A Flexemic Tagset for 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 notes certain weaknesses of current efforts aiming at the standardization of POS tagsets for morphologically rich languages and argues that, in order to achieve clear mappings *between* tagsets, it is necessary to have clear and formal rules of delimiting POSs and grammatical categories *within* any given tagset. An attempt at constructing such a tagset for Polish is presented.

1 Introduction

The aim of this article is to address one of the objectives of the EACL 2003 workshop on *Morphological Processing of Slavic Languages*, namely, to "try to reveal lexical structures necessary for morphological analysis and... discuss standardization efforts in the field that can, for instance, enable transfer of applied methods from one language to the other or inform the annotation of morphological information in corpora."

One admirable standardization effort in the field of Slavic part of speech (POS) tagging has been the Multext-East project (Erjavec, 2001), one of whose aims was to construct mutually compatible tagsets for 8 European languages, including 4 Slavic languages (originally Bulgarian, Czech and Slovene, later extended to Croatian); additionally, a Multext-East-style tagset for Russian was constructed at the University of Tübingen (`http://www.sfb441.uni-tuebingen.de/c1/tagset.html`). Those tagsets are based on a common repertoire of grammatical classes (POSs; e.g., 'verb', 'noun', 'adjective', etc.) and grammatical categories (e.g., 'case', 'person', 'gender', etc.), and each tagset uses just a subset of those grammatical classes and categories.

Despite the considerable success of Multext-East, and the apparent uniformity of the resulting tagsets, certain weaknesses of this approach are clear. First of all, the relative uniformity of the POS classes across the 8 languages was attained at the cost of introducing the grammatical category 'type' whose values reflect the considerable differences between POS systems of the languages involved. Second, it is not clear that various grammatical categories and their values have the same interpretation in each language; for example, it is rather surprising that only the Romanian tagset explicitly mentions strong and weak pronominal forms, it is not clear whether negative pronouns in Romanian, Slovene, Czech and Bulgarian are negative in the same sense of participating in Negative Concord, it is not clear why Romanian has negative adverbs while, say, Czech lacks them, etc. Finally, and most importantly from our point of view, the approach adopted by Multext-East does not clearly reflect cross-linguistic correspondences, such as the one mentioned in (Erjavec, 2001), that "in the Romanian case system the value 'direct' conflates 'nominative' and 'accusative', while the value 'oblique' conflates 'genitive' and 'dative'." Such correspondences are not exceptional, e.g., the at least three masculine genders of Polish (Mańczak, 1956; Saloni, 1976) are mapped into the single masculine gender of many

other languages, the dual and the plural numbers of some languages (Slovene, Czech) are mapped to plural of other languages, etc.

In more general terms, we have identified the following features of currently used tagsets for Slavic in general and Polish in particular which seem problematic from the point of view of their reusability and cross-linguistic applicability:

- uncritical adoption of traditional and sometimes ill-defined POS classes, such as 'pronoun' or vaguely delimited classes such as 'verb' or 'noun' (it is often not clear whether gerunds are 'verbs' or 'nouns' in such classifications);
- 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';
- 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);
- unclear segmentation rules (should so-called analytic tenses or reflexive verbs be treated as single units for the purpose of annotation?).

The main thesis of this paper is that, in order for a tagset to be reusable and comparable with similar tagsets for related languages, it must be based on a homogeneous set of clear formal (morphological and morphosyntactic) criteria. Only once such criteria for delimiting grammatical classes and categories are presented in detail, can those classes and categories be mapped to grammatical classes and categories of other similarly constructed tagsets.

The remainder of the paper presents such a tagset for Polish, developed within a Polish corpus project[1] and deployed by a stochastic tagger of Polish (Dębowski, 2003).

2 A Flexemic 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;[2] cf. 2.1;
- 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. 2.2;
- 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. 2.3.

2.1 Segmentation

By segmentation, or tokenization, we mean the task of splitting the input text into tokens, i.e., seg-

[1] *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.

[2] 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.

ments of texts which are subject to morphosyntactic tagging. We propose the following guidelines for segmentation (for a more complete discussion see our other article in this volume):

- 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., *mass-media*, *s-ka* = an abbreviation of *spółka* 'company', etc.) and apostrophe used in Polish when inflecting foreign names (e.g. *Lagrange'a*); they are given by a list.

Those guidelines do not preculde the situation where an orthographic word is split into several POS tokens. For example, in the case of Polish past tense finite verbs, the morpheme bearing information on person and number can be attached to the verb itself (1a) or to some other word within the sentence (1b). For that reason we always consider such a 'floating inflection' morpheme as a separate segment.[3]

(1) a. Dlaczego mi nie powiedziałaś?
Why I-*dat* not told be-you
'Why haven't you told me?'

b. Dlaczegoś mi nie powiedziała?
Why be-you I-*dat* not told

2.2 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.

What follows is the complete list of morphological categories assumed in the proposed tagset:

- ***number***: *sg*, *pl*;
- ***case***: *nom*, *acc*, *gen*, *dat*, *inst*, *loc*, *voc*;
- ***gender***: masculine personal *m1* (*facet*), masculine animate *m2* (*koń*), masculine inanimate *m3* (*stół*), feminine *f* (*kobieta, żyrafa, książka*), two neuter genders *n1* (*dziecko*), *n2* (*okno*), and three *plurale tantum* genders *p1* (*wujostwo*), *p2* (*drzwi*), *p3* (*okulary*);
- ***person***: *pri*, *sec*, *ter*;
- ***degree***: *pos*, *comp*, *sup*;
- ***aspect***: *imperf*, *perf*;
- ***negation***: *aff*, *neg*;
- ***accentability*** (Pol.: *akcentowość*): *akc*, *nakc*;
- ***post-prepositionality*** (Pol.: *poprzyimkowość*): *praep*, *npraep*;
- ***accommodability*** (Pol.: *akomodacyjność*): *congr*, *rec*;
- ***agglutination*** (Pol.: *aglutynacyjność*): *nagl*, *agl*;
- ***vocability*** (Pol.: *wokaliczność*): *wok*, *nwok*.

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;[4] we will not attempt to further justify this position here.

Negation is a category of various de-verbal classes, e.g., participles. Since we assume that the words *piszący* 'writing' and *niepiszący* 'not writing' have the same lemma *pisać* 'to write', these words have to be distinguished with this morphological category.

The category of accentability is used to differentiate accented forms of nominal pronouns (e.g. *jego*, *mnie*) from weak forms (*go*, *mi*). It roughly corresponds to the category of *clitic* used in Multext-East.

Post-prepositionality is another category of nominal pronouns. It differentiates special forms

[3] Segmentation, as understood in the present context, is discussed at length in (Przepiórkowski and Woliński, 2003).

[4] 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.

used only directly after a preposition (e.g., *niego*, *-ń*) from forms that can be used in other contexts (*jego*, *go*).

The category of accomodability is important for the description of Polish numeral-nominal phrase. Some Polish numerals have forms that agree in case with noun (marked *congr*), as well as forms that require a noun in genitive case (marked *rec*):

(2) Przyszli dwaj chłopcy.
came two-*nom.congr* boys-*nom*
'Two boys came.'

(3) Przyszło dwóch/dwu chłopców
came two-*nom.rec* boys-*gen*
'Two boys came.'

The need for the category of agglutination is a result of the way past tense verb forms are segmented (cf. (1) in sec. 2.1). For the majority of Polish verbs the form used for the first and the second person is the same as the third person form:

(4) a. Ty **przyszedł**eś.
you came
b. On **przyszedł**.
he came

But for some verbs these forms differ:

(5) a. Ty **niosł**(*nagl*)eś.
you carried
b. On **niósł**(*agl*).
he carried

Vocability distinguishes those 'floating' forms of the verb *być* 'to be' which attach to consonant-final forms (*wok*, e.g., *-em*) from the forms which attach to vowel-final forms (*nwok*, e.g., *-m*).

Various non-standard categories used above, such as post-prepositionality, accomodability and agglutination, are based on important work by Zygmunt Saloni and his colleagues (Saloni, 1976; Saloni, 1977; Gruszczyński and Saloni, 1978; Bień and Saloni, 1982).

2.3 Morphological Classes

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 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.

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, but not all, inflect for ***person***;
- some pronouns, but not all, 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*.

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.)

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'.[5] 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***.

For this reason, 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)
```

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

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. *powiedziała* in (1) 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
   NO: Has a *pri sg* form?
     YES: 8.2. **agglutinate**
          (*-(e)m*, *-(e)ś*, *-śmy*, *-ście*)
     NO: 8.3. **imperative**

Non-past verb forms correspond to present tense for imperfective verbs (e.g., *idę* 'I am going') and future tense for perfective ones (e.g., *pójdę* 'I will go').

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.

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 table on the next page presents the complete repertoire of grammatical classes and their respective inflectional ('$\oplus$') and lexical ('$\odot$') categories. Some more ephemeral classes not mentioned in the decision tree are briefly described below (a more complete description of a previous version of this tagset is available in (Woliński and Przepiórkowski, 2001)).

For Polish nouns of masculine personal (*m1*) gender a stylistically marked form is possible besides a 'regular' form for nominative and vocative

	number	case	gender	person	degree	aspect	negation	accent.	post-prep.	accom.	aggl.	vocab.
noun	⊕	⊕	⊙									
depreciative noun	⊙	⊕	⊙									
adjective	⊕	⊕	⊕		⊕							
ad-adjectival adjective												
post-prepositional adjective												
adverb					⊕							
numeral	⊙	⊕	⊕							⊕		
pronoun (non-3rd person)	⊙	⊕	⊕	⊙				⊕				
pronoun (3rd person)	⊕	⊕	⊕	⊙				⊕	⊕			
pronoun *siebie*		⊕										
non-past verb	⊕			⊕		⊙						
future *być*	⊕			⊕		⊙						
agglut. *być*	⊕			⊕		⊙						⊕
l-participle	⊕		⊕			⊙					⊕	
imperative	⊕			⊕		⊙						
-no/*-to*						⊙						
infinitive						⊙						
adv. pres. prtcp.						⊙						
adv. anter. prtcp.						⊙						
gerund	⊕	⊕	⊙			⊙	⊕					
adj. act. prtcp.	⊕	⊕	⊕			⊙	⊕					
adj. pass. prtcp.	⊕	⊕	⊕			⊙	⊕					
winien-like verb	⊕		⊕			⊙						
predicative												
preposition		⊙										
conjunction												
particle-adverb												
alien (nominal)	⊕	⊕	⊙									
alien (other)												

case in plural (e.g., *profesory* vs. *profesorowie*). These special forms do not fit in the scheme of regular nominal inflection, and so were moved to a separate flexeme for **depreciative noun**.

Ad-adjectival adjectives are special forms of adjectives used in compounds like ***angielsko***-*polski* 'English-Polish'. Moreover, some adjectives (e.g., *polski*) have a special form that is required after some prepositions (e.g., *po* ***polsku*** 'in Polish'). This form constitutes **post-prepositional adjective** flexeme.

A few verbs (e.g., *powinien* 'should') inflect in an atypical way and lack some verbal flexemes (e.g., **imperative** and **l-participle**). ***Winien*-like** flexeme gathers present tense forms of these verbs (which accept 'floating inflection').

The class of **predicative**s consists of verbs which do not inflect at all (e.g., *warto* 'be worth', *można* 'can/may', *trzeba* 'must').

3 Conclusions

Two tagsets can be compared and respective correspondences between their grammatical classes and categories can be found more easily when the definitions of those classes and categories are stated explicitly and formulated in terms of eas-

ily verifiable formal properties of particular wordforms, such as their inflectional, morphosyntactic and derivational characteristics, and their phonological or orthographic makeup.

We presented a tagset for Polish constructed with such criteria in mind. In particular, grammatical classes are understood as classes of flexemes, i.e., they are defined on the basis of, first of all, inflectional and, secondly, morphosyntactic properties of wordforms. Further distinctions, such as those between non-inflecting forms of verbal lexemes, are also made with the avoidance of any recourse to the semantic or pragmatic properties of such forms. This allowed us to evade the controversial issues of the exact extent of such semantically-defined traditional POSs as **numeral** and **pronoun**.

Despite the evasion of semantic criteria, the resulting set of grammatical classes bears surprising affinity to traditional POSs, with classes such as **noun** and **adjective** corresponding directly to traditional POSs, and other classes, such as **non-past verb**, **l-participle** or **gerund** being proper subclasses of such traditional POSs as **verb**. Because of this fine-grainedness of the current tagset we were able to evade the controversial issues of whether to classify gerunds as **noun**s or as **verb**s, and whether to classify adjectival participles as **adjective**s or as **verb**s.

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

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.

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.

Witold Mańczak. 1956. Ile jest rodzajów w polskim? *Język Polski*, XXXVI(2):116–121.

Adam Przepiórkowski and Marcin Woliński. 2003. The unbearable lightness of tagging: A case study in morphosyntactic tagging of Polish. EACL 2003, *4th International Workshop on Linguistically Interpreted Corpora (LINC-03)*.

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

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

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 and Adam Przepiórkowski. 2001. Projekt anotacji morfosynktaktycznej korpusu języka polskiego. IPI PAN Research Report 938, Institute of Computer Science, Polish Academy of Sciences.

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.