The Design of Syntactic Annotation Levels in the National Corpus of Polish

Katarzyna Głowińska, Adam Przepiórkowski

Institute of Computer Science, Polish Academy of Sciences
ul. Ordona 21, 01-237 Warsaw, Poland
k.glowinska@gmail.com, adamp@ipipan.waw.pl

Abstract

This paper presents the procedure of the syntactic annotation of the National Corpus of Polish. Syntactic annotation consists here of shallow parsing and manual post-editing of the results by annotators. The description concentrates on the delimitation of syntactic words and groups, as well as on problems encountered during the annotation process.

1. Introduction

National Corpus of Polish (Pol. *Narodowy Korpus Języka Polskiego*; NKJP; `http://nkjp.pl/`; Przepiórkowski *et al.* 2008) is a project carried out in 2008–2010, aiming at the creation of a 1-billion-word automatically annotated corpus of Polish, with a 1-million-word subcorpus annotated manually. The following levels of linguistic annotation are distinguished in the project: 1) segmentation into sentences, 2) segmentation into fine-grained word-level tokens, 3) morphosyntactic analysis, 4) coarse-grained syntactic words (e.g., analytical forms, constructions involving bound words, etc.), 5) syntactic groups, 6) named entities,[1] 7) word senses (for a limited number of ambiguous lexemes).

2. Aim

The aim of this paper is to present the design of the two strictly syntactic annotation levels 4) and 5), since — as described in detail in Sections 3. and 4. — they differ in interesting respects from the usual approach to syntactic annotation. The ensuing Section 5. presents the envisaged annotation procedure, Section 6. describes several problems encountered so far, and Section 7. concludes the paper.[2]

3. Syntactic words

Word-level segmentation (or tokenisation) in NKJP follows the approach of the previous large corpus of Polish, the IPI PAN Corpus (`http://korpus.pl/`; Przepiórkowski 2004), in assuming very fine-grained segmentation adhering to two segmentation principles: segments must be contiguous and they cannot overlap. For example, the analytical future tense form *będę szedł* 'will walk' is split into two segments in *Będę szybko szedł*, lit. 'I-will quickly walk', to satisfy the contiguity principle (note the intervening *szybko* 'quickly'). Similarly, in *Będę szedł i śpiewał*, lit. 'I-will walk and sing', there are arguably two analytical forms, *Będę szedł* and *Będę śpiewał*, which share the form *Będę*; also for this reason, the future auxiliary *Będę* must be treated as a separate segment. Since such auxiliaries must be treated as separate segments in some cases, they are assumed to always be separate segments.

To give another, perhaps more interesting, example: BAĆ SIĘ 'fear' and ZAŚMIAĆ SIĘ 'laugh out', are two so-called inherently reflexive verbs — there are no lexemes BAĆ or ZAŚMIAĆ, without the reflexive marker (RM) SIĘ. However, in *Bał się zaśmiać* 'He feared to laugh', just one realisation of the RM is the unmarked case. Again, if *Bał się* and *się zaśmiać* were treated as segments, they would overlap, contrary to one of the segmentation principles.

A tool used in the National Corpus of Polish, the morphological analyser Morfeusz (Woliński, 2006), tokenises texts according to the above principles and assigns morphosyntactic interpretations to such segments, adhering to the NKJP Tagset (Przepiórkowski, 2009a). Nevertheless, for further syntactic processing it is useful to distinguish a level of representation consisting of traditional word forms, including analytical tense and mood forms, reflexive verbs, discontinuous conjunctions, etc. This is the level of syntactic words.

In most cases syntactic words are co-extensive with word-level segments (henceforth, simply *segments*) and may bear the same morphosyntactic interpretation. However, there are also systematic differences between the segment-level NKJP Tagset and the tagset for syntactic words (henceforth, NKJP_{SW} Tagset). For example, at the segment level morphosyntactic interpretations do not contain information about tense, as the future tense of, e.g., *będę szedł* is the property of the whole syntactic word, rather than the segment *szedł*, which by itself may actually express the past tense. Similarly, at the segment level there was no need for the category of reflexivity and for a subdivision of conjunctions into different syntactic types, including discontinuous conjunctions.

Another important difference between the two tagsets consists in their granularity. Where the segment-level tagset distinguishes multiple verbal grammatical classes (roughly, parts of speech), the NKJP_{SW} Tagset is closer to the traditional parts of speech, thanks to assuming the traditional grammatical categories of tense and mode, absent (for

[1] Syntactic annotation is performed at the same time as annotation of named entities. This latter task is described in Savary *et al.* 2010.

[2] XML encoding of these syntactic levels is presented in detail in Przepiórkowski 2009b; see also Przepiórkowski and Bański 2009.

good reasons) in the segment-level NKJP Tagset. Despite this bow towards the tradition, both tagsets define grammatical classes and categories according to morphosyntactic and syntactic criteria only. This should be contrasted with tagsets directly reflecting the Latin tradition of defining parts of speech on the basis of mixed morphosyntactic, syntactic, semantic and pragmatic criteria, as in, e.g., MULTEXT-EAST (Erjavec 2004; see Przepiórkowski and Woliński 2003 for discussion).

Table 1 presents the complete NKJP_{SW} Tagset, given as a conservative modification of the segment-level NKJP Tagset. An excerpt from the NKJP Tagset is presented in Table 2, with the corresponding classes of the NKJP_{SW} Tagset in Table 1 boldfaced. Note that the four classes: `praet` (past participle), `bedzie` (future auxiliary or future of BYĆ 'be'), `fin` (finite form) and `impt` (imperative form) are replaced with one class named `Verbfin` and that the category `reflexivity` is added to all verbal classes, including active and passive participles.[3]

Adjc	=	
Conj	=	[cont]
Comp	=	
Interj	=	
Interp	=	
Qub[4]	=	[vocalicity]
Adv	=	degree
Imps	=	aspect reflexivity negation
Inf	=	aspect reflexivity negation
Pant	=	aspect reflexivity negation
Pcon	=	aspect reflexivity negation
Prep	=	case [vocalicity]
Siebie	=	case
Noun	=	number case gender [aspect] [reflexivity] [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 reflexivity negation
Ppas	=	number case gender aspect reflexivity negation
Verbfin	=	number person tense mood aspect reflexivity negation [gender]
Winien	=	number person gender tense mood aspect negation
Pred	=	tense mood aspect negation
Brev	=	fullstoppedness brev_pos

Table 1: Complete NKJP_{SW} Tagset specification

[3] In general, in order to differentiate morphosyntactic interpretations of syntactic words from that of segments, capitalised tags for grammatical classes are used in the NKJP_{SW} Tagset.

[4] `Qub` (*kublik* in Polish) is the tag for particle-adverb.

pact	=	number case gender aspect negation
ppas	=	number case gender aspect negation
praet	=	number gender aspect [agglutination]
bedzie	=	number person aspect
fin	=	number person aspect
impt	=	number person aspect

Table 2: A fragment of the NKJP Tagset specification

4. Syntactic groups

As is well known, the borderline between syntactic words or, more generally, multi-word expressions, on one hand, and syntactic groups,[5] on the other, is fuzzy. Various idiomatic expressions could equally well be treated as syntactic words or as syntactic groups. The general principle adopted here is that constructions which are defined with a reference to a specific orthographic or base form are treated as words, and more general constructions — as groups. For example, all the adverbs that match the pattern: `Prep` *po* + `adjp`[6] ended with *-sku* (e.g., *po babsku* 'like a woman', *po chamsku* 'like a lout', *po cudzoziemsku* 'like a foreigner') are syntactic words, as orthographic forms must be used to create the grammar rule.

Shallow (partial) approach to syntactic analysis is assumed here (Abney, 1991). For example, a nominal phrase that consists of a noun and a prepositional phrase, e.g., *mieszkanie z balkonem* 'a flat with a balcony', is always treated as two syntactic groups (*mieszkanie* and *z balkonem*), without an attempt to solve PP-attachment ambiguities.[7] On the other hand, note that there are compund prepositions in Polish (so called "secondary prepositions") that may consist of two prepositions and an intervening noun, e.g., *w przeciwieństwie do*, 'in contrast with'. They are treated as one syntactic word marked as `Prep`. So the phrase *w przeciwieństwie do brata* 'unlike his brother' is one PrepNG group, and not two PrepNG groups. An exception is also made for elective constructions, e.g., *jeden z najlepszych* 'one of the best', which are treated as one syntactic group.

Moreover, as usual in the shallow parsing paradigm, no use of a valence dictionary is made here, so there is no attempt either to identify complete verb phrases or to show dependency structure (as it is done in the Prague Dependency Treebank for Czech; `http://ufal.mff.cuni.cz/pdt2.0/`). Syntactic annotation in the National Corpus of Polish is limited to joining words together into constituents.

The following syntactic groups are distinguished in NKJP:

[5] In this paper, the terms *(syntactic) group* and *syntactic phrase* are treated as synonymous.

[6] `Adjp` is the tag from NKJP tagset that stands for post-prepositional adjective.

[7] There is a separate project carried out at the Institute of Computer Science, Polish Academy of Sciences, aiming at the creation of a full-fledged treebank of Polish, based on the material of NKJP.

- nominal group (NG): *pilot samobójca* ‘kamikaze pilot’, *król Francji* ‘the king of France’, *rząd i parlament* ‘government and parliament’, *czerwona sukienka* ‘red dress’, *nic ważnego* ‘nothing important’,
- numeral group (NumG): *pięć samochodów* ‘five cars’, *trzech spośród pisarzy* ‘three of the writers’,
- adjectival group (AdjG): *wyjątkowo piękna* ‘exceptionally beautiful’, *[jest] gotowy wyjechać* ‘[he is] ready to leave’,
- prepositional-nominal group (PrepNG): *nad głównym wejściem* ‘above the main entrance’,
- prepositional-adjectival group (PrepAdjG): *[wyglądasz] na zmęczonego* ‘[you look] tired’,
- prepositional-numeral group (PrepNumG): *[pracował] za dwóch* ‘[he did enough work] for two’, *[równanie] z dwoma niewiadomymi* ‘[an equation] with two unknowns’,
- adverbial group (AdvG): *gdzieś daleko* ‘somewhere far away’,
- discourse group (DisG): *no cóż* ‘oh well’, *moim zdaniem* ‘in my opinion’,
- subordinate clause (CG) (with subordinate conjunction): *[wiedział], że to już koniec* ‘[he knew] it was the end’,
- interrogative clause (KG): *[spytałem ojca], czy mogę iść do kina* ‘[I asked my father] whether I could go to the cinema’.

Figure 1 shows three levels of annotation: segments (tokens), syntactic words and syntactic groups. For each phrase syntactic and semantic heads are marked. In Figure 1, the syntactic head of each constituent is marked in green and the semantic head is marked with a triangle.

5. Annotation procedure

In case of morphosyntactic annotation, NKJP fully follows the best methodological practices (Przepiórkowski and Murzynowski, 2009): manual annotation is performed by two independent annotators and if they do not agree, a referee makes the final decision and perhaps modifies the guidelines. We claim that shallow syntactic annotation is a much simpler task than detailed morphosyntactic annotation, so a more automatic procedure should suffice to achieve high quality annotation. Syntactic annotation consist of shallow parsing and manual post-editing of the results by annotators.

The manually constructed grammar, both for syntactic words and for syntactic groups, is encoded in the shallow parsing system Spejd (`http://nlp.ipipan.waw.pl/Spejd/`; Buczyński and Przepiórkowski 2008), already successfully used for similar tasks (Buczyński and Wawer, 2008; Przepiórkowski, 2009c). Spejd rules form a cascade, with the output of one rule constituting the input of the next rule. An example of a particularly simple word-level rule identifying multi-segment adverbs such as *po ciemku* (‘in the dark’) and *po kryjomu* (‘in secret’), marking them as `Adv` and assigning them base forms such as PO CIEMKU and PO KRYJOMU, is given below:

```
Rule    "idiomatic expressions: po + ..."
Match:  [orth~"[Pp]o"]
        [orth~"ciemku|kryjomu|trochu"];
Eval:   word(Adv, "po " 2.orth);
```

Another example of a rule, clustering words together into a nominal group (NG) and marking the second element (either a `Noun` or the syntactic head of another nominal group) of the sequence as both syntactic and semantic head of the group, is presented below:

```
Rule    "NG: Adj + Noun"
Match:  [pos~"Adj|Pact|Ppas"]
        ([pos~"Noun"] | [type="NG"]);
Eval:   unify(case number gender,1,2);
        group(NG,2,2);
```

The iterative process of grammar development and manual post-editing is implemented: the initial grammar was applied to a sample of the 1-million word corpus and the results were subject to manual correction.[8] These corrections gave rise to the next version of the grammar, applied to the next corpus sample, etc. The evaluation of the grammar, as well as an estimation of the inter-annotator agreement, will be performed on the basis of the last sample of this subcorpus. The final grammar, attained at the end of the process of the manual correction of the 1-million word subcorpus, will be applied to the whole 1-billion word NKJP.

As usual in shallow parsing, and in order to maintain a high level of consistency, neither the shallow grammar nor the post-editors resolve PP-attachment ambiguities or similar ambiguities involving potentially post-modifying adjectival participles (cf. Section 6.5.). On the other hand, discontinuous phrases or syntactic words, not discovered automatically by the grammar, have to be manually adduced (cf. Section 6.4.).

6. Annotation related problems

The most important problems encountered so far were: group boundaries, multiword entities, abbreviations, discontinuous phrases and syntactic words, and active and passive participles modifying nouns.

6.1. Group boundaries

Normally, a syntactic group is the *longest* possible sequence of syntactic words that satisfies certain conditions, i.e., match a Spejd rule or a description in the annotation guidelines. However, it may happen that such a match actually contains two syntactic groups. In the sentence: *Będą*

[8] Manual post-editing is done via the TrEd editor (`http://ufal.mff.cuni.cz/~pajas/tred/`), adjusted to the needs of the National Corpus of Polish.

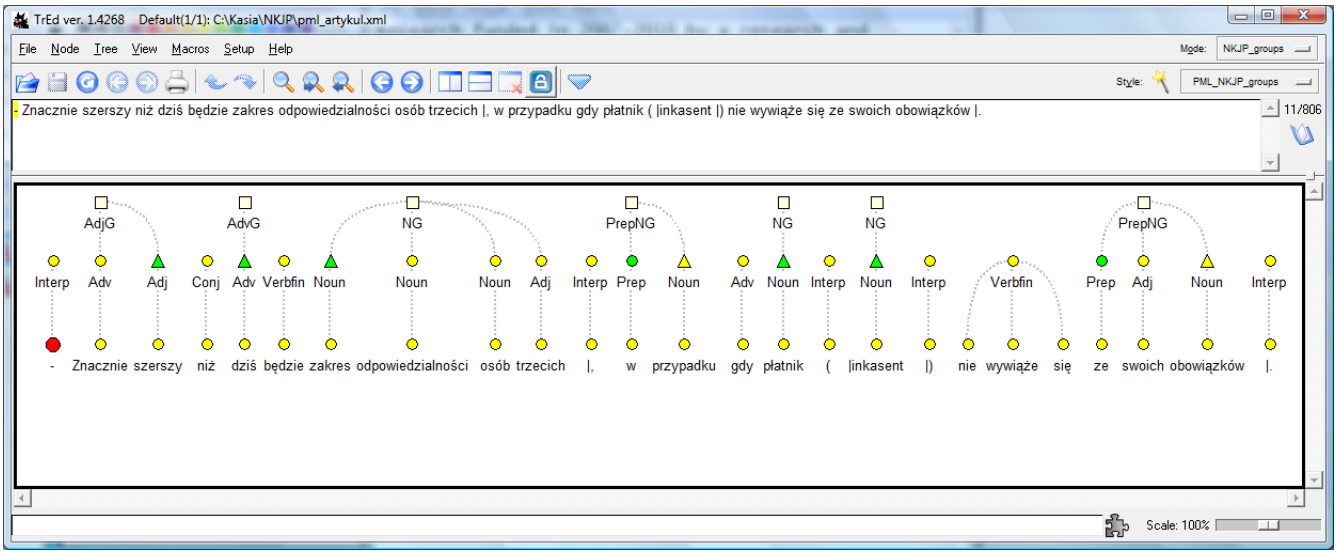

Figure 1: Example of syntactic annotation with the use of the TrEd editor

mogli dochodzić w postępowaniu cywilnym zapłaty podatku., lit. 'They could demand in civil proceedings the tax payment.', the parser identifies one prepositional-nominal group because a set of conditions specified by a Spejd rule are met. In particular, the word *zapłaty* 'payment' (in the genitive) could form a nominal group with the word *postępowanie* 'proceedings' (Noun+Noun_{gen} follows the pattern of expressions *król Francji* 'the king of France'). In fact, there are two syntactic groups: prepositional-nominal group (*w postępowaniu cywilnym*) and nominal group (*zapłaty podatku*) that fulfil two syntactic functions in the sentence: an adverbial of manner and a complement. This kind of problem is subject to manual correction.

6.2. Multiword entities

In the first step the list of about 1000 entities that can be categorized into multiword entities was created. Then grammatical classes were assigned to each entity, e.g.: adverbials (*po ciemku* 'in the dark', *na czczo* 'on an empty stomach'), particle-adverbs (*na pewno* 'for sure'), compound prepositions (*co do* 'as for'), compound conjunctions (*dlatego że* 'because'), discontinuous conjunctions (*nie tylko ... lecz także* 'not only ... but also'). Apart from that, there are expressions from foreign languages: for example, *au courant* and *curriculum vitae* are marked as Adv and Noun, respectively. In the last step, the rules for some entities were created to disambiguate the meaning of an entity in a given context. To give an example: in *Zgłupiał **do reszty***. 'He's **completely** out of his mind.' — *do reszty* is the multiword entity (Adv), while in *Stół nie pasuje **do reszty** mebli.* 'The table doesn't match **the rest** of the furniture.' — *do reszty* is the syntactic group (PrepNG).

6.3. Abbreviations

All abbreviations in the National Corpus of Polish are marked as Brev and their full forms are given as their base forms, but there is no morphological information (such as case or gender for nouns).

If an abbreviation stands for one word, (e.g., *r.* is the abbreviation of *rok* 'year', which appears in dates), it can be treated as the corresponding full form. For example, the phrase *w 1981 r.* 'in 1981' could be recognized as a prepositional-nominal group, where *r.* is regarded as a noun.

The situation is much more complicated when the abbreviation stands for two or more words, as in *pt.* = *pod tytułem* 'entitled', *kk* = *kodeks karny* 'the penal code', *itp.* = *i tym podobne* 'and the like'. There are three possible solutions to this problem:

- treat the abbreviation as the full form, e.g., *w br.* = *w bieżącym roku* 'in the current year' could be PrepNG (in this case *br.* should be marked as semantic head of the group, while in the full form only the noun *rok* 'year' would be marked),
- include the abbreviation in another syntactic group, e.g., *pt.* (and *kk*) usually follows the noun and could be attached to it (*wiersz pt. "Miłość"* 'a poem entitled 'Love' could be recognized as the nominal group),
- treat the abbreviation as being outside syntactic classification, e.g., *itp.*, when it does not belong to any syntactic group.

The approach adopted here is close to the first solution — abbreviations are treated as corresponding full forms, but they should still be marked as abbreviations. To this end, the brev_pos category appropriate to abbreviations was added, with values corresponding to grammatical classes of syntactic words (NOUN, ADJ, etc.; written in capitals for technical reasons) and to types of syntactic groups (NG, PrepNG, etc.). Some examples of syntactic tags for abbreviations mentioned above are given in Table 3.

6.4. Discontinuous phrases

A discountinuous phrase consists of at least two words separated by another word that does not belong to this phrase. A set of rules for such cases could be created but in some contexts manual corrections are necessary,

r.	=	Brev:pun:NOUN
br.	=	Brev:pun:NG
pt.	=	Brev:pun:PrepNG
kk	=	Brev:npun:NG

Table 3: Examples of tags for abbreviations

e.g., *Życzył „szczęścia" **zbiegłemu** poprzedniego wieczora z więzienia Maze **terroryście*** 'He wished luck to the **terrorist** who **escaped** from the Maze prison last night.'. As the word *zbiegłemu* is an adjective and clearly modifies (and has the same value of case, gender and number as) the noun *terroryście*, they should be joined together into a nominal group.

6.5. Active and passive participles

Active and passive adjectival participles, regarded as verb forms in NKJP, can modify the nouns in some contexts, e.g., *dymiące zgliszcza* 'smoking ruins', usually if they precede the nouns and have the same value of case, gender and number as the nouns. However, if an adjectival participle follows the noun, it is sometimes difficult to automatically resolve its attachment point and the right boundary of the group headed by the participle, e.g., *meldunki napływające z całego kraju* 'reports coming from the whole country'.
In the preliminary version of the grammar, only `Pact` and `Ppas` forms that precede `Noun` are included within the nominal group.

7. Conclusion

The most advanced linguistic annotation present in a Polish corpus is the low-level morphosyntactic annotation, available in the IPI PAN Corpus at `http://korpus.pl/` (and in the NKJP demo at `http://nkjp.pl/`). Within the National Corpus of Polish, syntactic annotation is applied in a conservative, step-wise manner, on top of morphosyntactic annotation. At the level of syntactic words the original NKJP Tagset is modified to allow for broader grammatical classes and more traditional grammatical categories, such as tense and mood. At the syntactic group level, only relatively small groups that can be identified with very high accuracy are marked, so that the shallow grammar resulting from the manual correction process can be reliably applied to the whole 1-billion word corpus. A full treebank annotation of the 1-million word subcorpus is carried out in a related project, again with the aim of developing a full-fledged deep grammar applicable to the whole NKJP. By the break of 2010/2011, these activities should converge in the existence of the first corpus of Polish containing multiple levels of linguistic annotation.

Acknowledgements Research funded in 2007–2010 by a research and development grant from the Polish Ministry of Science and Higher Education.

References

Abney, S. (1991). Parsing by chunks. In R. Berwick, S. Abney, and C. Tenny, editors, *Principle-Based Parsing*, pp. 257–278. Kluwer.

Buczyński, A. and Przepiórkowski, A. (2008). ♠ Demo: An Open Source Tool for Shallow Parsing and Morphosyntactic Disambiguation. In LREC (2008).

Buczyński, A. and Wawer, A. (2008). Shallow parsing in sentiment analysis of product reviews. In S. Kübler, J. Piskorski, and A. Przepiórkowski, editors, *Proceedings of the LREC 2008 Workshop on Partial Parsing: Between Chunking and Deep Parsing*, pp. 14–18, Marrakech. ELRA.

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*, pp. 1535–1538, Lisbon. ELRA.

LREC (2008). *Proceedings of the Sixth International Conference on Language Resources and Evaluation, LREC 2008*, Marrakech. ELRA.

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

Przepiórkowski, A. (2009a). A comparison of two morphosyntactic tagsets of Polish. In V. Koseska-Toszewa, L. Dimitrova, and R. Roszko, editors, *Representing Semantics in Digital Lexicography: Proceedings of MONDILEX Fourth Open Workshop*, pp. 138–144, Warsaw.

Przepiórkowski, A. (2009b). TEI P5 as an XML standard for treebank encoding. In M. Passarotti, A. Przepiórkowski, S. Raynaud, and F. Van Eynde, editors, *Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories (TLT 8)*, pp. 149–160, Milan, Italy. Forthcoming.

Przepiórkowski, A. (2009c). Towards the automatic acquisition of a valence dictionary for Polish. In M. Marciniak and A. Mykowiecka, editors, *Aspects of Natural Language Processing*, volume 5070 of *Lecture Notes in Computer Science*, pp. 191–210. Springer-Verlag, Berlin.

Przepiórkowski, A. and Bański, P. (2009). Which XML standards for multilevel corpus annotation? In Z. Vetulani, editor, *Proceedings of the 4th Language & Technology Conference*, pp. 245–250, Poznań, Poland.

Przepiórkowski, A. and Murzynowski, G. (2009). Manual annotation of the National Corpus of Polish with Anotatornia. In S. Goźdź-Roszkowski, editor, *The proceedings of Practical Applications in Language and Computers PALC 2009*, Frankfurt am Main. Peter Lang. Forthcoming.

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

Przepiórkowski, A., Górski, R. L., Lewandowska-Tomaszczyk, B., and Łaziński, M. (2008). Towards the National Corpus of Polish. In LREC (2008).

Savary, A., Waszczuk, J., and Przepiórkowski, A. (2010).

Towards the Annotation of Named Entities in the Polish National Corpus. To appear in *Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC 2010*, Malta. ELRA.

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