National Corpus of Polish

Adam Przepiórkowski[1,5], **Mirosław Bańko**[3,5], **Rafał L. Górski**[2], **Barbara Lewandowska-Tomaszczyk**[4], **Marek Łaziński**[3,5], **Piotr Pęzik**[4]

[1]Institute of Computer Science, Polish Academy of Sciences
[2]Institute of Polish Language, Polish Academy of Sciences
[3]Polish Scientific Publishers PWN [4]University of Łódź [5]University of Warsaw

Abstract

The paper presents the main results of the *National Corpus of Polish* project, which took place from December 2007 to June 2011, including: the sizes of the main corpus and various subcorpora; their makeups; linguistic and XML annotation; corpus annotation tools; corpus search engines; and applications. This is not a scientific article *per se*, but rather a starting point for anybody interested in the main achievements of this multifaceted project; the paper may also serve as a possible general reference to the National Corpus of Polish.

Keywords: corpora, Polish, representativeness, linguistic annotation, annotation tools, search engines, applications

1. Introduction

The aim of this paper is to present the main results of the *National Corpus of Polish* (Pol. *Narodowy Korpus Języka Polskiego*; NKJP; `http://nkjp.pl/`) project carried out between December 2007 and June 2011.[1] Four institutions took part in the project: two institutes of the Polish Academy of Sciences (PAS; Institute of Computer Science in Warsaw – the coordinator, and Institute of Polish Language in Cracow), the University of Łódź and the Polish Scientific Publishers PWN. More information on the origin of the project may be found in Przepiórkowski *et al.* 2008, 2010.

2. Size and composition

The whole NKJP corpus consists of about 1.8 billion (i.e., 1 800 000 000) segments. A segment, as used in the National Corpus of Polish, is a technical linguistic term: words are usually single segments, but in some special cases they are split into smaller segments, as in the three-segment word *biało*|-|*czerwony* (‘white-and-red) or *przyszli*|*by*|*śmy* ‘we would (have) come’. Also commas, exclamation marks, etc., are treated as separate segments. In general, the ratio of segments to traditional orthographic words is about 1.2, so, e.g., a 1.8-billion-segment corpus contains about 1.5 billion traditional words.

Arguably, there does not exist a sufficient number of Polish texts of various genres to compose a balanced corpus of this size. In NKJP, a subcorpus of 300 million segments was created and balanced roughly on the basis of readership statistics in Poland. In particular, this subcorpus was prepared with the following percentages of particular text genres in mind:

- 50%: journalism, including:
  - dailies (51% of journalism),
  - magazines (47%),
  - journalistic books (2%);
- 16%: fiction literature (prose, poetry, drama),
- 5.5%: non-fiction literature,
- 5.5%: instructive writing and textbooks,
- 2%: academic writing and textbooks,
- 4%: miscellaneous written (legal, advertisements, user manuals, letters) and unclassified written;
- 7%: born in the Internet (forums, chatrooms, mailing lists, etc.);
- 10%: spoken, including:
  - conversational,
  - spoken from the media,
  - quasi-spoken (incl. parliamentary transcripts).

In practice, due to the relatively small number of instructive writing and textbooks in NKJP, it turned out to be difficult to create even a 300-million-segment subcorpus balanced strictly in accordance with the above specification. On the other hand, the general balance between journalism, fiction, non-fiction, other written texts, Internet and spoken texts reflects the desiderata above.

NKJP also makes available two smaller subcorpora which follow these specifications precisely: a 100-million-word (i.e., roughly, 120-million-segment) balanced subcorpus and a 1-million-word *manually annotated* balanced subcorpus. In fact, while the whole NKJP and its larger subcorpora are – because of the restrictions imposed by the copyright law – only available for search (cf. § 5), the 1-million-word subcorpus, consisting of small samples from many texts, is available in the source form, from `http://clip.ipipan.waw.pl/`.

3. Annotation

All texts in the National Corpus of Polish are – mostly manually – annotated with metadata, i.e., contain information about their origin, the title, the author(s), etc. Moreover, all texts are – mostly automatically or semi-automatically – marked structurally, i.e., where applicable, divided into front matter, body and back matter, with the body split into chapters, sections, etc., down to the level of paragraphs.

Additionally, texts are annotated linguistically at the following levels:

- segmentation into sentences and word-level tokens (the latter called *segments* in NKJP; see above);
- morphosyntactic;
- syntactic;

[1] Originally, the project had been planned for 3 years, but was extended for another half a year.

- named entity recognition (NER);
- word sense disambiguation (WSD).

Technically, all kinds of annotation are encoded in XML, using mark-up schemata based on Text Encoding Initiative guidelines (version P5; `http://www.tei-c.org/`; Burnard and Bauman 2008). Within NKJP, a subset of TEI was carefully selected that can be used for the markup of corpora with multiple level linguistic annotation; see, e.g., Przepiórkowski and Bański 2009 and `http://nlp.ipipan.waw.pl/TEI4NKJP/`.

Obviously, manual linguistic annotation is labour-intensive, so only the 1-million-word subcorpus was annotated by linguists. At each level, the annotation followed the best practice in the field: each fragment was annotated independently by two people who were not allowed to communicate directly, and any differences in annotation were resolved by an experienced referee. Specifically, two tools were used for manual annotation, with slightly differing procedures:

- TrEd (`http://ufal.mff.cuni.cz/~pajas/tred/`; Pajas and Štěpánek 2008) was used for NER and for the syntactic annotation (cf. Waszczuk *et al.* 2010);
- Anotatornia (`http://nlp.ipipan.waw.pl/Anotatornia/`; Hajnicz *et al.* 2008), a tool developed extensively within NKJP, was used for all other levels (cf. Przepiórkowski and Murzynowski 2011).

In case of Anotatornia, the procedure was rather sophisticated. Firstly, text fragments were non-randomly distributed between different annotators in such a way that the time spent by any two annotators working together was minimised; the objective of such a distribution of texts was to lower the risk of co-learning of mistakes by two constantly cooperating annotators. Secondly, in cases the two annotators differed, they were shown *where* they differed (but not *how* they differed), in order to allow them to correct any obvious mistakes. Only in case of discrepancies left after this additional step did the referee step in.

It is also worth noting that syntactic annotation was performed on the results of morphosyntactic annotation, and – in the process – some errors were discovered which were left after this already very careful annotation procedure. Such error reports where again sent to the morphosyntactic referee, for final acceptance. As a result, we expect the quality of the morphosyntactic level to be extremely high.

The following subsections briefly characterise each linguistic annotation level.

3.1. Segmentation

Word-level segmentation has already been mentioned in § 2. In rare cases of segmentation ambiguities, as in *gdzieś*, which may be analysed as a single segment meaning 'somewhere' or two segments (*gdzie*|*ś*) meaning 'where-you', both variants are represented in the annotation, but only one of them is marked as correct in a given context.

Also sentence-level segmentation is relatively unproblematic, although some problems might be caused by sentences citing other sentences, e.g., equivalents of *"Come here!" he said.*, etc. In such cases, since sentences cannot be nested in the annotation scheme adopted in NKJP (but see § 3.3), only the maximal sentences are marked as such, so, e.g., *"Come here!"* in this example is not marked as a sentence.

3.2. Morphosyntax

The morphosyntactic tagset assumed in NKJP is based on that of the IPI PAN Corpus (Przepiórkowski, 2004); it assumes a flexemic view of parts of speech and a very detailed treatment of morphosyntactic categories (cf. Przepiórkowski and Woliński 2003), including such ephemeral categories as postprepositionality (for 3rd person pronouns) and accommodability (for numerals), apart from the more traditional categories of case, number, gender, etc. The differences between the two tagsets, summarised in Przepiórkowski 2009, stem mainly from the differences between the two version of the morphological analyser Morfeusz (Woliński, 2006) used in these corpora; the current Morfeusz SGJP is based on the data of the *Grammatical Dictionary of Polish* (Saloni *et al.*, 2007). Just as in the IPI PAN Corpus, any morphosyntactically ambiguous segments are annotated with all their interpretations, exactly one of which is selected as correct in a given context.

3.3. Syntax

Two syntactic sublevels are distinguished: syntactic words and syntactic groups (Głowińska and Przepiórkowski, 2010).

Syntactic words group the fine-grained segments into more traditional words, e.g., the three segments *przyszli*|*by*|*śmy* 'we would (have) come' are marked as a single word. Unlike segments, syntactic words may be non-contiguous and may even overlap. For example, *będę szedł i śpiewał* 'I-will walk and sing' is analysed as two syntactic words sharing the segment *będę*: *będę szedł* 'I-will walk' and *będę śpiewał* 'I-will sing'. Note that the latter word is non-contiguous: the two segments are separated by *szedł i*. Each syntactic word is annotated morphosyntactically, although the tagset for words differs a little from that for segments; for example, the categories of tense and mood are present only in the former, while the morphosyntactic class of ad-adjectival adjectives (e.g., *biało* in *biało*|-|*czerwony*; cf. § 2) is only present in the latter.

Syntactic groups are also understood roughly in a traditional way, as nominal groups, prepositional groups, clauses, etc., but only shallow syntactic annotation is present in NKJP. This means that no attempt is made to resolve potential structural ambiguities (e.g., cases of PP-attachment ambiguities). Nesting is avoided, i.e., usually only maximal nominal groups, prepositional groups, etc., in each clause are marked, although clauses themselves can be nested. An unusual feature of this annotation level is the distinction between syntactic heads and semantic heads (cf. Przepiórkowski 2006, 2008).

Technically, the semi-manual syntactic annotation was performed with a Spejd (Buczyński and Przepiórkowski, 2009) grammar and the subsequent manual correction of its results with the TrEd tool mentioned above.

3.4. NER

The repertoire of named entities annotated in NKJP is based on the proposal in TEI P5. The following types and subtypes of NEs are distinguished (Savary *et al.*, 2010):

- person name:
  - forename (e.g., *Lech*),
  - surname (e.g., *Wałęsa*),
  - additional name element (e.g., *Lwie Serce* 'Lionheart');
- organisation name (e.g., *Uniwersytet im. Adama Mickiewicza* 'Adam Mickiewicz University' – in Poznań);
- geographical name (e.g., *Warta* – the river running through Poznań);
- place name (referring to a geopolitical entity):
  - district (e.g., *Jeżyce*, a district of Poznań),
  - settlement (e.g., *Poznań*),
  - region (e.g., *Wielkopolska* the 'Greater Poland' voivodship),
  - country (e.g., *Polska* 'Poland'),
  - bloc (e.g., *Unia Europejska* 'European Union');
- time (e.g., *12.25*);
- date (e.g., *grudzień 1981* 'December 1981').

In the 1-million-word corpus, each named entity is annotated with its type (and subtype, if applicable), lemma, and – optionally – other information. A special feature of NE annotation in NKJP is that named entities may be nested, as in *Uniwersytet im. Adama Mickiewicza*, which should be annotated as an organisation, but *Adama Mickiewicza* within it should be marked as a person name, with *Adama* additionally marked as a forename, and *Mickiewicza* – as a surname.

Technically, the semi-manual annotation of the 1-million-word corpus consisted in the automatic parsing of texts with a named entity grammar written in SProUT (Piskorski, 2005) and the manual correction of the results with the use of TrEd.

3.5. WSD

WSD is the most experimental and preliminary level of annotation in NKJP: a little over 100 frequent and clearly ambiguous (homonymous rather than simply polysemous) lexemes were selected and all occurrences of their forms in the 1-million-word corpus were annotated with their appropriate meanings. In line with the recommendations in various papers in Agirre and Edmonds 2006, only rough senses were distinguished. To this end, all fine-grained senses in a traditional – albeit corpus-based – dictionary (*Inny słownik języka polskiego*; Bańko 2000) were grouped into coarse-grained senses used by the annotators.

4. Tools

Various tools were developed within NKJP for the automatic annotation of the whole corpus.

4.1. Segmentation and morphosyntax: PANTERA

The creation of a new tagger within NKJP was necessitated by the fact that the only publicly available tagger for Polish, TaKIPI (Piasecki, 2007), used for the annotation of the IPI PAN Corpus, is to some extent based on manually created rules and cannot be easily extended to new tagsets.

A new morphosyntactic tagger developed in NKJP, PANTERA (`http://code.google.com/p/pantera-tagger/`; Acedański 2010), is a transformation-based tagger[2] (Brill, 1993), but it extends the transformation-based methodology to the case of complex morphological tagsets. The tagger makes use of the Morfeusz SGJP morphological analyser mentioned above, and it also incorporates the guesser module of TaKIPI (Broda *et al.*, 2008), as well as the sentence segmentation rules of Miłkowski and Lipski 2009.

Because of the differences in tagsets, it is not trivial to compare the quality of TaKIPI and PANTERA, but – following the methodology of Karwańska and Przepiórkowski 2011 – PANTERA seems to fare a little better than TaKIPI and gives the accuracy of about 93%, as evaluated on the 1-million-word manually annotated subcorpus.

The morphosyntactic annotation of PANTERA is available via the Poliqarp search engine (cf. § 5).

4.2. WSD Development Environment

The automatic Word Sense Disambiguation was performed within NKJP with the help of WSDDE (`http://nlp.ipipan.waw.pl/WSDDE/`; Młodzki and Przepiórkowski 2009), a platform for experimenting with various WSD algorithms and parameters. 215 experiments were performed for each of the 106 ambiguous lexemes under consideration, differing in the kinds of features used, the feature selection algorithm and the machine learning approach. For the implementation of various feature selection and machine learning algorithms, WSDDE relies on the WEKA library (Witten and Frank, 2005).

The best machine learning algorithms turned out to be: Naïve Bayes (the algorithm of choice for 51 lexemes), Bayes Net (32 lexemes) and the K* instance-based classifier (10 lexemes). As expected, InfoGain turned out to be the best feature selection method, often supported by a method called CfsSubsetEval.

The best features differed widely among the 106 lexemes, but typically the most important features were:

- the presence or absence of a given lexeme in the context of the disambiguated word (e.g., the presence of ZGŁASZAĆ 'put forward' in the context of UWAGA 'remark, attention');
- the grammatical number of the disambiguated word;
- less obviously, the gender and the number of the following word,
- the kind of preposition, if any, following the disambiguated word, the case of the following word, etc., i.e., features hinting at the valence frame of the word in question; etc.

The average disambiguation accuracy, measured accordingly to the standard 10-fold cross-validation methodology, is 91.48%,[3] i.e., significantly higher than the in-

[2] Hence the Polish name "Polskiej Akademii Nauk Tager Ekstrahujacy Reguły Automatycznie" and the acronym PANTERA.

[3] It is virtually the same, 91.46%, for the leave-one-out cross-validation.

formed baseline consisting in always selecting the most frequent sense (78.3%) and directly comparable to the best results for English.

4.3. Tools for NER and syntactic annotation

As mentioned above, manually constructed grammars were used for the initial syntactic and NER annotation of the 1-million-word corpus (Waszczuk *et al.*, 2010). A modified version of the Spejd grammar was also used for the annotation of the complete NKJP.

On the other hand, for the NER annotation of the complete NKJP, a separate tool was trained on the manually annotated corpus. A method known as Joined Label Tagging (Alex *et al.*, 2007) was applied, a variant of the well-known IOB 'chunking-as-tagging' approach, and the HMM-like linear-chain Conditional Random Field algorithm was used as the classifier.

While the evaluation methodology is described in detail in Savary and Piskorski 2011, let us note that the results of the automatic classifier are significantly better than those of the manually constructed grammar: the average precision over various types of named entities is 0.83 (it is lowest for organisation names, 0.70, and highest for person names and temporal expressions, 0.86, as well as for derivations from NEs, 0.87; this should be compared to around 0.72 for the grammar), while the average recall is 0.76 (from 0.65 for organisations to 0.80 for person names and temporal expressions; compare to around 0.36 for the grammar).

5. Search engines

Two search engines are currently employed at `http://nkjp.pl/`: Poliqarp and the PELCRA engine.

Poliqarp (`http://poliqarp.sourceforge.net/`; Przepiórkowski *et al.* 2004) was created for the IPI PAN Corpus and substantially further developed since then, also within NKJP. Poliqarp has a very rich query syntax[4], loosely based on that of the IMS Corpus Workbench, which makes it possible to query for sequences of orthographic forms, lemmata and morphosyntactic information (provided by PANTERA), using the full power of regular expressions. The price for this expressive power is the need to learn the specific syntax of the query language.

While Poliqarp uses its own internal representation for indexing the corpus, the PELCRA search engine (`http://nkjp.uni.lodz.pl/`; cf., e.g., Pęzik 2011) is based on the combination of Apache Lucene (`http://lucene.apache.org/`) and standard relational database technologies, which results in its high efficiency. The search interface does not make the morphosyntactic annotation available, but – by making use of Morfologik (`http://morfologik.blogspot.com/`) under the bonnet – it allows for lemma-based search. A strong feature of the PELCRA search engine is its general ease of use and the highly developed collocation component.

Unfortunately, at the moment neither engine is able to access the levels of linguistic annotation higher than morphosyntax; constructing such an engine is left for future work.

6. Applications and Conclusion

The first partial results of the project were made available a few months after its inception and were intensively used by the team at the Institute of Polish Language PAS working on a new large dictionary of Polish (Żmigrodzki *et al.*, 2007); in fact, NKJP is the main empirical base of this dictionary. Another NKJP-based dictionary – of eponymous phrases – is due to be published by the Warsaw University Press. An important subproject of NKJP was *Words of the Day*, presenting and commenting every day those words whose frequency of occurrence in the Polish dailies was significantly higher than could be expected from their frequencies in a longer reference period.

Apart from lexicographic applications, the results of the project are already extensively employed – and, in some cases, further developed – in Natural Language Processing projects carried out at the Institute of Computer Science PAS, University of Łódź, Wrocław University of Technology and – doubtlessly – other places. Although the project ended only recently, resources created within NKJP have also already been used in language teaching, theoretical linguistics, psycholinguistics and translation studies.

We hope that the results of the National Corpus of Polish will continue to be useful in the years to come.

References

Acedański, S. (2010). A morphosyntactic Brill tagger for inflectional languages. In H. Loftsson, E. Rögnvaldsson, and S. Helgadóttir, editors, *Advances in Natural Language Processing: Proceedings of the 7th International Conference on Natural Language Processing, IceTAL 2010, Reykjavík, Iceland*, volume 6233 of *Lecture Notes in Artificial Intelligence*, pages 3–14, Heidelberg. Springer-Verlag.

Agirre, E. and Edmonds, P., editors (2006). *Word Sense Disambiguation: Algorithms and Applications*, volume 33 of *Text, Speech and Language Technology*. Springer, Dordrecht.

Alex, B., Haddow, B., and Grover, C. (2007). Recognising nested named entities in biomedical text. In *Proceedings of the BioNLP Workshop at ACL 2007*, pages 65–72, Prague.

Bańko, M., editor (2000). *Inny słownik języka polskiego*. Wydawnictwo Naukowe PWN, Warsaw.

Brill, E. (1993). *A Corpus-Based Approach to Language Learning*. Ph. D. dissertation, University of Pennsylvania.

Broda, B., Piasecki, M., and Radziszewski, A. (2008). Towards a set of general purpose morphosyntactic tools for Polish. In M. A. Kłopotek, A. Przepiórkowski, S. T. Wierzchoń, and K. Trojanowski, editors, *Intelligent Information Systems*, pages 441–450. Akademicka Oficyna Wydawnicza EXIT, Warsaw.

Buczyński, A. and Przepiórkowski, A. (2009). Spejd: A shallow processing and morphological disambiguation tool. In Z. Vetulani and H. Uszkoreit, editors, *Human Language Technology: Challenges of the Information Society*, volume 5603 of *Lecture Notes in Artificial Intelligence*, pages 131–141. Springer-Verlag, Berlin.

Burnard, L. and Bauman, S., editors (2008). *TEI*

[4] See `http://nkjp.pl/poliqarp/help/en.html`.

P5: Guidelines for Electronic Text Encoding and Interchange. Oxford. http://www.tei-c.org/Guidelines/P5/.

Goźdź-Roszkowski, S., editor (2011). *Explorations across Languages and Corpora: PALC 2009*, Frankfurt am Main. Peter Lang.

Głowińska, K. and Przepiórkowski, A. (2010). The design of syntactic annotation levels in the National Corpus of Polish. In LREC (2010).

Hajnicz, E., Murzynowski, G., and Woliński, M. (2008). ANOTATORNIA – lingwistyczna baza danych. In *Materiały V konferencji naukowej InfoBazy 2008, Systemy * Aplikacje * Usługi*, pages 168–173, Gdańsk. Centrum Informatyczne TASK, Politechnika Gdańska.

Karwańska, D. and Przepiórkowski, A. (2011). On the evaluation of two Polish taggers. In Goźdź-Roszkowski (2011), pages 105–113.

LREC (2010). *Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC 2010*, Valletta, Malta. ELRA.

Miłkowski, M. and Lipski, J. (2009). Using SRX standard for sentence segmentation in LanguageTool. In Vetulani (2009), pages 556–560.

Młodzki, R. and Przepiórkowski, A. (2009). The WSD development environment. In Vetulani (2009), pages 185–189.

Pajas, P. and Štěpánek, J. (2008). Recent advances in a feature-rich framework for treebank annotation. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*, pages 673–680, Manchester.

Piasecki, M. (2007). Polish tagger TaKIPI: Rule based construction and optimisation. *Task Quarterly*, **11**(1–2), 151–167.

Piskorski, J. (2005). Named-entity recognition for Polish with SProUT. In L. Bolc, Z. Michalewicz, and T. Nishida, editors, *Intelligent Media Technology for Communicative Intelligence, Second International Workshop, IMTCI 2004, Warsaw, Poland, September 13-14, 2004, Revised Selected Papers*, volume 3490 of *Lecture Notes in Computer Science*. Springer-Verlag.

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

Przepiórkowski, A. (2006). On heads and coordination in a partial treebank. In J. Hajič and J. Nivre, editors, *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT 2006)*, pages 163–174, Prague.

Przepiórkowski, A. (2008). *Powierzchniowe przetwarzanie języka polskiego*. Akademicka Oficyna Wydawnicza EXIT, Warsaw.

Przepiórkowski, A. (2009). 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*, pages 138–144, Warsaw.

Przepiórkowski, A. and Bański, P. (2009). Which XML standards for multilevel corpus annotation? In Vetulani (2009), pages 245–250.

Przepiórkowski, A. and Murzynowski, G. (2011). Manual annotation of the National Corpus of Polish with Anotatornia. In Goźdź-Roszkowski (2011), pages 95–103.

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

Przepiórkowski, A., Krynicki, Z., Dębowski, Ł., Woliński, M., Janus, D., and Bański, P. (2004). A search tool for corpora with positional tagsets and ambiguities. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004*, pages 1235–1238, Lisbon. ELRA.

Przepiórkowski, A., Górski, R. L., Lewandowska-Tomaszczyk, B., and Łaziński, M. (2008). Towards the National Corpus of Polish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation, LREC 2008*, Marrakech. ELRA.

Przepiórkowski, A., Górski, R. L., Łaziński, M., and Pęzik, P. (2010). Recent developments in the National Corpus of Polish. In LREC (2010).

Przepiórkowski, A., Bańko, M., Górski, R. L., and Lewandowska-Tomaszczyk, B., editors (2011). *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN, Warsaw. Forthcoming.

Pęzik, P. (2011). Providing corpus feedback for translators with the PELCRA search engine for NKJP. In Goźdź-Roszkowski (2011), pages 135–144.

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

Savary, A. and Piskorski, J. (2011). Language resources for named entity annotation in the National Corpus of Polish. To appear in Control and Cybernetics.

Savary, A., Waszczuk, J., and Przepiórkowski, A. (2010). Towards the annotation of named entities in the National Corpus of Polish. In LREC (2010).

Vetulani, Z., editor (2009). *Proceedings of the 4th Language & Technology Conference*, Poznań, Poland.

Waszczuk, J., Głowińska, K., Savary, A., and Przepiórkowski, A. (2010). Tools and methodologies for annotating syntax and named entities in the National Corpus of Polish. In *Proceedings of the International Multiconference on Computer Science and Information Technology (IMCSIT 2010): Computational Linguistics – Applications (CLA'10)*, pages 531–539, Wisła, Poland. PTI.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2nd edition. http://www.cs.waikato.ac.nz/ml/weka/.

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, pages 503–512. Springer-Verlag, Berlin.

Żmigrodzki, P., Bańko, M., Dunaj, B., and Przybylska, R. (2007). Koncepcja *Wielkiego słownika języka polskiego* — przybliżenie drugie. In P. Żmigrodzki and R. Przybylska, editors, *Nowe studia leksykograficzne*, pages 9–21. Lexis, Cracow.