A comparison of two morphosyntactic tagsets of Polish[⋆]

July 15, 2009

Adam Przepiórkowski

Institute of Computer Science, Polish Academy of Sciences

Abstract. The aim of this paper is to present the main differences between the IPI PAN Tagset, used for the morphosyntactic annotation of the IPI PAN Corpus of Polish, and the NKJP Tagset, employed in the National Corpus of Polish.

1 Introduction

Morphosyntactic tagsets, i.e., formal specifications of morphosyntactic interpretations assigned to words in a given language, are usually developed for the purpose of the morphosyntactic annotation of corpora. While presentations of morphosyntactic systems of various languages found in textbooks and grammars may be sufficient for many linguistic purposes, the task of assigning a morphosyntactic tag (in short: tag) to each word in a large corpus requires a codification of such a system. The resulting tagset must exhaustively specify the repertoire of grammatical classes (parts of speech) assumed for the language, morphosyntactic categories appropriate for particular classes, and possible values of these categories.

A tagset of Polish called the IPI PAN Tagset was proposed in a series of papers (in English: Przepiórkowski and Woliński 2003a,b; in Polish: Woliński 2003 and Przepiórkowski 2003; summarised in the bilingual publication Przepiórkowski 2004a,b) within the IPI PAN Corpus (`http://korpus.pl/`) project.[1] Since then, the tagset has been used in a number of projects, including various projects carried out by the Linguistic Engineering Group at the Institute of Computer Science, Polish Academy of Sciences, as well as, e.g., in the Polish WordNet project (Piasecki et al. 2009; `http://plwordnet.pwr.wroc.pl/`), it inspired the tagset used in the Morfologik dictionary (`http://morfologik.blogspot.com/`), and it influenced the common tagset for a Polish-Ukrainian Parallel Corpus (Kotsyba et al., 2008).[2] A relatively conservative extension of the tagset is proposed in Broda et al. 2008.

At the time of its creation in 2004, the IPI PAN Corpus was the largest corpus of Polish, the only one that was linguistically annotated. However, there were two other independently developed corpora in public existence, namely, the PELCRA Corpus of Polish (`http://korpus.ia.uni.lodz.pl/`) and the PWN Corpus of Polish (`http://korpus.pwn.pl/`), as well as a non-public corpus developed at the Institute of Polish Language, Polish Academy of Sciences, and a small corpus of Polish developed in the 1960s (`http://www.mimuw.edu.pl/polszczyzna/pl196x/`). In 2007, the stakeholders in all large corpus efforts decided to combine their forces and a project was launched with the aim of merging the existing corpora and extending them to a 1-billion word *National Corpus of Polish* (henceforth, NCP or, in Polish, NKJP for *Narodowy Korpus Języka Polskiego*); see `http://nkjp.pl/`.

NCP is being annotated at various linguistic levels, including morphosyntax, named entities, syntax and limited word sense disambiguation. At the morphosyntactic level, NCP adopts the

[⋆] The work described in this article was carried out within the project *Narodowy Korpus Języka Polskiego* funded by the National Ministry of Science and Higher Education (grant number R17 003 03). This publication is supported by the European FP7 project project *MONDILEX*.

[1] IPI PAN is the Polish acronym of *Instytut Podstaw Informatyki Polskiej Akademii Nauk* 'Institute of Computer Science, Polish Academy of Sciences', where the IPI PAN Corpus project was carried out.

[2] Some comparisons of the IPI PAN Tagset to MULTEXT-East (Erjavec, 2001, 2004) tagsets may be found in Dimitrova et al. 2009 and Derzhanski and Kotsyba 2009.

main assumptions of the IPI PAN Tagset, including the morphosyntactic definition of grammatical classes (e.g., a numeral is defined on the basis of its morphosyntactic behaviour, not in the traditional semantic terms) inspired by works of Zygmunt Saloni and his colleagues (see, e.g., Saloni and Świdziński 2001 for a summary) and the detailed flexemic approach to the delimitation of grammatical classes (e.g., infinitive and finite verb are two separate classes, as they have different inflectional characteristics) following work by Janusz S. Bień (1991).

Nevertheless, some modifications of the IPI PAN Tagset were necessary, both for theoretical and for practical reasons. The tagset resulting from these modifications and used in the NCP annotation is called the NKJP Tagset. The aim of this paper is to describe and — where necessary — justify the differences between the IPI PAN Tagset ($\mathrm{T}_{\mathrm{IPI}}$ in brief) and the NKJP Tagset (henceforth, $\mathrm{T}_{\mathrm{NKJP}}$).[3]

2 Differences

Within NCP, a 1-million word corpus is being annotated manually. Manual annotation is one of the most expensive corpus building tasks, and one way to reduce the cost is to annotate the corpus automatically and only correct or disambiguate the automatic annotation manually. For the morphosyntactic annotation, a new version of the morphological analyser Morfeusz (Woliński, 2006) is used in NCP, which is based on the linguistic data described in Saloni et al. 2007. Some of the differences between $\mathrm{T}_{\mathrm{IPI}}$ and $\mathrm{T}_{\mathrm{NKJP}}$ stem from the availability of new linguistic information in this version of Morfeusz.

2.1 New non-inflecting classes

The main criterion for distinguishing grammatical classes in $\mathrm{T}_{\mathrm{IPI}}$ is morphosyntactic, i.e., inflection and agreement. According to this criterion, all non-inflecting (f)lexemes fall into one bag, so an additional — distributional — criterion must be applied to distinguish, e.g., prepositions from conjunctions, and only a few traditional non-inflecting categories are posited in $\mathrm{T}_{\mathrm{IPI}}$. With the benefit of hindsight it seems that these classes are too coarse-grained, so four additional non-inflecting classes are carved out in $\mathrm{T}_{\mathrm{NKJP}}$ from those present in $\mathrm{T}_{\mathrm{IPI}}$.

Interjection In principle any word may be used as an interjection, but for the purpose of $\mathrm{T}_{\mathrm{NKJP}}$ interjection (interj) is understood rather narrowly. A segment (i.e., a word-level token receiving a morphosyntactic interpretation) is marked as an interjection, if one of the following holds:

- it may only be used as an interjection, e.g., segments such as *ach*, *och*, *oj*,
- if the same form has other interpretations, they are not related to the interjection use of that form, e.g., *a* (which may also be a conjunction or an abbreviation),
- it is onomatopoeic, e.g., *mu* or *kukuryku*.

Examples of segments which may be used interjectively but are not marked as interjections are *tak* 'yes' and *kurwa* 'whore'.

Subordinate conjunction Where $\mathrm{T}_{\mathrm{IPI}}$ only recognised conjunctions (Pol. *spójniki*), $\mathrm{T}_{\mathrm{NKJP}}$ differentiates between coordinate conjunctions (Pol. *spójniki równorzędne*; conj), e.g., *i*, *lub* and *oraz*, and subordinate conjunctions (Pol. *spójniki podrzędne*), sometimes called complementisers (comp), e.g., *że*, *aby*, *bowiem*. It is clear that these two non-inflecting classes have very different syntactic behaviour.

[3] A detailed presentation of $\mathrm{T}_{\mathrm{NKJP}}$ may be found in the guidelines for annotators (Przepiórkowski, 2009); a stable version of these guidelines will be made available at http://nkjp.pl/.

Predicative adjective There are three adjectival classes in $\mathrm{T}_{\mathrm{IPI}}$: the usual inflecting adjectives (adj), ad-adjectival adjectives (adja), e.g., *polsko* 'Polish' in *polsko-niemiecki* 'Polish-German', and post-prepositional adjectives (adjp), e.g., *polsku* in *po polsku* 'in Polish'. To these, $\mathrm{T}_{\mathrm{NKJP}}$ adds another non-inflecting adjectival class, namely, the class of one-form lexemes consisting of forms which may only be used in predicative contexts (adjc)[4], e.g., *zdrów* 'healthy' (cf. *On wydaje się zdrów* 'He seems healthy', but not **zdrów człowiek* 'healthy man') or *ciekaw* 'curious' (e.g., *Jestem ciekaw* 'I am curious', but not **ciekaw człowiek* 'curious man').

Bound word The segmentation principles of $\mathrm{T}_{\mathrm{IPI}}$, adopted in $\mathrm{T}_{\mathrm{NKJP}}$, rule that there are no segments containing spaces, so, e.g., *po trochu* 'little by little' cannot be treated as one segment. But the form *trochu* in contemporary Polish is a bound word, occurring in this construction only, so there is no reason to treat it as a noun or an adjective — any decision would have to be arbitrary. In $\mathrm{T}_{\mathrm{NKJP}}$, such indeterminate bound words are marked as burk, with the name of the class inspired by Derwojedowa and Rudolf 2003.

2.2 Abbreviations

Abbreviations play an important role in the task of automatic segmentation of text into sentences: a full stop after an abbreviation may, but need not, also signal the end of a sentence, so each abbreviation should be marked for whether it requires a full stop or not.

Unlike in $\mathrm{T}_{\mathrm{IPI}}$, there is a separate abbreviation class (brev) in $\mathrm{T}_{\mathrm{NKJP}}$. There is a technical category associated with this class, "fullstoppedness", which may take one of two values: pun (the abbreviation segment should be followed by a full stop) and npun (the segment does not have to be followed by a full stop).

The lemma for a segment marked as brev is the full dictionary form of the abbreviation, e.g., for *np* (*na przykład* 'for example'), the tag should be brev:pun (*np* should be followed by a full stop) and its lemma should be NA PRZYKŁAD. For the segment *dr*, on the other hand, the lemma will always be DOKTOR, but the tag should be — in accordance with Polish orthographic rules — either brev:pun (e.g., in masculine accusative) or brev:npun (e.g., in nominative).

2.3 Adverbs and particles

In $\mathrm{T}_{\mathrm{IPI}}$, the class of particle-adverbs (qub), separate from the class of adverbs (adv), is considered an "else" class: if a segment does not fit any other class, it is annotated as qub. With the addition of several non-inflectional classes (see above), the need for such an "else" class diminishes, so in $\mathrm{T}_{\mathrm{NKJP}}$ this class is defined in a constructive way. It may contain various particles (described in more detail in Przepiórkowski 2009), the reflexive marker SIĘ, ad-numeral operators such as OKOŁO 'around' and BLISKO 'almost', and intensifiers such as JEDYNIE 'only' and NAWET 'even'.

On the other hand, the class of adverbs is larger in $\mathrm{T}_{\mathrm{NKJP}}$ than in $\mathrm{T}_{\mathrm{IPI}}$; adverbs in $\mathrm{T}_{\mathrm{NKJP}}$ are implicitly split into two subclasses:

1. de-adjectival or gradable adverbs, e.g., DŁUGO 'long' and BARDZO 'very', which are always specified for degree (positive, pos, in case of de-adjectival adverbs which are not synthetically gradable); this subclass in $\mathrm{T}_{\mathrm{NKJP}}$ corresponds closely to the whole adv class in $\mathrm{T}_{\mathrm{IPI}}$;
2. traditional adverbs which are neither de-adjectival nor gradable, e.g., GDZIE 'where' and WCZORAJ 'yesterday'; they are not marked for degree in $\mathrm{T}_{\mathrm{NKJP}}$; in $\mathrm{T}_{\mathrm{IPI}}$ they belong to the class qub.

2.4 Other differences

Apart from the substantial differences listed above, there is a number of minor differences between the two tagsets, mentioned below.

[4] The mnemotechnics of adjc is 'adjective after the copula', although such forms may occur in various predicative environments, not only copular, also as secondary predicates.

Alien elements There are two technical classes in $\mathrm{T}_{\mathrm{IPI}}$ corresponding to various “alien” elements in texts, mostly foreign language expressions and passages: xxs for those segments which occupy a nominal position and, hence, may be assigned case, number and gender, and xxx for other foreign expressions. In $\mathrm{T}_{\mathrm{NKJP}}$ there is only one “alien” class, xxx, for those segments which do not enter into relations with other (non-alien) segments in the sentence. This class is used mostly for annotating longer foreign expressions or whole passages in a foreign language. Other foreign segments, which enter into relations with other elements of the sentence, i.e., also those occupying nominal positions, should be marked in the usual way, as nouns, adverbs, etc.

Collective numerals Although some of the $\mathrm{T}_{\mathrm{IPI}}$ publications listed above mention the class of collective numerals, numcol, that class was absent from the tagset actually used for the annotation of the IPI PAN Corpus and it is reintroduced in $\mathrm{T}_{\mathrm{NKJP}}$.

Comparative degree Since comp is used in $\mathrm{T}_{\mathrm{NKJP}}$ as the name for the class of complementisers, the comparative degree is marked as com in this tagset, in contradistinction to comp used for that purpose in $\mathrm{T}_{\mathrm{IPI}}$.

3 Conclusion

Since $\mathrm{T}_{\mathrm{IPI}}$ is relatively widely used, the modifications in $\mathrm{T}_{\mathrm{NKJP}}$ were kept to the minimum and consist mainly in adding a few classes for non-inflecting elements and the removal of a hardly ever used class xxs. Both tagsets are well-documented, so we hope that an adaptation of existing tools to the new $\mathrm{T}_{\mathrm{NKJP}}$ will turn out to be a manageable task. To further facilitate that task, the appendix below contains a specification of $\mathrm{T}_{\mathrm{NKJP}}$.

Appendix: NKJP Tagset

In the following specification of $\mathrm{T}_{\mathrm{NKJP}}$, section [ATTR] lists all morphosyntactic categories and their possible values, while section [POS] specifies grammatical classes and categories appropriate for these classes. For example, any noun must be marked as subst:*number*:*case*:*gender*, where, e.g., *number* must be replaced by one of the possible values of this category, i.e., sg or pl. Hence, a full tag for the form *lampę* ‘lamp’ should be subst:sg:acc:f.

Some categories are optional for some classes, e.g., only some prepositions (such as w) have a vocalic (*we*) and a non-vocalic (*w*) form, so the segment *we* could be marked as prep:acc:wok, while the tag of, e.g., *na* could be prep:acc.

At the end of the specification some constraints are listed which should be respected by any tools used for the processing of this tagset.

All grammatical classes and categories not mentioned above are described in $\mathrm{T}_{\mathrm{IPI}}$ publications listed in section 1.

```
## NKJP Tagset (version 1.0 of 23 June 2009)

[ATTR]

number                  = sg pl
case                    = nom gen dat acc inst loc voc
gender                  = m1 m2 m3 f n
person                  = pri sec ter
degree                  = pos com sup
aspect                  = imperf perf
negation                = aff neg
accommodability         = congr rec
```

```
accentability           = akc nakc
post-prepositionality   = npraep praep
agglutination           = agl nagl
vocalicity              = nwok wok

fullstoppedness         = pun npun


[POS]

adja    =
adjp    =
adjc    =
conj    =
comp    =
interp  =
pred    =
xxx     =
adv     = [degree]
imps    = aspect
inf     = aspect
pant    = aspect
pcon    = aspect
qub     = [vocalicity]
prep    = case [vocalicity]
siebie  = case
subst   = number case gender
depr    = number case gender
ger     = number case gender aspect negation
ppron12 = number case gender person [accentability]
ppron3  = number case gender person [accentability] [post-prepositionality]
num     = number case gender accommodability
numcol  = number case gender accommodability
adj     = number case gender degree
pact    = number case gender aspect negation
ppas    = number case gender aspect negation
winien  = number gender aspect
praet   = number gender aspect [agglutination]
bedzie  = number person aspect
fin     = number person aspect
impt    = number person aspect
aglt    = number person aspect vocalicity

brev    = fullstoppedness
burk    =
interj  =

## This class should not appear in the results of manual annotation:

ign     =


## Non-defeasible constraints:
##
```

```
## siebie --> base = siebie
## siebie --> case IN gen dat acc inst loc
## pant --> aspect = perf
## pcon --> aspect = imperf
## pact --> aspect = imperf
## ger --> gender = n
## depr --> number = pl
## depr --> gender = m2
## depr --> case IN nom voc acc
## numcol --> gender IN n m1
## aglt --> aspect = imperf
## bedzie --> aspect = imperf
## impt --> number:person IN sg:sec pl:pri pl:sec
## prep --> case IN nom gen dat acc inst loc


## Defeasible constraints:
##
## ger --> number = sg
## num --> number = pl
```

References

Bień, J. S. (1991). *Koncepcja słownikowej informacji morfologicznej i jej komputerowej weryfikacji*, volume 383 of *Rozprawy Uniwersytetu Warszawskiego*. Wydawnictwa Uniwersytetu Warszawskiego, Warsaw. `http://bc.klf.uw.edu.pl/12/`.

Broda, B., Piasecki, M., and Radziszewski, A. (2008). Towards a set of general purpose morphosyntactic tools for Polish. In Kłopotek et al. (2008), pages 445–454.

Derwojedowa, M. and Rudolf, M. (2003). Czy Burkina to dziewczyna i co o tym sądzą ich królewskie mości, czyli o jednostkach leksykalnych pewnego typu. *Poradnik Językowy*, 5:39–49.

Derzhanski, I. and Kotsyba, N. (2009). Towards a consistent morphological tagset for Slavic languages: Extending MULTEXT-East for Polish, Ukrainian and Belarusian. In Garabík, R., editor, *Metalanguage and Encoding Scheme Design for Digital Lexicography: Proceedings of MONDILEX Third Open Workshop*, pages 9–26, Bratislava.

Dimitrova, L., Koseska-Toszewa, V., Derzhanski, I., and Roszko, R. (2009). Annotation of parallel corpora (on the example of the Bulgarian-Polish parallel corpus). In Shyrokov, V. and Dimitrova, L., editors, *Organization and Development of Digital Lexical Resources: Proceedings of MONDILEX Second Open Workshop*, pages 47–54, Kiev. National Academy of Sciences of Ukraine, Ukrainian Lingua-Information Fund.

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

Erjavec, T. (2004). MULTEXT-East version 3: Multilingual morphosyntactic specifications, lexicons and corpora. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004*, pages 1535–1538, Lisbon. ELRA.

Kotsyba, N., Shypnivska, O., and Turska, M. (2008). Principles of organising a common morphological tagset for PolUKR (Polish-Ukrainian Parallel Corpus). In Kłopotek et al. (2008), pages 475–484.

Kłopotek, M. A., Przepiórkowski, A., Wierzchoń, S. T., and Trojanowski, K., editors (2008). *Intelligent Information Systems.* Akademicka Oficyna Wydawnicza EXIT, Warsaw.

Piasecki, M., Szpakowicz, S., and Broda, B. (2009). *A Wordnet from the Ground Up.* Oficyna Wydawnicza Politechniki Wrocławskiej. To appear.

Przepiórkowski, A. (2003). Składniowe uwarunkowania znakowania morfosyntaktycznego w korpusie IPI PAN. *Polonica*, XXII–XXIII:57–76.

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

Przepiórkowski, A. (2004b). *Korpus IPI PAN. Wersja wstępna.* Institute of Computer Science, Polish Academy of Sciences, Warsaw.

Przepiórkowski, A. (2009). Zasady znakowania morfosyntaktycznego w NKJP. Version 1.03 of 29 June 2009.

Przepiórkowski, A. and Woliński, M. (2003a). A flexemic tagset for Polish. In *Proceedings of Morphological Processing of Slavic Languages, EACL 2003*, pages 33–40, Budapest.

Przepiórkowski, A. and Woliński, M. (2003b). 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*, pages 109–116.

Saloni, Z., Gruszczyński, W., Woliński, M., and Wołosz, R. (2007). *Słownik gramatyczny języka polskiego.* Wiedza Powszechna, Warsaw.

Saloni, Z. and Świdziński, M. (2001). *Składnia współczesnego języka polskiego.* Wydawnictwo Naukowe PWN, Warsaw, 5th edition.

Woliński, M. (2003). System znaczników morfosyntaktycznych w korpusie IPI PAN. *Polonica*, XXII–XXIII:39–55.

Woliński, M. (2006). Morfeusz — a practical tool for the morphological analysis of Polish. In Kłopotek, M. A., Wierzchoń, S. T., and Trojanowski, K., editors, *Intelligent Information Processing and Web Mining*, Advances in Soft Computing, pages 511–520. Springer-Verlag, Berlin.