusz S. Bień and Zygmunt Saloni (Saloni, 1974, 1977, 1981, 1988; Gruszczyński and Saloni, 1978; Bień and Saloni, 1982; Bień, 1991), and it is fully described in Przepiórkowski 2004 (and in references cited therein). The corpus is available for search at `http://korpus.pl/`, and a 100-million unbalanced subcorpus in its textual (XML) form is made available for research purposes on request.

The IPI PAN Corpus was created within a national project carried out at ICS PAS from 2001 to 2004. Apart from the corpus itself, one of the main results of the project was a corpus indexing and search engine, *Poliqarp* (Przepiórkowski *et al.*, 2004), which makes available a rich query language described in Przepiórkowski 2004. *Poliqarp* has subsequently been extended with some statistical functionalities (Buczyński, 2007b) and with some limited syntactic querying possibilities (Przepiórkowski, 2008b; Janus and Przepiórkowski, 2007). *Poliqarp* is avilable on the GNU GPL licence from `http://korpus.pl/`.

Another important result of that project was probably the first morphosyntactic tagger of Polish (Dębowski, 2003, 2004), used to annotate the first edition of the IPI PAN Corpus. That tagger never reached the stage where public distribution was seriously considered. Another tagger (Piasecki and Godlewski, 2006b,a), developed at another national project carried out at ICS PAS and maintained at the Wrocław University of Technology, is currently being prepared for public release on the GNU GPL licence.

The IPI PAN Corpus is not the only corpus of Polish in existance. Another corpus, which is considered to be carefully balanced, the PWN Corpus of Polish (`http://korpus.pwn.pl/`), contains over 100 million words, of which only a 7.5 million sample is freely available for search. The third corpus, the PELCRA Corpus of Polish (`http://korpus.ia.uni.lodz.pl/`), also contains about 100 million words, all of which are publicly searchable.

In December 2007 a new 3-year national project started (`http://nkjp.pl/`), whose aim is to bring together all previous major corpus initiatives and develop a very large national reference corpus of Polish, as well as various language tools. The project is coordinated by ICS PAS and its main goals are described in Przepiórkowski *et al.* 2008a.

3 Grammars and Parsers

One of the traditional areas of research within the Linguistic Engineering Group of ICS PAS is syntactic processing (Bolc and Mykowiecka, 1992), especially, with the use of unification grammars (Marciniak *et al.*, 1993; Kupść *et al.*, 1995). These activities resulted in numerous publications, including four Ph.D. dissertations (Mykowiecka, 1999; Przepiórkowski, 1999; Kupść, 2000; Marciniak, 2001), and in a few prototype toy implementations. Przepiórkowski *et al.* 2002 combines various partial results into a coherent grammar of fragments of Polish, although the resulting parser, due to the lack of effectiveness of the underlying formalism and to the empirical deficiencies, never reached the stage of a practical language processing tool.

A large scale grammar of Polish by Świdziński (1992), extending the grammar of Szpakowicz (1978, 1986), was efficiently implemented at ICS PAS by Woliński (2004, 2005). Some empirical evaluation of the grammar has been carried out (Świdziński, 1996; Ogrodniczuk, 2006), although a more comprehensive evaluation of the grammar and its implementation is still called for. The parser, named *Świgra*, is publicly available at `http://nlp.ipipan.waw.pl/~wolinski/swigra/`.

More recently, a new partial parsing formalism, *Spejd*, has been developed (Przepiórkowski, 2007a; Przepiórkowski and Buczyński, 2007), implemented (Buczyński, 2007a; Buczyński and Przepiórkowski, 2008) and made publicly available on the GNU GPL licence (`http://nlp.ipipan.waw.pl/Spejd/`). One unique feature of *Spejd* is its seamless combination of functionalities normally associated with two separate tools: a rule-based morphosyntactic dismabiguator and a cascaded shallow parser. A partial grammar of Polish is currently being developed (Przepiórkowski, 2007b, 2008a), and there are plans to make it publicly available.

4 Syntactic and Semantic Lexica

Two national projects, recently and currently carried out at ICS PAS, are concerned with the automatic acquisition of valence dictionaries, i.e., lexica containing information about the possible types of arguments of verbs and possibly other predicates, from large morphosyntactically annotated corpora.

The first of these projects, which ended in March 2008, focussed on the syntactic information. Early results of that project are reported in Fast and Przepiórkowski 2005, as well as Przepiórkowski and Fast 2005, while more recent results — in Dębowski and Woliński 2007, and Dębowski 2007. Although the quality of the extracted valence information is still far from satisfactory, the resulting valence dictionary will be made publicly available, in the hope of being useful at least for some applications, and perhaps as a starting point for further work. Such “further work” is carried out within the second project, running from November 2007 to November 2009, which aims to build on the results of the syntactic valence acquisition by adducing semantic valence information (Hajnicz, 2007).

5 Interoperability

Since January 2008, ICS PAS is a member of the CLARIN (Common Language Resources Infrastructure; `http://www.clarin.eu/`) project. According to the CLARIN grant agreement, the main technical objective of the project with respect to language resources and tools is *to provide a detailed specification of the infrastructure, agreement on data and interoperability standards to be adopted, and a running, validated prototype based on these specifications* (CLARIN, page 6). The key concept of the project is *interoperability.*

Wikipedia defines *interoperability* as *a property referring to the ability of diverse systems and organizations to work together (inter-operate).*[2] Specifically, in case of software, *the term interoperability is used to describe the capability of different programs to exchange data via a common set of exchange formats, to read and write the same file formats, and to use the same protocols... The lack of interoperability can be a consequence of a lack of attention to standardization during the design of a program.*

The resources and tools presented in the previous sections have been created with regard to various standards enhancing their interoperability. The most trivial example of that is the early adoption of the Unicode standard and the UTF-8 character encoding back in 2001, at the inception of the IPI PAN Corpus project, when the two most popular encoding schemes were ISO-8859-2 and the Windows code page 1250. While the former is an ISO standard, it does not allow for the representation of various characters do not occur in the Polish alphabet but which do occur in Polish texts, e.g., Greek letters. Moreover, representing IPI PAN Corpus texts with the use of one of these two “regional” pages would create problems for researchers working in other “code page regions”, exactly the kinds of problems that gave rise to the importance of interoperability.

Slightly less trivially, it was decided early on that the basic format for the IPI PAN Corpus should be the XML Corpus Encoding Standard (XCES; Ide *et al.* 2000), a specialisation of the guidelines of the Text Encoding Initiative. This was a lucky choice, as currently XCES is still probably the best corpus annotation standard around (Suderman and Ide, 2006; Stührenberg, 2007), although it may be replaced by the *Linguistic Annotation Framework* (LAF) in the near future (Ide and Romary, 2004a,b; Ide, 2007).

Both standards, UTF-8 and XCES, are also understood by two of the tools introduced above: the shallow parsing system *Spejd* and the corpus indexer and search tool *Poliqarp.* The latter tool is used in a Portuguese corpus project (Barreto *et al.*, 2006), and has been successfully applied (at the University of Regensburg) to the indexing and searching of an Old Russian corpus. The publicly available XCES sample of the IPI PAN Corpus has been utilised in numerous projects in Poland, including the valence projects presented above, as well as in Croatia, Germany, Great Britain and Spain. Also the tagger mentioned in §2, to be publicly released soon, is XCES-aware and, for this reason, could be easily adopted to the tasks carried out within the LT4eL project (cf. §1), where the XCES standard is also used. Clearly, the idea of interoperability is at play here,

[2] `http://en.wikipedia.org/wiki/Interoperability`

even if only at a small scale.

Of course, the strive towards interoperability should not consist in the blind assumption of any proposed standards. For example, in Przepiórkowski and Woliński 2003 we argue — on the linguistic grounds — against the adoption of the popular MULTEXT-East tagset scheme (Erjavec, 2001) to Polish, instead opting for a more principled solution in agreement with the Polish structuralist linguistic tradition.

Given the interplay of linguistic, technical, historical and political arguments for or against any standard, the path towards the full interoperability of language resources and tools will inevitably turn out to be thorny, but overcoming the dangers lurking there will be a part of the exciting CLARIN adventure.

References

Florbela Barreto, António Branco, Eduardo Ferreira, Amália Mendes, Maria Fernanda Nascimento, Filipe Nunes, and João Silva (2006), Open Resources and Tools for the Shallow Processing of Portuguese: The TagShare Project, in *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006*, ELRA, Genoa.

Janusz S. Bień (1991), *Koncepcja słownikowej informacji morfologicznej i jej komputerowej weryfikacji*, volume 383 of *Rozprawy Uniwersytetu Warszawskiego*, Wydawnictwa Uniwersytetu Warszawskiego, Warsaw.

Janusz S. Bień and Zygmunt Saloni (1982), Pojęcie wyrazu morfologicznego i jego zastosowanie do opisu fleksji polskiej (wersja wstępna), *Prace Filologiczne*, XXXI:31–45.

Leonard Bolc and Agnieszka Mykowiecka (1992), *Podstawy przetwarzania języka naturalnego. Wybrane metody formalnego zapisu składni*, Akademicka Oficyna Wydawnicza RM, Warsaw.

Aleksander Buczyński (2007a), An Implementation of Combined Partial Parser and Morphosyntactic Disambiguator, in *Proceedings of the ACL 2007 Student Research Workshop*, pp. 13–18, Prague.

Aleksander Buczyński (2007b), Statistical extension to the Poliqarp search engine, in Ville V. Nurmi and Dmitry Sustretov, editors, *Proceedings of the Twelfth ESSLLI Student Session*, pp. 47–54, Dublin.

Aleksander Buczyński and Adam Przepiórkowski (2008), ♠ Demo: An Open Source Tool for Shallow Parsing and Morphosyntactic Disambiguation, in LREC (b), forthcoming.

CLARIN (2007), CLARIN: Grant agreement for Combination of Collaborative Project and Coordination and Support Actions. Annex I – "Description of Work", grant agreement no.: 212230. Date of approval of Annex I: 31 October 2007.

Łukasz Degórski, Michał Marcińczuk, and Adam Przepiórkowski (2008), Definition extraction using a sequential combination of baseline grammars and machine learning classifiers, in LREC (b), forthcoming.

Łukasz Dębowski (2003), A reconfigurable stochastic tagger for languages with complex tag structure, in *Proceedings of* Morphological Processing of Slavic Languages*, EACL 2003.*

Łukasz Dębowski (2004), Trigram morphosyntactic tagger for Polish, in Mieczysław A.

Kłopotek, Sławomir T. Wierzchoń, and Krzysztof Trojanowski, editors, *Intelligent Information Processing and Web Mining*, Advances in Soft Computing, pp. 409–413, Springer-Verlag, Berlin.

Łukasz Dębowski (2007), Valence extraction using the EM selection and co-occurrence matrices, arXiv:0711.4475v2 [cs.CL] 5 Dec 2007.

Łukasz Dębowski and Marcin Woliński (2007), Argument co-occurence matrix as a description of verb valence, in Vetulani (2007), pp. 260–264.

Tomaž Erjavec, editor (2001), *Specifications and Notation for MULTEXT-East Lexicon Encoding*, Ljubljana.

Jakub Fast and Adam Przepiórkowski (2005), Automatic Extraction of Polish Verb Subcategorization: An Evaluation of Common Statistics, in Zygmunt Vetulani, editor, *Proceedings of the 2nd Language & Technology Conference*, pp. 191–195, Poznań, Poland.

Włodzimierz Gruszczyński and Zygmunt Saloni (1978), Składnia grup liczebnikowych we współczesnym języku polskim, *Studia Gramatyczne*, II:17–42.

Elżbieta Hajnicz (2007), Extending syntactic valence dictionary for Polish with semantic categories, Referat wygłoszony na konferencji *7th European Conference on Formal Description of Slavic Languages*, Lipsk, Niemcy, 30 listopada 2007.

Nancy Ide (2007), Annotation Science: From Theory to Practice and Use, in Rehm *et al.* (2007), pp. 3–7.

Nancy Ide, Patrice Bonhomme, and Laurent Romary (2000), XCES: An XML-based Standard for Linguistic Corpora, in *Proceedings of the Third International Conference on Language Resources and Evaluation, LREC 2000*, pp. 825–830, ELRA, Athens.

Nancy Ide and Laurent Romary (2004a), International Standard for a Linguistic Annotation Framework, *Natural Language Engineering*, 10:211–225.

Nancy Ide and Laurent Romary (2004b), A Registry of Standard Data Categories for Linguistic Annotation, in LREC (a), pp. 135–139.

Daniel Janus and Adam Przepiórkowski (2007), Poliqarp: An open source corpus indexer and search engine with syntactic extensions, in *Proceedings of the ACL 2007 Demo and Poster Sessions*, pp. 85–88, Prague.

Łukasz Kobyliński and Adam Przepiórkowski (2008), Definition Extraction with Balanced Random Forests, in progress.

Anna Kupść (2000), *An HPSG Grammar of Polish Clitics*, Ph. D. dissertation, Polish Academy of Sciences and Université Paris 7.

Anna Kupść, Małgorzata Marciniak, Agnieszka Mykowiecka, and Adam Przepiórkowski (1995), Formal Analysis of Polish in HPSG, in Mirosław Dąbrowski, Maciej Michalewicz, and Zbigniew Raś, editors, *Intelligent Information Systems. Proceedings of the Fourth Workshop on Intelligent Information Systems*, pp. 295–305, Wydawnictwa IPI PAN, Augustów, Poland.

LREC (2004a), *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004*, ELRA, Lisbon.

LREC (2008b), *Proceedings of the Sixth International Conference on Language Resources and Evaluation, LREC 2008*, ELRA, Marrakech.

Małgorzata Marciniak (2001), *Algorytmy implementacyjne syntaktycznych reguł koreferencji zaimków dla języka polskiego w terminach HPSG*, Ph. D. dissertation, Polish Academy of Sciences.

Małgorzata Marciniak, Agnieszka Mykowiecka, and Elżbieta Dobryjanowicz

(1993), Wprowadzenie do gramatyk unifikacyjnych, IPI PAN Research Report 736, Institute of Computer Science, Polish Academy of Sciences, Warsaw.

Małgorzata Marciniak, Agnieszka Mykowiecka, Adam Przepiórkowski, and Anna Kupść (2003), An HPSG-Annotated Test Suite for Polish, in Anne Abeillé, editor, *Treebanks: Building and Using Parsed Corpora*, volume 20 of *Text, Speech and Language Technology*, pp. 129–146, Kluwer, Dordrecht.

Agnieszka Mykowiecka (1999), *Opis składniowy polskich konstrukcji względnych w formalizmie HPSG*, Ph. D. dissertation, Polish Academy of Sciences.

Agnieszka Mykowiecka, Anna Kupść, Małgorzata Marciniak, and Jakub Piskorski (2007a), Resources for Information Extraction from Polish texts, in Vetulani (2007).

Agnieszka Mykowiecka, Krzysztof Marasek, Małgorzata Marciniak, Joanna Rabiega-Wiśniewska, and Ryszard Gubrynowicz (2007b), Annotation of Polish spoken dialogs in LUNA project, in Vetulani (2007).

Tomasz Obrębski (2002), *Automatyczna analiza składniowa języka polskiego z wykorzystaniem gramatyki zależnościowej*, Ph. D. dissertation, Institute of Computer Science, Polish Academy of Sciences, Warsaw.

Maciej Ogrodniczuk (2006), *Weryfikacja korpusu wypowiedników polskich (z wykorzystaniem gramatyki formalnej Świdzińskiego)*, Ph. D. dissertation, Warsaw University, Warsaw.

Maciej Piasecki and Grzegorz Godlewski (2006a), Effective Architecture of the Polish Tagger, in Petr Sojka, Ivan Kopeček, and Karel Pala, editors, *Text, Speech and Dialogue: 9th International Conference, TSD 2006, Brno, Czech Republic, September 2006*, volume 4188 of *Lecture Notes in Artificial Intelligence*, pp. 213–220, Springer-Verlag, Berlin.

Maciej Piasecki and Grzegorz Godlewski (2006b), Reductionistic, Tree and Rule Based Tagger for Polish, in Mieczysław A. Kłopotek, Sławomir T. Wierzchoń, and Krzysztof Trojanowski, editors, *Intelligent Information Processing and Web Mining*, Advances in Soft Computing, pp. 531–540, Springer-Verlag, Berlin.

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

Adam Przepiórkowski (2004), *The IPI PAN Corpus: Preliminary version*, Institute of Computer Science, Polish Academy of Sciences, Warsaw.

Adam Przepiórkowski (2007a), A Preliminary Formalism for Simultaneous Rule-Based Tagging and Partial Parsing, in Rehm *et al.* (2007), pp. 81–90.

Adam Przepiórkowski (2007b), Towards a Partial Grammar of Polish for Valence Extraction, in *Proceedings of Grammar and Corpora 2007, Liblice, Czech Republic*, forthcoming.

Adam Przepiórkowski (2008a), *Powierzchniowe przetwarzanie języka polskiego*, Akademicka Oficyna Wydawnicza EXIT, Warsaw, forthcoming.

Adam Przepiórkowski (2008b), Zapytania składniowe w wyszukiwarce korpusowej Poliqarp, in Grażyna Habrajska, editor, *Rozmowy o komunikacji*, Leksem, Łódź, forthcoming.

Adam Przepiórkowski, Piotr Bański, Łukasz Dębowski, Elżbieta Hajnicz, and Marcin Woliński (2003), Konstrukcja korpusu IPI PAN, *Polonica*, XXII–XXIII:33–38.

Adam Przepiórkowski and Aleksander Buczyński (2007), ♠: Shallow Parsing and Disambiguation Engine, in Vetulani (2007), pp. 340–344.

Adam Przepiórkowski, Łukasz Degórski, Miroslav Spousta, Kiril Simov, Petya Osenova, Lothar Lemnitzer, Vladislav Kuboň, and Beata Wójtowicz (2007a), Towards the automatic extraction of definitions in Slavic, in Jakub Piskorski, Bruno Pouliquen, Ralf Steinberger, and Hristo Tanev, editors, *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing at ACL 2007*, pp. 43–50, Prague.

Adam Przepiórkowski, Łukasz Degórski, and Beata Wójtowicz (2007b), On the evaluation of Polish definition extraction grammars, in Vetulani (2007), pp. 473–477.

Adam Przepiórkowski and Jakub Fast (2005), Baseline Experiments in the Extraction of Polish Valence Frames, in Mieczysław A. Kłopotek, Sławomir T. Wierzchoń, and Krzysztof Trojanowski, editors, *Intelligent Information Processing and Web Mining*, Advances in Soft Computing, pp. 511–520, Springer-Verlag, Berlin.

Adam Przepiórkowski, Rafal L. Górski, Barbara Lewandowska-Tomaszczyk, and Marek Łaziński (2008a), Towards the National Corpus of Polish, in LREC (b), forthcoming.

Adam Przepiórkowski, Zygmunt Krynicki, Łukasz Dębowski, Marcin Woliński, Daniel Janus, and Piotr Bański (2004), A Search Tool for Corpora with Positional Tagsets and Ambiguities, in LREC (a), pp. 1235–1238.

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

Adam Przepiórkowski, Michał Marcińczuk, and Łukasz Degórski (2008b), Dealing with Small, Noisy and Imbalanced Data: Machine Learning or Manual Grammars?, submitted to a conference.

Adam Przepiórkowski and Marcin Woliński (2003), A Flexemic Tagset for Polish, in *Proceedings of Morphological Processing of Slavic Languages, EACL 2003*, pp. 33–40, Budapest.

Georg Rehm, Andreas Witt, and Lothar Lemnitzer, editors (2007), *Data Structures for Linguistic Resources and Applications: Proceedings of the Biennial GLDV Conference 2007*, Gunter Narr Verlag, Tübingen.

Zygmunt Saloni (1974), Klasyfikacja gramatyczna leksemów polskich, *Język Polski*, LIV:3–13, 93–101.

Zygmunt Saloni (1977), Kategorie gramatyczne liczebników we współczesnym języku polskim, *Studia Gramatyczne*, I:145–173.

Zygmunt Saloni (1981), Uwagi o opisie fleksyjnym tzw. zaimków rzeczownych, *Folia Linguistica*, 2:265–271.

Zygmunt Saloni (1988), O tzw. formach nieosobowych [rzeczowników] męskoosobowych we współczesnej polszczyźnie, *Biuletyn Polskiego Towarzystwa Językoznawczego*, XLI:155–166.

Maik Stührenberg (2007), Texttechnological Standards — An Overview, in Rehm *et al.* (2007), pp. 157–166.

Keith Suderman and Nancy Ide (2006), Layering and Merging Linguistic Annotations, in *Proceedings of the 5th Workshop on NLP and XML (NLPXML-2006): Multi-Dimensional Markup in Natural Language Processing*, pp. 89–92, ACL, Trento.

Stanisław Szpakowicz (1978), *Automatyczna analiza składniowa zdań pisanych*, Ph. D. dissertation, Warsaw University, Warsaw.

Stanisław Szpakowicz (1986), *Formalny opis składniowy zdań polskich*, Wydawnictwa Uniwersytetu Warszawskiego, Warsaw.

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

Marek Świdziński (1996), *Własności składniowe wypowiedników polskich*, Dom Wydawniczy Elipsa, Warsaw.

Zygmunt Vetulani, editor (2007), *Proceedings of the 3rd Language & Technology Conference*, Poznań, Poland.

Marcin Woliński (2004), *Komputerowa weryfikacja gramatyki Świdzińskiego*, Ph. D. dissertation, Institute of Computer Science, Polish Academy of Sciences, Warsaw.

Marcin Woliński (2005), An efficient implementation of a large grammar of Polish, *Archives of Control Sciences*, 15(3):251–258.