Recent Developments in the National Corpus of Polish

Adam Przepiórkowski[1,5], **Rafał L. Górski**[2], **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 aim of the paper is to present recent — as of March 2010 — developments in the construction of the National Corpus of Polish (NKJP). The NKJP project was launched at the very end of 2007 and it is aimed at compiling a large, linguistically annotated corpus of contemporary Polish by the end of 2010. Out of the total pool of 1 billion words of text data collected in the project, a 300 million word balanced corpus will be selected to match a set of predefined representativeness criteria. This present paper outlines a number of recent developments in the NKJP project, including: 1) the design of text encoding XML schemata for various levels of linguistic information, 2) a new tool for manual annotation at various levels, 3) numerous improvements in search tools. As the work on NKJP progresses, it becomes clear that this project serves as an important testbed for linguistic annotation and interoperability standards. We believe that our recent experiences will prove relevant to future large-scale language resource compilation efforts.

1. Introduction

For Polish, the most represented Slavic language of the EU, there still does not exist a national corpus, i.e., a large, balanced and publicly available corpus, which would be at least morphosyntactically annotated. Currently, there exist three contemporary[1] Polish corpora which are — to various extents — publicly available. The largest and the only one that is fully morphosyntactically annotated is the IPI PAN Corpus (`http://korpus.pl/`; Przepiórkowski 2004), containing over 250 million segments (over 200 million orthographic words), but — as a whole — it is rather badly balanced.[2] Another corpus, which is considered to be carefully balanced, the PWN Corpus of Polish (`http://korpus.pwn.pl/`), contains over 100 million words, of which only 7,5 million sample is freely available for search. The third corpus, the PELCRA Corpus of Polish (`http://korpus.ia.uni.lodz.pl/`), also contains about 100 million words, all of which are publicly searchable.

The aim of the paper is to present recent — as of March 2010 — developments in the construction of the *National Corpus of Polish* (Pol. *Narodowy Korpus Języka Polskiego*; NKJP; `http://nkjp.pl/`). The project, funded by the Polish Ministry of Science and Higher Education (project number: R17 003 03), was launched at the very end of 2007 and will run until the end of 2010.

Background description of the project may be found in Przepiórkowski *et al.* 2008 (in LREC 2008).

[1]Another — much smaller and dated, but historically very important — corpus is available in its entirety: `http://www.mimuw.edu.pl/polszczyzna/pl196x/`, a 0.5-million word corpus created in the 1960s as the empirical basis of a frequency dictionary (Kurcz *et al.* 1990).

[2]There exists a 30-million segment subcorpus of the IPI PAN Corpus which is relatively balanced.

2. Recent developments

2.1. Acquisition of spoken data

Within the NKJP text type taxonomy, we made a clear distinction between casual spoken discourse and other registers of spoken language, such as scripted speech, or even relatively unrestricted registers of speech used in some radio and television shows. Not only does casual spoken discourse exhibit a number of unique lexical and syntactic characteristics, but it also it seems to have a very distinctive topical and pragmatic discourse structure, which often reflects the implicitness in the treatment of certain topics by inner circles of speakers, usually friends or family members, sharing a common background of experiences and beliefs.

A considerable amount of effort has therefore been invested in the acquisition of casual spoken data. Due to the technical and legal aspects, the acquisition of such data has posed a particular challenge since the very beginning of the project. A number of acquisition agents were equipped with high quality voice recorders and trained in the process of acquiring and transcribing samples of spoken data into a temporary format (which was later converted into TEI P5 compliant XML). For legal reasons, the persons were asked to focus on collecting samples of conversations with their friends and family members, who were first made aware of, and agreed to the possibility of being taped for the purposes of the project. A formal legal agreement was then obtained to include transcripts of the recordings in the NKJP. For privacy reasons, some names and personal details mentioned in the recordings had to be edited. The transcripts have not so far been aligned with the recordings, as it would have greatly limited the representation of this register of Polish in the corpus, but the

availability of digitized recordings leaves open such a possibility.
To date, some 900 000 words of casual spoken discourse have been acquired and transcribed. Given the 600 000 words of similar data contributed to the project from the PELCRA corpus, we expect to collect nearly 2 million words of informal conversational Polish in the NKJP project. At the same time, we have been collecting and transcribing samples of spoken media discourse, in order to meet the 10% spoken data representation criterion in the 300 million word balanced subcorpus to be compiled by the end of the project.

2.2. Text encoding

The need for text encoding standards for language resources (LRs) is widely acknowledged: within the International Standards Organization (ISO) Technical Committee 37 / Subcommittee 4 (TC 37 / SC 4; `http://www.tc37sc4.org/`) work in this area has been going on since early 2000s, and working groups devoted to this issue have been set up in two current pan-European projects, CLARIN (`http://www.clarin.eu/`) and FLaReNet (`http://www.flarenet.eu/`). It is obvious that standards are necessary for the interoperability of tools and for the facilitation of data exchange between projects, but they are needed also within projects, especially where multiple partners and multiple levels of linguistic data are involved.
NKJP is committed to following current standards and best practices, but it turns out that the choice of text encoding for multiple layers of linguistic annotation is far from clear. Przepiórkowski and Bański 2009a contains an overview of recent and current standards and best practices, at various stages of development. The conclusion of this paper is that the guidelines of the Text Encoding Initiative (Burnard and Bauman 2008; `http://www.tei-c.org/`) should be followed, as it is a mature and carefully maintained *de facto* standard with a rich user base. Various proposed ISO TC 37 / SC 4 standards, including Morphosyntactic Annotation Framework (MAF), Syntactic Annotation Framework (SynAF) and Linguistic Annotation Framework (LAF), are still under development and, especially SynAF and LAF, have the form of general models rather than specific off-the-shelf solutions.
Nevertheless, when selecting from the rich toolbox provided by TEI, an attempt has been made to follow recommendations of these proposed ISO standards, as well as other common XML formats, including TIGER-XML (Mengel and Lezius 2000) and PAULA (Dipper 2005). This work, described in more detail in Przepiórkowski and Bański 2009b,a, as well as in Przepiórkowski 2009b, resulted in TEI P5 XML schemata encoding data models largely isomorphic with or at least mappable to those formats.

2.3. Annotation tools

A 1-million word subcorpus of NKJP is being annotated manually and it will be used for training automatic annotation tools. Anotatornia, a tool for the manual annotation of word senses developed within a previous project at ICS PAS (Hajnicz *et al.* 2008), has been extended and extensively modified to allow for the manual addition of sentence boundaries, word-level segmentation, morphosyntactic annotation and word sense disambiguation (Przepiórkowski and Murzynowski 2009).
At each level, annotation is adduced by two linguists, connecting with the server via a web interface. In case of differences, both are notified that the other annotator has made a different decision at a given place, but they are not informed about each other's decision. Each annotator may change their own annotation, or confirm it. If the discrepancy persists, a referee makes the final decision and suggests a modification of the annotation guidelines (Przepiórkowski 2009c), if necessary.
Currently annotation is performed at the levels of sentence segmentation, word-level segmentation, morphosyntax, and word sense disambiguation. Work on the annotation of syntactic words, syntactic groups and named entities was started in November 2009; it is described in detail in Głowińska and Przepiórkowski 2010 (syntactic words, syntactic groups) and in Savary *et al.* 2010 (named entities). The results of the first experiments in word sense disambiguation within the current project are presented in Młodzki and Przepiórkowski 2009.
For the manual morphosyntactic annotation, each segment is automatically marked with all interpretations known to the new version of Morfeusz (Woliński 2006), a morphosyntactic dictionary of Polish based on the data of the *Słownik gramatyczny języka polskiego* ('Grammatical Dictionary of Polish'; Saloni *et al.* 2007). The task of the annotator is to select the right interpretation or add the correct interpretation, if it is not among those proposed by Morfeusz. In the process, various deficiencies of Morfeusz and the underlying grammatical data base have been identified and corrected.
The morphosyntactic tagset used in NKJP is a modified version of the IPI PAN Tagset (Przepiórkowski and Woliński 2003a,b). The differences between the two tagsets are described — and a formal specification of the NKJP Tagset presented — in Przepiórkowski 2009a.

2.4. Search tools and NKJP demo

Since developing an efficient search tool able to manage a 1-billion corpus is a potentially high-risk task, two approaches are pursued in parallel.
The first approach, developed by the PELCRA team at the University of Łódź, is based on the combination of Apache Lucene (`http://lucene.apache.org/`) and relational database technologies.

This approach has scaled well with the size of the corpus and it currently supports a corpus containing one billion words without a noticeable decrease in performance. Apart from scalability, this search engine also focuses on providing convenient access to concordance search results in a variety of output formats, including downloadable Microsoft Excel XML spreadsheets (also supported by the Open Office suite) and integrated browser plugins. Moreover, the results of all search operations can be saved and shared in the form of compressed URLs. Having typed in a complex query, the user can click the URL button to generate a compressed reference to the result set. As an example, the query for the different inflections of the noun prawda, with additional sorting and grouping options is compressed to the following form:

`http://nkjp.uni.lodz.pl/?q=yjsz8hc`

Such a compressed link can be easily shared or saved for future reference. Typing in the compressed link in a web browser's address bar brings up precisely the original result screen together with the completed search form. The PELCRA search engine for the NKJP can also be used to generate time series graphs illustrating the relative popularity of a word or phrase in the different time periods represented in the corpus. Another type of visualization is the register distribution plot — a dynamically generated bar chart showing the frequency of occurrences matching the query in the different registers represented in the corpus. A separate module of this search engine facilitates the extraction of two, and multi-word collocations based on the node query provided. The general accessibility of the PELCRA search tool is further enhanced by the availability of a programmatic interface for automated remote access to the corpus. It has yet to be seen to what extent this approach will accommodate more complex types of linguistic search at various levels of annotation.

The second approach is based on Poliqarp (Janus and Przepiórkowski 2007a,b), a dedicated search engine developed at ICS PAS and currently serving a corpus of 250 million segments: while Poliqarp involves a very expressive query language, currently further expanded to accommodate syntactic queries, it is not clear how well it scales with the size of the corpus. So far, modifications of Poliqarp within NKJP consisted in developing a new corpus compiler, translating the TEI-based XML encoding of texts to an efficient binary format. End-user improvements include more specific error messages in case of malformed queries, and an option to randomise search results.

Both search engines are successfully employed in the NKJP Demo (`http://nkjp.pl/` → EN → DEMO), which currently contains about 1 billion (10^9) words.

2.5. Words of the Day

The 'Words of the Day' (Polish: *Słowa dnia*) is a subproject of the National Corpus of Polish which aims to monitor temporary popularity of words used in major Polish newspapers — to find words that quickly become popular and then fall into disuse. The program constantly monitors a list of RSS news feeds of several newspapers. Each time a new article appears on one of those feeds, it is downloaded, extracted via HTML scraping, and analyzed morphologically to produce a word frequency map.

Every midnight, the frequencies of words that appeared in previous day's articles are compared with frequencies of words in a reference period (articles from the previous month). Several statistical tests (chi-square and log-likelihood) are used to assess significance of a word in a day relative to the reference period. This collection of words is then published on the Web.

It is also possible to fine-tune the publication by hand, using a simple Web interface. The editor is free to correct the program's decisions, as well as add comments on individual words and categorize them. In addition, it is possible to create publications that cover many days: an arbitrary test period and reference period can be chosen.

A similar program monitors every week the frequencies of words published in the Polish local weeklies — members of the Association of the Local Press (*Stowarzyszenie Gazet Lokalnych*). The weekly frequencies are published on the web site of the association: `www.gazetylokalne.pl`.

3. Conclusion

As any "recent developments" overview publication, this paper describes work in progress. Intensive work within the National Corpus of Polish project concerns all levels of corpus development: from data acquisition, through text encoding and linguistic annotation, to efficient corpus search engines. As the work on NKJP progresses, it becomes clear that this project serves as an important testbed for linguistic annotation and interoperability standards. We believe that our recent experiences will prove relevant to future large-scale language resource compilation efforts.

4. Acknowledgements

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

5. References

Burnard, L. and Bauman, S., editors (2008). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Oxford. `http://www.tei-c.org/Guidelines/P5/`.

Dipper, S. (2005). Stand-off representation and exploitation of multi-level linguistic annotation. In *Proceedings of Berliner XML Tage 2005 (BXML 2005)*, pages 39–50, Berlin.

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

Goźdź-Roszkowski, S., editor (2009). *The proceedings of Practical Applications in Language and Computers PALC 2009*, Frankfurt am Main. Peter Lang. Forthcoming.

Górski, R. L. (2009). The representativeness of NKJP. Talk delivered at *Practical Applications in Language and Computers (PALC 2009)*, Łódź, April 2009.

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.

Janus, D. and Przepiórkowski, A. (2007a). Poliqarp 1.0: Some technical aspects of a linguistic search engine for large corpora. In J. Waliński, K. Kredens, and S. Goźdź-Roszkowski, editors, *The proceedings of Practical Applications in Language and Computers PALC 2005*, Frankfurt am Main. Peter Lang.

Janus, D. and Przepiórkowski, A. (2007b). Poliqarp: An open source corpus indexer and search engine with syntactic extensions. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, pages 85–88, Prague.

Kurcz, I., Lewicki, A., Sambor, J., Szafran, K., and Woronczak, J. (1990). *Słownik frekwencyjny polszczyzny współczesnej*. Wydawnictwo Instytutu Języka Polskiego PAN, Cracow.

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

Mengel, A. and Lezius, W. (2000). An XML-based encoding format for syntactically annotated corpora. In *Proceedings of the Second International Conference on Language Resources and Evaluation, LREC 2000*, pages 121–126, Athens. ELRA.

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

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*, pages 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)*, pages 149–160, Milan, Italy. Forthcoming.

Przepiórkowski, A. (2009c). Zasady znakowania morfosyntaktycznego w NKJP. Unpublished manuscript, ICS PAS. Version 1.23 of 27 September 2009.

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

Przepiórkowski, A. and Bański, P. (2009b). XML text interchange format in the National Corpus of Polish. In Goźdź-Roszkowski (2009). Forthcoming.

Przepiórkowski, A. and Murzynowski, G. (2009). Manual annotation of the National Corpus of Polish with Anotatornia. In Goźdź-Roszkowski (2009). Forthcoming.

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.

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.

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

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.

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 511–520. Springer-Verlag, Berlin.