Towards the Design of a Syntactico-Semantic Lexicon for Polish

Adam Przepiórkowski

Institute of Computer Science, Polish Academy of Sciences, ul. Ordona 21, 01-237 Warsaw, Poland

Abstract. This article presents the design of a syntactico-semantic dictionary for Polish, i.e., a valence dictionary enriched with certain semantic informations. Valence dictionaries, specifying the number and morphosyntactic form of arguments of verbs, are useful in many Natural Language Processing applications, including deep parsing, e.g., for the purpose of machine translation, shallow parsing, e.g., for the purpose of information extraction, and rule-based morphosyntactic disambiguation, e.g., for the purpose of corpus annotation. An approach based on recent results in formal and computational linguistics is proposed, which takes into consideration the morphosyntactic and syntactic structure of Polish and which avoids various known problems of previous valence dictionaries, some of them stemming from their impoverished theoretical framework, unable to take proper care of the syntax-semantics interface, case variations and raising predicates. An implementation of a grammar of Polish deploying the ideas presented here is currently under development.

1 Introduction

The aim of this article is to present important aspects of the design of a syntactico-semantic dictionary for Polish.[1] This dictionary is being developed with the intention of forming the lexical basis of a parser[2] of Polish developed at the Institute of Computer Science, Polish Academy of Sciences, [15], but it is meant to be reusable in other systems and applications.

The following section, §2, briefly discusses the nature of valence dictionaries and their usefulness in Natural Language Processing (NLP). Section §3 summarises various problems with the design of the existing valence dictionaries for Polish; such problems are more extensively discussed in [14]. The main section of the article, §4, presents the core ideas of the valence dictionary proposed here; because of its higher emphasis on lexical semantics than is usual in valence dictionaries, we will call it a syntactico-semantic dictionary. Finally, §5 briefly describes an implementation of an interface between such a valence dictionary of lexemes and an external morphological analyser, and §6 contains some concluding remarks.

[1] For space reasons, we cannot compare here the design presented in this article with the design of existing machine-readable lexica for other languages.

[2] The term *parser* is used here with the meaning of 'implementation of a grammar of a particular language', rather than 'platform for implementing such grammars'.

2 Valence Dictionaries in NLP

Valence dictionaries contain information about the valences, or argument lists, of verbs and perhaps other argument-taking predicates. For example, for the English verb *know*, a valence dictionary will provide the information that this verb takes a nominative subject and an object which can be realised either as an accusative nominal phrase (NP), e.g., *I know* ***him***, or a clause introduced by the complementiser *that*, e.g., *I know* ***that he came***, or a clause introduced by a question word, e.g., *I know* ***why he came***, among other options.

Within NLP, valence dictionaries are most obviously useful in the task of constructing so-called deep parsers, i.e., parsers which find the full syntactic and possibly some semantic structure for natural language sentences. Such parsers are used, e.g., in some Machine Translation or Question-Answering Systems. Valence information is also useful in shallow parsing, where only certain aspects of the structure of a sentence are taken into account, e.g., only noun phrases or only predicates and their arguments. Information Extraction is one of the typical application areas of such shallow parsers. Finally, and perhaps surprisingly, such valence information is useful for the task of part of speech (POS) disambiguation, e.g., for the purpose of corpus annotation or speech recognition. A high profile example of a rule-based tagger making use of valence information is ENGCG [23].

The immediate purpose of the work reported here is to further develop an existing constraint-based prototype parser for Polish described in [15] and to enrich it with a well-designed realistic syntactico-semantic dictionary. Nevertheless, it is our hope that the usefulness of the valence dictionary designed along the lines reported here will extend far beyond this immediate application.

3 Valence Dictionaries for Polish

To the best of our knowledge, there are three publicly available dictionaries containing valence information:

- *Słownik syntaktyczno-generatywny czasowników polskich*, [10], published in 5 volumes which appeared between 1980 and 1992; this is probably the most extensive existing source of valence information of Polish verbs;
- *Inny słownik języka polskiego*, [1], a 2-volume general dictionary of Polish which contains various grammatical characteristics of lexemes and their meanings, including valence information;
- *Słownik walencyjny czasowników niemieckich i polskich*, [7], a valence dictionary of German verbs and their Polish counterparts, rather modest both with respect to the number of lexemes and the exhaustiveness of valence information.

The three dictionaries listed above are not available in a machine readable form. A dictionary which does have a machine readable form, but is not publicly available at the time of writing this article, is:

- a *Syntactic Dictionary of Polish Verbs*, [21], an unpublished list of valences of some Polish verbs.

A discussion of the kinds of information made available in these dictionaries can be found in [14]. The following paragraphs summarise some of the deficiencies of these dictionaries from the point of view of potential NLP applications.

Availability as MRDs From the NLP point of view, the most obvious deficiency of the existing valence dictionaries for Polish is the fact that they are not available in machine readable form. Since such dictionaries usually use non-trivial typesetting conventions, they cannot be easily converted to the electronic form using existing OCR software. At the time of writing this article, there are at least two projects (one academic and one commercial) aiming at converting [10] to an electronic form, but the results of these efforts are not yet available, and it is not clear whether they will be publicly available at all.

Syntax-Semantics Interface The dictionaries listed above do not provide any information about the correspondence between the verb's (or, in general, the predicate's) semantic arguments, and its syntactic arguments. Making explicit which syntactic arguments correspond to which semantic roles is important, e.g., in such tasks as Information Extraction and Machine Translation. Two examples illustrating that such correspondence is not trivial and must be stated in the dictionaries are so-called *psych verbs* and the *raising/control* distinction.

Take two psychological predicates, *przestraszyć* 'to frighten' and *przestraszyć się* 'to get scared of'. At a certain level of granularity, they express the same semantic relation, namely about x frightening y. However, the two semantic arguments of this relation, x and y, are realised in two syntactically different ways: in case of *przestraszyć*, x is realised as the (usually nominative) subject and y — as a (usually accusative) object, while in case of *przestraszyć się*, x is realised a genitive object, while y — as the subject.

Similarly, verbs such as *zacząć* 'start' and *próbować* 'try', with similar syntactic arguments (the subject and an infinitival complement), have different semantic arguments. In fact, verbs such as *zacząć*, called raising verbs, are usually treated as semantically mono-valent, with the semantic argument corresponding to the proposition expressed by the infinitival complement, while verbs such as *próbować*, called control verbs or equi verbs, have two semantic arguments, directly corresponding to the syntactic arguments.

Insufficient Formal Apparatus Raising verbs such as *zacząć* mentioned above have the peculiar property of taking just the kind of subject that is expected by the infinitival verbs they combine with. For example, in *Zaczęło padać*, lit. 'Started to rain', the form *zaczęło* does not combine with a subject precisely because the verb *padać* 'to rain' does not expect a subject. Similarly, in *Janka zaczęło dziwić, że pada*, 'It started to surprise John that it is raining', lit. 'John started to surprise that rains', the subordinate clause *że pada* 'that it is raining' can be shown to be the subject of *zaczęło* (cf. [20]). But the only reason that *zaczęło* takes a sentential subject here is that the infinitival verb, *dziwić* 'to surprise', expects such a sentential subject. Valence dictionaries listed above do not have at their disposal a formal apparatus capable of describing such dependencies.

Overly Specific Information Finally, although valence dictionaries are dictionaries of lexemes and the information they provide should be true for all forms of a given lexeme, they usually contain case information which is true only for some forms of the lexeme. For example, the direct object of verbs is usually specified as accusative, even though it is realised as genitive in case of gerundial forms in *-nie/-cie*, often assumed to belong to the verbal lexeme, as nominative in case of passive participles, and as genitive in the scope of verbal negation (roughly speaking). Thus, such valence dictionaries implicitly rely on the users' knowledge about morphosyntactically induced case variations of the lexeme's arguments, instead of specifying such information in an explicit manner.

4 An HPSG-Based Syntactico-Semantic Lexicon

The view of a lexicon presented here is based on standard representations and mechanisms used in Head-driven Phrase Structure Grammar (HPSG), slightly modified to reflect the morphosyntactic and syntactic structure of Polish. This is an obvious choice given the immediate aim of the present endeavour, namely, to provide a syntactico-semantic lexicon for an HPSG-based parser, but also because of the rare combination of positive traits of HPSG: it is 1) a full-fledged linguistic theory, [11,15], 2) with a sound and expressive underlying logical formalism, [18], 3) and a number of computational implementations and general platforms for implementing HPSG-like grammars, [2,4,8].

We assume that syntactic words, i.e., words as they occur in syntactic structures, have four levels of representation corresponding to the argument structure, represented as the values of four attributes:[3]

[3] Because of space constraints, the structures presented in this section are necessarily simplified and they ignore a number of relevant issues, such as the representation of optionality and obligatoriness of arguments.

- CONTENT — the value of this attribute represents the semantic predicate and its semantic argument structure; constant for all forms of a lexeme;
- ARG-ST — basic syntactic argument structure; constant for all forms;
- DEPS — form-specific *dependent* structure, perhaps also including adjuncts [3]; locus of case assignment, binding, etc.;
- VALENCE — represents only overtly realised dependents; this level of representation is redundant, as syntactic tree may be mapped directly from DEPS.

On the view of lexicon assumed here, lexical entries are underspecified descriptions of syntactic words. In particular, lexical entries specify the values of CONTENT and ARG-ST, but not of DEPS or VALENCE. For example, the (first approximations of the) lexical entries for verbs *przestraszyć* and *przestraszyć się* discussed in §3 above are given in (1)–(2):

(1)
$$\begin{bmatrix} word \\ \text{MORPH} & \begin{bmatrix} morph \\ \text{LEMMA} & \text{przestraszyć} \end{bmatrix} \\ \text{SYNSEM} & \begin{bmatrix} synsem \\ \text{ARG-ST} & \langle \text{NP}[\text{CASE } str]_{\boxed{2}}, \text{NP}[\text{CASE } str]_{\boxed{1}} \rangle \\ \text{CONTENT} & \begin{bmatrix} frighten \\ \text{EXPERIENCER } \boxed{1} \\ \text{CAUSE } \boxed{2} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

(2)
$$\begin{bmatrix} word \\ \text{MORPH} & \begin{bmatrix} morph \\ \text{LEMMA} & \text{przestraszyć się} \end{bmatrix} \\ \text{SYNSEM} & \begin{bmatrix} synsem \\ \text{ARG-ST} & \langle \text{NP}[\text{CASE } str]_{\boxed{1}}, \text{NP}[\text{CASE } gen]_{\boxed{2}} \rangle \\ \text{CONTENT} & \begin{bmatrix} frighten \\ \text{EXPERIENCER } \boxed{1} \\ \text{CAUSE } \boxed{2} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Note that the CONTENT values of these verbs are the same but the mapping of semantic arguments, 1 and 2, into the syntactic arguments in the ARG-ST list differs in a way discussed in §3; e.g., the EXPERIENCER argument of *przestraszyć* in (1) corresponds to the second element on its ARG-ST list, i.e., to its complement, while in case of *przestraszyć się* in (2), it corresponds to the first ARG-ST element, i.e., to its subject.

Note also that the syntactic arguments of *przestraszyć* are specified as nominal phrases (which is an approximation, given that they can be realised also as, e.g., numeral phrases) and that they are underspecified with respect to the value of CASE: *str*(uctural) means that the exact morphological case is assigned via general principles, on the basis of the specific form of the lexeme and its syntactic context, cf. [12]. For example, in case of personal forms, as in *Burza przestraszyła Ewę*, lit. 'Storm.NOM frightened Ewa.ACC', the structural case of the subject will be resolved to the nominative, while the structural case of the complement will be resolved to the accusative. On the other hand, in case of the genitive of negation sentence *Burza nie przestraszyła Ewy*, lit. 'Storm.NOM not frightened Ewa.GEN', the structural

case of the complement will be resolved to the genitive, cf. [13]. Note, however, that the complement of the verb *przestraszyć się* in (2) is lexically specified as genitive — its case does not depend on the form of the verb or on the verb's syntactic environment.

The 'deep' syntactic arguments in ARG-ST may be mapped to different 'surface' argument structures in DEPS for different forms of the lexeme. For example, (3)–(4) illustrate possible mappings of the deep syntactic arguments of *przestraszyć* for the active personal form *przestraszyła* and for the passive participle *przestraszony*. This ARG-ST to DEPS mapping is not, however, encoded within lexical entries — such mapping may be specified by general morphosyntactic rules.

(3)
$$\begin{bmatrix} \textit{word} \\ \text{PHON przestraszyła} \\ \text{MORPH} \begin{bmatrix} \textit{morph} \\ \text{LEMMA przestraszyć} \end{bmatrix} \\ \text{SYNSEM} \begin{bmatrix} \textit{synsem} \\ \text{DEPS } \langle \boxed{3}, \boxed{4} \rangle \\ \text{ARG-ST } \langle \boxed{3}\text{NP}[\text{CASE } \textit{nom}]_{\boxed{2}}, \boxed{4}\text{NP}[\text{CASE } \textit{acc}]_{\boxed{1}} \rangle \\ \text{CONTENT} \begin{bmatrix} \textit{frighten} \\ \text{EXPERIENCER } \boxed{1} \\ \text{CAUSE } \boxed{2} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

(4)
$$\begin{bmatrix} \textit{word} \\ \text{PHON przestraszony} \\ \text{MORPH} \begin{bmatrix} \textit{morph} \\ \text{LEMMA przestraszyć} \end{bmatrix} \\ \text{SYNSEM} \begin{bmatrix} \textit{synsem} \\ \text{DEPS } \langle \boxed{4}, \text{PP}[\text{przez}+\boxed{3}] \rangle \\ \text{ARG-ST } \langle \boxed{3}\text{NP}[\text{CASE } \textit{acc}]_{\boxed{2}}, \boxed{4}\text{NP}[\text{CASE } \textit{nom}]_{\boxed{1}} \rangle \\ \text{CONTENT} \begin{bmatrix} \textit{frighten} \\ \text{EXPERIENCER } \boxed{1} \\ \text{CAUSE } \boxed{2} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

The structures (3)–(4) above also illustrate the results of the application of case assignment rules: such rules are assumed to operate at the level of DEPS and, simplifying, they assign the nominative case to the first structural element of DEPS of a verb and the accusative to all other structural dependents of a non-negated verb, as well as to the structural dependents of prepositions. Hence, the 'deep subject' [3] is assigned the nominative case in the active form and the accusative case in the passive form, where it is mapped into the argument of the preposition *przez* 'by'. It is important to remember, however, that such case assignment rules operate outside the realm of the lexicon.

Finally, the elements of DEPS are mapped, again outside the lexicon, into the VALENCE list, which contains only those surface arguments which are realised in the local syntactic tree. For example, in case of the pro-drop sentence *Przestraszyła Ewę*, lit. 'Frightened.2.SG.FEM Ewa.ACC', i.e., 'She frightened Ewa', the VALENCE value corresponding to the DEPS list in (3) will only contain the complement [4], while in case of sentences such as *Tomka, to wydaje mi się, że przestraszyła(, ale mnie nie)*, lit. 'Tom.ACC Particle seems

me.DAT ReflexiveMarker that frightened.2.SG.FEM (but me.ACC not)', i.e., 'As for Tom, it seems to me that she frightened (him) (but she didn't frighten me)', the VALENCE list of *przestraszyła* is empty: the subject is pro-dropped, while the object is extracted and realised in a higher local tree.

Before concluding the discussion of how the underspecified lexical entries, containing only the values of CONTENT and ARG-ST, give rise to a relatively large number of surface argument structures and their concrete realisations, let us briefly look at the representation of raising verbs, discussed in §3.

(5)
$$\begin{bmatrix} \textit{word} \\ \text{MORPH} \ \begin{bmatrix} \textit{morph} \\ \text{LEMMA zacząć} \end{bmatrix} \\ \text{SYNSEM} \ \begin{bmatrix} \textit{synsem} \\ \text{ARG-ST} \ \langle \boxed{0}, \text{VP} \begin{bmatrix} \text{CATEGORY } \textit{inf} \\ \text{CONTENT } \boxed{1} \\ \text{DEPS } \langle \boxed{0}, \ldots \rangle \end{bmatrix} \rangle \\ \text{CONTENT} \begin{bmatrix} \textit{start} \\ \text{ARG } \boxed{1} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

According to this simplified lexical entry of *zacząć* 'start', this verb has only one semantic argument, $\boxed{1}$, but two deep syntactic arguments: an infinitival verbal (VP) complement, and a subject, $\boxed{0}$, which, however, is not morphosyntactically specified in this lexical entry, but rather is identified with the surface subject of the infinitival complement. This way, whatever constraints are imposed on the realisation of the subject of the lower verb, will carry over to the subject of this raising verb, in accordance with the discussion in §3.

5 Lexemes and Forms

The view of the lexicon presented above, based on standard generative linguistic assumptions, treats the lexicon as a dictionary of *lexemes*, not as a dictionary of *forms*. However, in the process of parsing a natural language sentence, the parser needs to tokenize the input, i.e., split it into wordforms, and, for each such form, find the syntactic and semantic characteristics of this form as encoded in the lexical entry of the lexeme which the form belongs to. This section briefly explains how this problem of mapping forms to lexemes has been solved in the current (December 2003) version of a toy parser for Polish developed by the author.

The parser is an implementation of a relatively small HPSG grammar based on [15]. The grammar has been implemented in TRALE [8], a comprehensive general-purpose platform for implementing HPSG-like grammars, one of the leading platforms of this kind.

The grammar contains 5 very general binary syntactic rules: 2 rules for the realization of arguments (to the left and to the right from the syntactic head), 2 rules for the realization of adjuncts, and 1 rule for the realization of so-called fillers, i.e., extracted elements. Note that HPSG phrase structure rules are really rule schemata, and a large realistic HPSG grammar might

contain as few as 10–12 such rules. Even though this is a relatively small grammar, it contains the preliminary treatment of a few non-trivial linguistic phenomena, including long-distance extraction and adjunction, as well as various agreement and case assignment rules.

At the moment, only a handful of lexemes are implemented, as much time and effort has been devoted to the problem of mapping between forms and lexemes. TRALE does not have any built-in mechanisms for providing an access to an external morphological analyser, so it was necessary to create some glue code for this purpose.

The external morphological analyser used in the implementation effort described here is *Morfeusz*, developed by Marcin Woliński on the basis of linguistic data provided by Zygmunt Saloni, especially, his database of Polish verbs, [19], and the stemming rules published as [22]. The analyser uses the tagset described in [16,17], and it is currently employed in a morphosyntactic tagger [5,6] and in some initial Information Extraction efforts for Polish [9]. The analyser is implemented as a C library.

The glue code between the morphological analyser and the parser consists of a function in C which provides access to *Morfeusz* and suitably packages the morphological analyses returned by *Morfeusz* for particular wordforms,[4] as well as much more substantial code in SICStus Prolog, the programming language in which the TRALE system is implemented. The role of the Prolog code is to pass the input string to *Morfeusz* and to translate its morphosyntactic analyses into the corresponding HPSG structures. More specifically:

- a new Prolog predicate, `an/1`, is used for parsing a string,
- `an/1` calls the predicate `tokenise/2`, which passes the string to *Morfeusz* and returns a Prolog representation of the morphological analyses, e.g., for *faceta*, a form of *facet* 'guy' ambiguous between the accusative and the genitive: `[i(faceta,[m(facet,subst:sg:gen^acc:m1)])]`,
- the TRALE predicate for accessing the lexicon of forms, `lex/2`, is redefined in such a way that it treats the `i/2` terms illustrated above as wordforms and translates the morphological information into the corresponding HPSG structures, preserving local ambiguities where possible; e.g., `[i(faceta,[m(facet,subst:sg:gen^acc:m1)])]` is translated into the following structure, where *accgen* is a supertype of *acc* and *gen*:

(6)
$$\begin{bmatrix} word \\ \text{PHON } \langle faceta \rangle \\ \text{MORFEUSZ } \begin{bmatrix} \text{LEMMA } facet \\ \text{MORPH } subst{:}sg{:}gen\hat{}acc{:}m1 \end{bmatrix} \\ \text{SS|HEAD|CAT } \begin{bmatrix} subst \\ \text{AGR } \begin{bmatrix} c\_agr \\ \text{CASE } accgen \\ \text{GEND } m1 \\ \text{NUMB } sg \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

[4] The author is grateful to Marcin Woliński for his help with the C part of the interface.

- in the process, a relevant lexeme is identified on the basis of the form's lemma (*facet* in the example above) and its grammatical class (*subst*, i.e., noun), and the information provided by that lexeme is added to the lexical entry, e.g.:

(7)
$$\begin{bmatrix} word \\ \text{PHON } \langle facet a \rangle \\ \text{MORFEUSZ } \begin{bmatrix} \text{LEMMA } facet \\ \text{MORPH } subst{:}sg{:}gen\hat{}acc{:}m1 \end{bmatrix} \\ \text{SS|HEAD } \begin{bmatrix} \text{CAT } \begin{bmatrix} subst \\ \text{AGR } \begin{bmatrix} c\_agr \\ \text{CASE } accgen \\ \text{GEND } m1 \\ \text{NUMB } sg \end{bmatrix} \end{bmatrix} \\ \text{ARG-ST } \langle\rangle \\ \text{CONTENT } \begin{bmatrix} guy \\ \text{INDEX } \boxed{0} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

6 Concluding Remarks

The view of the syntactico-semantic lexicon presented here eschews all the deficiencies of the design of valence dictionaries discussed in §3: it is being created in a machine readable form, it takes into account the syntax-semantics interface, it has at its disposal a formal apparatus expressive enough to adequately describe raising verbs and other complex lexemes, and it ensures that the information provided in lexical entries of lexemes is true for all forms of these lexemes. It is our hope that the work whose early stages are reported here will eventually lead to the creation of a reusable large-scale syntactico-semantic lexicon of Polish.

References

1. Mirosław Bańko, editor. *Inny słownik języka polskiego*. Wydawnictwo Naukowe PWN, Warsaw, 2000.
2. Leonard Bolc, Krzysztof Czuba, Anna Kupść, Małgorzata Marciniak, Agnieszka Mykowiecka, and Adam Przepiórkowski. A survey of systems for implementing HPSG grammars. IPI PAN Research Report 814, Institute of Computer Science, Polish Academy of Sciences, 1996.
3. Gosse Bouma, Robert Malouf, and Ivan A. Sag. Satisfying constraints on extraction and adjunction. *Natural Language and Linguistic Theory*, 19(1):1–65, 2001.
4. Ann Copestake. *Implementing Typed Feature Structure Grammars*. CSLI Publications, Stanford, CA, 2002.
5. Łukasz Dębowski. A reconfigurable stochastic tagger for languages with complex tag structure. In *Proceedings of* Morphological Processing of Slavic Languages, *EACL 2003*, 2003.
6. Łukasz Dębowski. Trigram morphosyntactic tagger for Polish. In *Proceedings of IIS:IIPWM 2004*, 2003.

7. Norbert Morciniec, Lesław Cirko, and Ryszard Ziobro. *Słownik walencyjny czasowników niemieckich i polskich / Wörterbuch zur Valenz Deutscher und Polnischer Verben*. Wydawnictwo Uniwersytetu Wrocławskiego, Wrocław, 1995.
8. Gerald Penn, Detmar Meurers, Kordula De Kuthy, Mohammad Haji-Abdolhosseini, Venessa Metcalf, Stevan Müller, and Holger Wunsch. *Trale Milca Environment v. 2.5.0. User's Manual (Draft)*, May 2003.
9. Jakub Piskorski, Peter Homola, Małorzata Marciniak, Agnieszka Mykowiecka, Adam Przepiórkowski, and Marcin Woliński. Information extraction for Polish using the SProUT platform. In *Proceedings of IIS:IIPWM 2004*, 2003.
10. Kazimierz Polański, editor. *Słownik syntaktyczno-generatywny czasowników polskich*. Zakład Narodowy im. Ossolińskich / Instytut Języka Polskiego PAN, Wrocław / Kraków, 1980–1992.
11. Carl Pollard and Ivan A. Sag. *Head-driven Phrase Structure Grammar*. Chicago University Press / CSLI Publications, Chicago, IL, 1994.
12. Adam Przepiórkowski. *Case Assignment and the Complement-Adjunct Dichotomy: A Non-Configurational Constraint-Based Approach*. Ph. D. dissertation, Universität Tübingen, Germany, 1999.
13. Adam Przepiórkowski. Long distance genitive of negation in Polish. *Journal of Slavic Linguistics*, 8:151–189, 2000.
14. Adam Przepiórkowski. On the computational usability of valence dictionaries for Polish. IPI PAN Research Report 971, Institute of Computer Science, Polish Academy of Sciences, 2003. To appear in Proceedings of *Slovko2003*.
15. Adam Przepiórkowski, Anna Kupść, Małgorzata Marciniak, and Agnieszka Mykowiecka. *Formalny opis języka polskiego: Teoria i implementacja*. Akademicka Oficyna Wydawnicza EXIT, Warsaw, 2002.
16. Adam Przepiórkowski and Marcin Woliński. A flexemic tagset for Polish. In *Proceedings of* Morphological Processing of Slavic Languages, *EACL 2003*, 2003.
17. Adam Przepiórkowski and Marcin Woliński. 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*, 2003.
18. Frank Richter. *A Mathematical Formalism for Linguistic Theories with an Application in Head-Driven Phrase Structure Grammar*. Ph. D. dissertation, Universität Tübingen, 2000.
19. Zygmunt Saloni. *Czasownik polski. Odmiana, słownik*. Wiedza Powszechna, Warsaw, 2001.
20. Marek Świdziński. Dalsze kłopoty z bezokolicznikiem. In Jadwiga Sambor, Jadwiga Linde-Usiekniewicz, and Romuald Huszcza, editors, *Językoznawstwo synchroniczne i diachroniczne*, pages 303–314. Wydawnictwa Uniwersytetu Warszawskiego, Warsaw, 1993.
21. Marek Świdziński. Syntactic dictionary of Polish verbs. Ms., University of Warsaw and Universiteit van Amsterdam, 1994.
22. Jan Tokarski. *Schematyczny indeks* a tergo *polskich form wyrazowych*. Wydawnictwo Naukowe PWN, Warsaw, 1993. Elaborated and edited by Zygmunt Saloni.
23. Atro Voutilainen. Morphological disambiguation. In F. Karlsson, A. Voutilainen, J. Heikkilä, and A. Anttila, editors, *Constraint Grammar: A Language-Independent Systsem for Parsing Unrestricted Text*, pages 165–284. Mouton de Gruyter, Berlin, 1995.