The Unbearable Lightness of Tagging*
A Case Study in Morphosyntactic Tagging of Polish

Adam Przepiórkowski
Institute of Computer Science
Polish Academy of Sciences
adamp@ipipan.waw.pl

Marcin Woliński
Institute of Computer Science
Polish Academy of Sciences
wolinski@ipipan.waw.pl

Abstract

The article takes a step back and examines the notion of *part of speech* (POS), arguing that POS tagsets should be constructed more carefully and, in effect, should be light in at least three senses: 1) they should pay less heed to the traditionally ill-defined notion of POS, 2) they should adopt clear POS delimitation criteria based on solely formal (morphological and morphosyntactic) properties, and 3) tags should be assigned to light units, typically not longer than orthographic words. A tagset for Polish constructed on the basis of such criteria is presented.

1 Introduction

Morphosyntactic, or part of speech (POS), tagging is often considered to be an uninteresting aspect of natural language processing (NLP); after all, robust morphological analyzers and good-accuracy disambiguators exist for many languages, while the same cannot be said about, e.g., comprehensive computational grammars or dialogue models.[1] Even within corpus linguistics, morphological annotation is considered a done deal, with much annotation work focusing on higher levels of linguistic representation (mainly syntax and, more recently, semantics).

While there exist many morphological analyzers for Polish and other Slavic languages which are certainly useful and robust, we argue here that they often are linguistically naïve, which has the practical consequence of being suited just for the one specific task at hand. The main aim of this paper is to argue for the need for a clear design of POS tagsets on the basis of transparent morphosyntactic criteria.

The following section, sec. 2, discusses various features of current tagsets which seem problematic from the point of view of linguistic theory and reusability. Then, section 3 presents a tagset for Polish designed with the aim of avoiding those problems. Finally, section 4 concludes the article.

2 Traditional POS Tagsets

Morphological classes, or parts of speech, assumed within various tagsets are usually taken over more-or-less verbatim from traditional grammars. For example, the Multext-East (Erjavec, 2001) tagset for Czech assumes the following parts of speech: **noun**, **verb**, **adjective**, **pronoun**, **adverb**, **adposition**, **conjunction**, **numeral**, **interjection**, **residual**, **abbreviation** and **particle**.

While tagsets based on such POSs are well-grounded in linguistic tradition, they do not represent a logically valid classification of wordforms in the sense that the criteria which seem to underlie these classes do not always allow to uniquely classify a given word. We will support this criticism with two examples.

*With apologies to Milan Kundera.

[1] To avoid terminological confusion, we assume here that a POS *tagger* has the combined functionality of a *morphological analyzer* (which may produce ambiguous results for a given wordform) and a POS *disambiguator* (which selects the 'right' tag(s) for a given context).

Let us first of all consider the classes **pronoun** and **adjective**. The former is morphosyntactically very heterogeneous:

- some pronouns inflect for ***gender*** (e.g., the demonstrative pronoun *ten*, the possessive pronoun *mój*, but not the interrogative pronoun *kto* or the negative pronoun *nikt*);
- some pronouns (the so-called personal pronouns), but not all, inflect for ***person***;
- some pronouns, but not all, inflect for ***number*** (e.g., the pronouns *kto* 'who' and *co* 'what' do not inflect for ***number***);
- the short reflexive pronoun *się* does not overtly inflect at all, although it may be construed as a weak form of the anaphoric pronoun *siebie*;
- the anaphoric pronoun *siebie* overtly inflects for ***case*** only and it is the only pronoun with such morphosyntactic properties.

It seems that the class of **pronoun**s is defined mainly, if not solely, on the basis of semantic intuition. On the other hand, **adjective**s are well-defined morphosyntactically, as the forms inflecting for ***gender***, ***number*** and ***case***, but not, say, ***person*** or ***voice***.

Now, according to these definitions, it is not clear, whether so-called possessive pronouns, such as *mój* 'my' should be classified as **pronoun**s or **adjective**s: semantically they belong to the former class, while morphosyntactically — to the latter. (Traditionally, it is classified as a pronoun, of course.)

A very similar problem arises in case of so-called ordinal numerals, e.g., *trzeci* 'third', which are morphosyntactically indistinguishable from ordinary adjectives and which, nevertheless, are traditionally classified as numerals, not as adjectives.

Another, and perhaps more serious example concerns so-called *-nie/-cie* gerunds, i.e., *substantiva verbalia* (Puzynina, 1969) such as *pić*::*picie* 'to drink::drinking', *browsować*::*browsowanie* 'to browse::browsing'.[2] These are nominal forms in the sense that they have ***gender*** (always *n2*) and inflect for ***case*** and, potentially, for ***number***, but they are also productively related to verbs, have the category of ***aspect*** and inflect for ***negation***. As such, they do not comfortably fit into the traditional class **noun**, whose members do not have ***aspect*** or ***negation***, nor do they belong to the class **verb**, whose members have no ***case***. A similar difficulty is encountered also in case of adjectival participles, which — apart from the adjectival inflectional categories of ***gender***, ***number*** and ***case*** — also inflect for ***negation*** and have ***aspect***.

All those problems stem from the uncritical adoption of traditional and sometimes ill-defined POS classes, such as 'pronoun' or vaguely delimited classes such as 'verb' or 'noun', and from the fact that POS classes and categories are often chosen on the basis of a mix of morphological, syntactic and semantic criteria, e.g., 'gender' in Slavic is sometimes defined on the basis of mixed morphosyntactic and semantic properties, and so are 'pronoun' and 'numeral'. Apart from those, we have identified some other problems with traditional approaches to POS tagging:

- mixing morphosyntactic annotation with what might be called dictionary annotation; e.g., tagsets often include tags for proper names or morphosyntactically transparent collocations, which — in our opinion — do not belong to the realm of POS annotation;
- sometimes the priorities of such mixed criteria are unclear, e.g., should the preposition *of* in *District of Columbia* be tagged as an ordinary preposition, or should it have the 'proper' tag as it is a part of a proper name?
- ignoring the finer points of the morphosyntactic system of a given language, e.g., the multitude of genders in languages such as Polish, or categories such as 'post-prepositionality' and 'accommodability' (see below);

[2]The second pair illustrates the productivity of the gerundial derivational rule: *browsować* is, of course, a very recent borrowing.

- unclear segmentation rules (should so-called analytic tenses or reflexive verbs be treated as single units for the purpose of annotation?); segmentation rules are often ignored in descriptions of tagsets, even though, as we argue at length in section 3.1 below, they are integral and often crucial part of tagset design.

In this paper we argue for a clear delimitation of morphosyntactic tagging, where morphosyntactic tagsets are based only on well-defined morphological criteria. Such tagsets are 'light' in at least three senses:

- they partially evade the burden of linguistic tradition;
- they ignore semantic, pragmatic and — to a large extent — syntactic information;
- tags are assigned to very light units, typically single orthographic words.

The remainder of the paper presents such a tagset for Polish, developed within a Polish corpus project[3] and deployed by a stochastic tagger of Polish (Dębowski, 2003).

3 A Light Tagset for Polish

The tagset presented in this section is based on the following design assumptions:

- what is being tagged is a single orthographic word or, in some well-defined cases, a part thereof; multi-word constructions, even those sometimes considered to be morphological formations (so-called analytic forms) or dictionary entries (proper names), should be considered by a different level of processing;[4] cf. 3.1;
- grammatical categories reflect various oppositions in the morphological system, even those oppositions which pertain to single grammatical classes and are not recognized by traditional grammars; cf. 3.2;
- the main criteria for delimiting grammatical classes are morphological (how a given form inflects; e.g., nouns inflect for case, but not for gender) and morphosyntactic (in which categories it agrees with other forms; e.g., Polish nouns do not inflect for gender but they agree in gender with adjectives and verbs); semantic criteria are eschewed; cf. 3.3.

3.1 Segmentation

By segmentation, or tokenization, we mean the task of splitting the input text into tokens, i.e., segments of texts which are subject to morphosyntactic tagging. Such tokens must minimally have the following two properties:

- tokens must be contiguous;
- tokens must be disjoint.

These assumptions seem trivial, but when taken seriously, they turn out to have some interesting consequences.

In case of inherently reflexive verbs, such as *bać się* 'to be afraid', the reflexive marker (RM) *się* is sometimes analyzed as being a morphological part of the reflexive verb, i.e., according to such a view, the complex *bać się* should have just one tag assigned. This, however, would violate the disjointness property above, as (1) illustrates.

(1) Boję się głośno roześmiać.
fear-*rv*-I RM loudly laugh-*inf.rv*
'I'm afraid to laugh loudly.'

This sentence exemplifies the so-called haplology of the Polish reflexive marker (Kupść, 1999): just one reflexive marker *się* occurs with two inherently reflexive verbs: *boję się* and *roześmiać się*. If inherently reflexive verbs were to be segmented jointly with their reflexive markers, the tokenizer would have to interpret whether *się* is part of the 'word' *boję się*, or the 'word' *roześmiać się*.[5]

[3] *An Annotated Internet-Accessible Corpus of Written Polish (with Emphasis on NLP Applications)*, a 3-year project financed by the State Committee for Scientific Research.

[4] In case of proper names, there exist many dedicated algorithms and systems for finding them in texts, often developed within the Message Understanding Conference series.

[5] Because of the criterion of contiguity it would have to choose the former alternative in this case.

Thus, it seems reasonable to tokenize the reflexive marker separately, and to interpret it at a level aware of such linguistic phenomena as haplology.

Of course, splitting reflexive marker from the corresponding inherently reflexive verb is also required to satisfy the criterion of contiguity: in Polish, the reflexive marker may be separated from the verb by an in principle unlimited number of words. A purer case of an application of the disjointness criterion is the haplology of full-stop, where the sentence-final dot may also be an inherent part of an abbreviation which happens to be the last word in this sentence:

(2) Widziałem Tomka, Janka itp.
saw-I Tom, John etc.
'I saw Tom, John, etc.'

The two criteria mentioned above still leave much room for maneuver. In order for the result of segmentation to be maximally transparent, we propose the following guidelines:

- tokens do not contain white space;
- tokens either are punctuation marks or do not contain any punctuation marks;
- an exception to the previous guideline are certain words containing the hyphen (e.g., *Daimler-Benz*, *mass-media*, *s-ka* = an abbreviation of *spółka* 'company', etc.) and apostrophe used when inflecting foreign names (e.g. *Lagrange'a*); they are given by a list.

Note that it does not follow from the guidelines above that orthographic words cannot be further split into POS tokens, but — again — the cases where such intra-word segmentation occurs should be well-defined.

We propose to split orthographic words when they contain what sometimes is called *mobile* or *floating inflection*:

(3) a. Dawno nie widziała**m** Janka.
long time not saw-I John
'I haven't seen John for a long time.'
b. Dawno**m** nie widziała Janka.

(4) a. Kiedyś poszedł***bym*** tam.
once would go-I there
'I'd go there once.'
b. Kiedyś ***bym*** tam poszedł.

It is clear that in the b. examples above, the detached morphemes *-m* (bearing person and number information) and *bym* (i.e., the subjunctive particle *by* and the bound morpheme *-m*) play the same role as in the corresponding a. examples.

While the example in (3b) can be considered dated and hard to spot in real texts, there are Polish sentences where the detachment of movable inflection from the verb is obligatory:

(5) Chciałbym, żeby**ś** przyszedł.
would like-I that-you come
'I'd like you to come.'

Even though there are only a few words (*żeby*, *aby*, *gdyby*, ...) that occur in such constructs, sentences of this type are quite common.

The 'floating inflections' in Polish should be treated as weak forms of the verb *być*. These are the same forms as the one in the following sentence (Saloni and Świdziński, 1998):

(6) Świniaś!
pig be-you
'You're a pig!'

In fact, such floating inflections have been reanalyzed in recent linguistic literature as auxiliaries, i.e., essentially syntactic elements (Borsley and Rivero, 1994; Bański, 2000).[6] For these reasons, we propose to tokenize orthographic wordforms such as *poszedłbym* into three POS tokens: *poszedł*, *by* and *m*. This in effect means that, for Polish, segmentation must be treated as an inherent part of morphological analysis.[7]

Arguments can also be given for splitting the negative prefix *nie* from participles, despite orthographic tradition, because they play the same morphosyntactic role as the verbal negative marker *nie*, e.g., participate in negative concord (Przepiórkowski and Kupść, 1999) and trigger the

[6]This is an oversimplification; see the work cited here for details.

[7]To simplify the processing, a 'presegmentation' phase is conceivable where text is split into orthographic words, which could be further split during the morphological processing.

so-called genitive of negation (Przepiórkowski, 2000):

(7) a. Janek pisze (*żadną) książkę.
John writes no-*acc* book-*acc*
'John is writing a book / *no book.'

b. Janek **nie** pisze (**żadnej**) książki.
John not writes no-***gen*** book-***gen***

(8) a. Janek, piszący (*żadną) książkę...
John writing no-*acc* book-*acc*...

b. Janek, **nie**piszący (**żadnej**) książki...
John not-writing no-***gen*** book-***gen***...

However, as argued in (Kupść and Przepiórkowski, 2002), Polish verbal negation should rather be treated as verbal prefixation, contrary to what Polish orthography suggests, so for the purposes of the tagset presented here, we assume that negated participles are single tokens, distinguished from their non-negated counterparts via the morphological category of negation.

3.2 Morphological Categories

Although we proposed ignoring some information often present in tagsets, e.g., the 'proper noun' vs. 'common noun' distinction, we argue that morphological categories should be taken seriously and should be as detailed as possible. For this reason, apart from the traditional categories of ***gender***, ***person***, ***number***, ***case***, ***degree*** and ***aspect***, we assume the following less-standard grammatical categories:

- ***negation***: a category of various de-verbal classes, e.g., participles such as *(nie)piszący* '(not) writing'; the relavant values are *aff* and *neg*;
- ***accentability*** (Polish: *akcentowość*): a category of nominal pronouns; *akc* (*jego*), *nakc* (*go*);
- ***post-prepositionality*** (Polish: *poprzyimkowość*): a category of nominal pronouns; *praep* (*niego*, *-ń*), *npraep* (*jego*, *go*);
- ***accommodability*** (Polish: *akomodacyjność*): a category of numerals; *congr* (*dwaj*, *trzej*), *rec* (*dwóch*, *trzech*);
- ***agglutination*** (Polish: *aglutynacyjność*): *nagl* (*niósł*), *agl* (*niosł-*);
- ***vocability*** (Polish: *wokaliczność*): *wok* (*-em*, *-eś*), *nwok* (*-m*, *-ś*).

Those categories, although non-standard, are based on important work by Zygmunt Saloni and his colleagues (Saloni, 1976; Saloni, 1977; Gruszczyński and Saloni, 1978; Bień and Saloni, 1982).

For completeness, the values of the more traditional grammatical categories are presented below:

- ***number***: *sg*, *pl*;
- ***case***: *nom*, *acc*, *gen*, *dat*, *inst*, *loc*, *voc*;
- ***person***: *pri*, *sec*, *ter*;
- ***degree***: *pos*, *comp*, *sup*;
- ***aspect***: *imperf*, *perf*;

The one traditional category ommitted above is ***gender***:

- ***gender***: three masculine genders *m1* (*facet*), *m2* (*koń*), *m3* (*stół*), the feminine gender *f* (*kobieta, żyrafa, książka*), two neuter genders *n1* (*dziecko*), *n2* (*okno*), and three *plurale tantum* genders *p1* (*wujostwo*), *p2* (*drzwi*), *p3* (*okulary*).

It may seem surprising, at first, to see 9 gender values in an Indo-European language (as opposed to, say, a Bantu language), but this position is well argued for by (Saloni, 1976), who distinguishes those genders on the basis of agreement with adjectives and numerals;[8] we will not attempt to further justify this position here.

[8]Elsewhere, we propose reducing the number of genders, essentially, by factoring out the number information (Woliński, 2001) or the information about agreement with numerals (Przepiórkowski et al., 2002), but for the purposes of this tagset we assume the original repertoire of genders proposed by Saloni.

3.3 Morphological Classes

For the reasons amply discussed in section 2, and following the general approach of (Saloni, 1974) and (Bień, 1991), we propose to derive the notion of grammatical class from the notion of *flexeme* introduced by Bień, where flexeme is understood as a morphosyntactically homogeneous set of forms belonging to the same lexeme.

For example, a typical Polish verbal lexeme contains a number of personal forms, a number of impersonal forms, as well as, depending on a particular understanding of the notion of lexeme, various deverbal forms, such as participles and gerunds. These forms have very different morphosyntactic properties: finite non-past tense forms have the inflectional categories of person and number, adjectival participles have the inflectional properties of non-gradable adjectives and, additionally, inflect for negation and have aspect, gerunds inflect for case and, at least potentially, for number, but not for person, etc. Ideally, flexemes are subsets of such lexemes consisting of those forms which have the same inflectional properties: all verbal forms of given lexeme with the inflectional category of person and number are grouped into one flexeme, other forms belonging to this lexeme, but with adjectival inflectional properties, are grouped into another flexeme, those forms, which inflect for case but not for gender are grouped into a gerundial flexeme, etc. Each of such flexemes is characterized by a set of grammatical categories it inflects for and, perhaps, a set of grammatical categories it has lexically set (e.g., the gender of nouns).

Now, given the notion of flexeme, it is natural to define grammatical classes as *flexemic classes*, i.e., classes of flexemes with the same inflectional characteristics. For example, the grammatical class **non-past verb** contains exactly those flexemes which inflect for person and number, and nothing else, and which also have the lexical category of aspect; the class **noun** contains exactly those flexemes which inflect for number and case, and have gender; the class **gerund** contains exactly those flexemes which inflect for number, case and negation, and have lexical gender (always neuter, *n2*, in case of gerunds) and aspect; etc.

It should be noted that, despite the way flexemes have been defined above, the notion of lexeme is of only secondary importance here: it is invoked for the purpose of assigning a lemma to a given form (e.g., a gerundial form such as *przyjściem* 'coming-*inst*' will be lemmatized to the infinitival form *przyjść* 'to come': even though the form *przyjść* does not belong to the *flexeme* of *przyjściem*, it does belong to the *lexeme* containing *przyjściem*). Moreover, just as in case of deciding whether two forms belong to the same lexeme, also classification of two wordforms to the same flexeme requires some semantic intuition: thus, e.g., *pies* 'dog-*nom*' and *psem* 'dog-*inst*' belong to the same (f)lexeme, and so do *rok* 'year-*sg*' and *lata* 'year-*pl*', but *pies* 'dog' and *suka* 'bitch' do not.

The basic classification of flexemes into grammatical ('flexemic') classes is given by the following decision tree:

```
Inflects for case?
YES: Inflects for negation?
  YES: Inflects for gender?
    YES: 1. adjectival participle
    NO:  2. gerund
  NO:  Inflects for gender?
    YES: Has person?
      YES: 3. nominal pronoun
      NO:  Inflects for number?
        YES: 4. adjective
        NO:  5. numeral
    NO:  6. noun
NO: Inflects for gender?
  YES: 7. l-participle
  NO:  Inflects for number?
    YES: 8. (inflecting verbal forms)
    NO:  9. ('non-inflecting' verbal
         forms, adverbs, prepositions,
         conjunctions)
```

Note that most of the classes in the 'inflects for case' branch of the tree already are reasonable POSs, i.e., they correspond to traditional POSs (**noun**, **adjective**, **numeral**) or to their well-defined subsets (**nominal pronoun**, **gerund**, **adjectival participle**). It is important to realize, however, that these classes are defined mainly on the basis of the inflectional properties of their members; e.g., the class **numeral** is much narrower here than traditionally, as it does not include so-called ordinal numerals (which, morphosyntactically, are adjectives).

On the other hand, in the 'does not inflect for case' branch, only the 'inflects for gender' class corresponds to an intuitive set of forms, namely, to

so-called *l-participles* or *past participles*, i.e., verbal forms hosting 'floating inflections'; cf. *widziała* and *poszedł* in (3)–(4) above.

The class 8. above can be further partitioned according to the following criteria:

8. Has a *ter* (i.e., 3rd person) form?
   YES: 8.1. **non-past** forms, e.g., *idę* 'I am going', *pójdę* 'I will go'
   NO: Has a *pri sg* form?
     YES: 8.2. **agglutinate** (*-(e)m, -(e)ś, -śmy, -ście*)
     NO: 8.3. **imperative**

Further, we will remove from the class of **noun**s the flexeme of the strong reflexive pronoun *siebie*, which does not inflect for number and does not have overt gender:

6. Inflects for ***number***?
   YES: 6.1. true **noun**
   NO: 6.2. **siebie**

Moreover, inflectional class marked as 9. can be further split according to non-inflectional morphosyntactic properties of its members in the following way:

9. Has ***aspect***?
   YES: 9.1. non-inflecting verbal forms
   NO: Inflects for ***degree*** or derived from **adjective**?
     YES: 9.2. **adverb**
     NO: 9.3. **preposition**, **conjunction**, etc.

Note that, in order to arrive at a class close to the traditional class of **adverb**s, we had to define this class disjunctively; it should contain all adverbs inflecting for degree, at least one of which does not seem to be derived from an adjective (*bardzo* 'very'), as well as all de-adjectival adverbs, some of which do not (synthetically) inflect for degree (e.g., *antywirusowo* 'anti-virus-like', **antywirusowiej*).

If our purpose were to define a purely flexemic tagset for Polish, we would have to stop here (and remove the 'derived from **adjective**' disjunct from the subtree above). For example, it is impossible to distinguish the impersonal *-no/-to* form, the infinitive, and adverbial participle of the same lexeme on the basis of their morphosyntactic properties alone: they all lack any inflectional categories and have the lexical category of ***aspect***. For this reason, we will further partition the class 9.1. above on the basis of purely orthographic (or phonetic) information:

9.1. Ends in *-no* or *-to*?
   YES: 9.1.1. impersonal ***-no/-to*** forms (e.g., *chodzono* 'one used to walk/go', *pito* 'one used to drink')
   NO: Ends in *-ąc* or *-szy*?
     YES: 9.1.2. **adverbial participle** (e.g., *czytając* 'reading', *przeczytawszy* 'having read')
     NO: 9.1.3. **infinitive** form (e.g., *iść* 'to go'); should end in *-c* or *-ć*

Finally, the class 9.3. consists of those word-forms which do not inflect, and do not have ***aspect***, i.e.:

9.3.1. **conjunction**
9.3.2. **preposition**
9.3.3. **particle-adverb**

The first two classes are closed classes, which can be defined extensionally, by enumerating them. All other non-inflecting, non-aspectual and non-de-adjectival single-form flexemes fall into the **particle-adverb** class.

The full tagset is presented in (Przepiórkowski and Woliński, 2003).

4 Conclusions

We argued above for a 'light' approach to POS tagging, where POS tags reflect solely morphosyntactic information, without paying any heed to semantic and pragmatic information. This approach leads to well-defined POS classes with clear tests of being a member of a class based, first of all, on inflectional properties of particular forms and, secondly, on other morphosyntactic and orthographic/phonetic features. We included a detailed feasibility study showing that this approach is well-suited to Polish, a Slavic language with rich morphology. Despite this 'lightness', the morphosyntactic information in the tagset we arrived at is more detailed in most, if not all, tagsets for Polish.

This approach may be difficult to accept from the point of view of linguistic tradition (hence the title of this paper), as it does not allow to define classes such as 'pronoun' or 'numeral' in the traditional sense of these terms. We claim, however, that this is a feature of our approach, not a bug: the traditional notions 'pronoun' and 'numeral' are semantic in nature and should be confined to the semantic level of processing.

Acknowledgments

The tagset described here was highly influenced by many discussions with Łukasz Dębowski, by the insightful comments we received from Zygmunt Saloni, and by the various remarks from Elżbieta Hajnicz, Monika Korczakowska and Beata Wierzchołowska. The research reported here was partly supported by the KBN (State Committee for Scientific Research) grant 7 T11C 043 20.

References

Piotr Bański. 2000. *Morphological and Prosodic Analysis of Auxiliary Clitics in Polish and English*. Ph. D. dissertation, University of 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.

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

Robert D. Borsley and Adam Przepiórkowski, editors. 1999. *Slavic in Head-Driven Phrase Structure Grammar*. CSLI Publications, Stanford, CA.

Robert D. Borsley and María Luisa Rivero. 1994. Clitic auxiliaries and incorporation in Polish. *Natural Language and Linguistic Theory*, 12:373–422.

Łukasz Dębowski. 2003. Reconfigurable stochastic tagger for languages with complex tag structure. EACL 2003, *Morphological Processing of Slavic Languages*.

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

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

Anna Kupść and Adam Przepiórkowski. 2002. Morphological aspects of verbal negation in Polish. In *Proceedings of the Second European Conference on Formal Description of Slavic Languages, Potsdam, Germany, November 20–22, 1997*.

Anna Kupść. 1999. Haplology of the Polish reflexive marker. In Borsley and Przepiórkowski (Borsley and Przepiórkowski, 1999), pages 91–124.

Adam Przepiórkowski and Anna Kupść. 1999. Eventuality negation and negative concord in Polish and Italian. In Borsley and Przepiórkowski (Borsley and Przepiórkowski, 1999), pages 211–246.

Adam Przepiórkowski and Marcin Woliński. 2003. A flexemic tagset for Polish. EACL 2003, *Morphological Processing of Slavic Languages*.

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

Adam Przepiórkowski. 2000. Long distance genitive of negation in Polish. *Journal of Slavic Linguistics*, 8:151–189.

Jadwiga Puzynina. 1969. *Nazwy czynności we współczesnym języku polskim*. Wydawnictwo Naukowe PWN, Warsaw.

Zygmunt Saloni and Marek Świdziński. 1998. *Składnia współczesnego języka polskiego*. Wydawnictwo Naukowe PWN, Warsaw, 4th (changed) edition.

Zygmunt Saloni. 1974. Klasyfikacja gramatyczna leksemów polskich. *Język Polski*, LIV(1):3–13.

Zygmunt Saloni. 1976. Kategoria rodzaju we współczesnym języku polskim. In *Kategorie gramatyczne grup imiennych we współczesnym języku polskim*, pages 41–75. Ossolineum, Wrocław.

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

Marcin Woliński. 2001. Rodzajów w polszczyźnie jest osiem. In Włodzimierz Gruszczyński, Urszula Andrejewicz, Mirosław Bańko, and Dorota Kopcińska, editors, *Nie bez znaczenia... Prace ofiarowane Profesorowi Zygmuntowi Saloniemu z okazji jubileuszu 15000 dni pracy naukowej*, pages 303–305. Wydawnictwo Uniwersytetu Białostockiego, Białystok.