The Unbearable Lightness of Tagging[*]
Case Study in Polish Morphology[†]

Adam Przepiórkowski and Marcin Woliński

March 15, 2002

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 linguistics representation (mainly syntax).

While there exist many morphological analyzers for Polish and other Slavic languagages which are certainly useful and robust, we argue here that they often are linguistically naïve, which has the practical consequence of lack of reusability of such tools. We have identified the following features of currently used tagsets which seem problematic from the point of view of linguistic theory and reusability:

- 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[2] 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 is *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 *depreciation* 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?).

[*] With apologies to Milan Kundera.

[†] We are grateful to Łukasz Dębowski for many helpful discussions.

[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 for a given context).

[2] Another terminological note: by POS *classes* we mean sets of morphosyntactically interpreted wordforms, essentially, partitions of the set of all bilateral wordforms of a given language, e.g., the traditional classes 'verb', 'noun', 'adjective', etc.; by POS *categories* we mean morphosyntactic properties of wordforms belonging to particular classes, e.g., *case* for nouns and adjectives, but not for verbs.

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

The rest of the paper presents a light tagset for Polish developed within a project aiming at constructing a large annotated corpus of Polish and tools for its annotation and Internet access,[3] §2 and contrasts it with more traditional tagsets proposed for Slavic languages.

2 A Light Tagset for Polish

The tagset presented in this section is based on the following 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]
- the main criteria for delimiting grammatical classes are morphological (how a given form inflects; e.g., nouns inflect for case) 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);
- the secondary criterion is orthographic; in some cases a POS class may be defined extensionally, by enumeration of its elements;
- POS annotation should be as detailed as possible, and not just confined to the repertoir of traditional morphological categories (and their traditional values).

2.1 Segmentation

By segmentation we mean the task of splitting the input text into basic tokens which can be morphosyntactically tagged. This process of tokenization should have at least the following two properties:

- tokens should be contiguous;
- tokenization should not involve any interpretation (disambiguation).

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 morphological tag assigned. This, however, would violate the 'no interpretation' property above, as (1) illustrates.

(1) Boję się głośno roześmiać.
fear-RV-I RM loudly laugh-INF.RV
'I'm afraid to laugh loudly.'

[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, project number 7 T11C 043 20.

[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.

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. 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] i.e., it would have to choose between two wrong alternatives. It seems reasonable to tokenize the reflexive marker separately instead, 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 'no interpretation' 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.); 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 morpheme *-m*) play the same role as in the corresponding a. examples. In fact, such floating inflections have been reanalyzed in recent linguistic literature as auxiliaries, i.e., essentially syntactic elements (Borsley and Rivero, 1994; Borsley, 1999; Bański, 2000).[6] For these reasons, we propose to tokenize orthographic wordforms such as *poszedłbym* into three POS tokens: *poszedł*, *by* and *m*.

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

(5) 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**

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

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

(6) 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, for the purposes of the tagset presented here, we assume that that negated participles are single tokens, distinguished from their non-negated counterparts via the morphological category of negation.

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. 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* in (6); the relavant values are AFF and NEG;
- depreciation (Polish: *deprecjatywność*): a category of nominative and accusative M1 (see gender below) nouns; NDEPR (*chłopi*), DEPR (*chłopy*);
- 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*, *dlaczego*), AGL (*niosł-*, *dlaczegó-*);
- vocability (Polish: *wokaliczność*): WOK (*-em*, *-eś*, *ze*), NWOK (*-m*, *-ś*, *z*).

Those categories, although non-standard, are based on important work by Zygmunt Saloni and his colleagues (Saloni, 1976, 1977; 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 Indioeuropean 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;[7] we will not attempt to further justify this position here.

[7] We proposed elsewhere limiting the number of genders to 8, essentially by factoring out the number information (Woliński, 2001; Przepiórkowski *et al.*, 2001), but here we assume Saloni's repertoir of genders.

2.3 Morphological Classes

2.3.1 Tradition-Driven 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[8] and the Multext-East-style tagset for Russian at the University of Tübingen[9] assume 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, i.e., 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.[10]

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.

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.[11]

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, also called *substantiva verbalia* (Puzynina, 1969), *gerundives* (Tajsner, 1990) and *verbal nouns* (Rozwadowska, 1997), e.g., *pić*::*picie* 'to drink::drinking', *browsować*::*browsowanie* 'to browse::browsing'.[12] These are nominal forms in the sense that they have gender (N2) and inflect for case and, potentially, for number, but they are also productively related to verbs and 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** (its members have no case).[13]

2.3.2 Morphosyntactically-Driven Tagset

Following the general approach of Saloni (1974) and Bień (1991), we propose to delimit parts of speech on the basis of morphosyntactic and distributional properties, constructing the criteria so that:

- they give unambiguous results for any given (bilateral) word;
- they still reflect the traditional parts of speech, to the extent to which it is possible without sacrificing the transparency of the classification.

We will classify forms occurring in natural language, Polish in this case, first of all according to their *inflectional* properties; the first rough classification is presented below as a decision tree:

[8] `http://nl.ijs.si/ME/CD/docs/mte-d11f/node34.html#SECTION00440000000000000000`.

[9] `http://www.sfb441.uni-tuebingen.de/c1/tagset.html`

[10] Although we discuss tagsets for Czech and Russian, the examples below will come from Polish.

[11] Some of them, but not all, also inflect for degree.

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

[13] A similar difficulty is encountered in case of adjectival participles, which — apart from the adjectival inflectional categories of gender, number and case — also inflect for negation and have aspect.

```
Inflects for case?
YES: Inflects for negation?
     YES: Inflects for gender?
          YES: 1. adjectival participle
          NO:  2. gerund
     NO:  Inflects for gender?
          YES: Inflects for person?
               YES: 3. nominal pronoun
               NO:  Inflects for number?
                    YES: 4. adjective
                    NO:  5. numeral
          NO:  6. noun
NO: Inflects for gender?
    YES: 7. pseudo-participle
    NO:  Inflects for number?
         YES: 8. (various inflecting verbal forms)
         NO:  9. (various ‘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 POS's, i.e., they correspond to traditional POS's (**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 solely 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). Again, this is the straightforward consequence of our decision to have a ‘light’ tagset, abstracting away from semantics.

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. tensed (non-past) forms (e.g., idę ‘I am going’,
             pójdę ‘I will go’, będę ‘I will be’)
   NO: Is one of -(e)m, -(e)ś, -śmy, -ście?
       YES: 8.2. agglutinate (‘floating inflection’)
       NO:  8.3. imperative forms
```

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*).

Furthermore, the class 9.3. consists of those wordforms which do not inflect, do not have aspect and are not de-adjectival, i.e.:

- **conjunction**s, which can in turn be divided into

  - coordinating conjunctions, e.g., *i*, *lub*, etc.;
  - subordinating conjunctions (or complementizers), e.g., *że*, *aby*, etc.;
- **preposition**s, e.g., *w*, *na*, etc.;
- **other** forms.

It is possible to distinguish the two kinds of conjunctions and the prepositions syntactically, on the basis of their subcategorization properties, but — since they constitute closed classes — it is easier to define them extensionally, by enumerating them, and to assume that all other non-inflecting, non-aspectual and non-de-adjectival forms fall into the **other** class.

Finally, the class 9.1. may be further partitioned on the basis of purely orthographical (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’, corresponds to the German ‘es wurde getrunken’)
     NO:  Ends in -ąc or -szy?
          YES: 9.1.2. adverbial participles (e.g., czytając
               ‘reading’, przeczytawszy ‘having read’)
          NO:  9.1.3. infinitive form (e.g., iść ‘to
               go’); should end in -c or -ć
```

2.3.3 Summary

On the basis of purely morphosyntactic and orthographic (phonetic) properties, we have identified 17 classes of wordforms in Polish: **noun**, **nominal pronoun**, **gerund**, **numeral**, **adjective**, **adverb**, **conjunction**, **preposition**, **tensed (non-past)** forms, ***-no/-to*** forms, **imperative**, **agglutinate**, **infinitive**, **pseudo-participle**, **adjectival participle**, **adverbial participle**, **other**. Moreover, also on the basis of morphosyntactic and orthographic features, **conjunction**s may be subdivided into coordinating conjunctions and complementizers, **adjectival participle**s can be split into active and passive, and **adverbial participle**s may be partitioned into anterior and contemporary.

Note that this classification corresponds relatively well to the traditional POS's; apart from classes such as **noun**, **adjective**, **adverb**, **preposition** and **conjunction**, various verbal classes can be grouped together into a single **verb** class. On the other hand, we have limited the classes of **pronoun**s and **numeral**s to those forms which can be morphologically distinguished from forms belonging to the other classes. Moreover, depending on particular needs, **gerund**s may be grouped together with **noun**s or together with **verb**s.

It should also be noted that the approach sketched here leaves some room for interpretation, mainly due to the vagueness of the term *inflect*. For example, should wordforms such as *emu*, traditionally regarded to be non-inflecting nouns, be categorized as **other** (because they do not visibly inflect at all, are not de-adverbial, etc.)? No, not necessarily: on the basis of their distributional behavior, they may be claimed to inflect for gender, number and case, although not overtly so. Similarly, indefinite numerals such as *dużo* and *mało*, which do not overtly inflect, may be analyzed as highly syncretic **numeral**s.

3 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.

References

Bański, P. (2000). *Morphological and Prosodic Analysis of Auxiliary Clitics in Polish and English*. Ph. D. dissertation, University of Warsaw.

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

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

Borsley, R. D. (1999). Weak auxiliaries, complex verbs and inflected complementizers. In Borsley and Przepiórkowski (1999), pages 29–59.

Borsley, R. D. and Przepiórkowski, A., editors (1999). *Slavic in Head-Driven Phrase Structure Grammar*. CSLI Publications, Stanford, CA.

Borsley, R. D. and Rivero, M. L. (1994). Clitic auxiliaries and incorporation in Polish. *Natural Language and Linguistic Theory*, **12**, 373–422.

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

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

Przepiórkowski, A. (2000). Long distance genitive of negation in Polish. *Journal of Slavic Linguistics*, **8**, 151–189.

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

Przepiórkowski, A., Kupść, A., Marciniak, M., and Mykowiecka, A. (2001). *Formalny opis języka polskiego: Teoria i implementacja*. Akademicka Oficyna Wydawnicza. In progress.

Puzynina, J. (1969). *Nazwy czynności we współczesnym języku polskim*. Wydawnictwo Naukowe PWN, Warsaw.

Rozwadowska, B. T. (1997). *Towards a Unified Theory of Nominalizations. External and Internal Eventualities*. Wydawnictwo Uniwersytetu Wrocławskiego, Wrocław.

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

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

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

Tajsner, P. (1990). *Case Marking in English and Polish: A Government and Binding Study*. Ph. D. dissertation, Adam Mickiewicz University, Poznań.

Woliński, M. (2001). Rodzajów w polszczyźnie jest osiem. In *Nie bez znaczenia... Prace ofiarowane Profesorowi Zygmuntowi Saloniemu z okazji jubileuszu 15000 dni pracy naukowej*, pages 303–305. Wydawnictwo Uniwersytetu Białostockiego, Białystok.