Recent Developments in the National Corpus of Polish

Adam Przepiórkowski[1,5], Rafał L. Górski[2], Marek Łaziński[3,5], and 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 July 2009 — developments in the construction of the National Corpus of Polish. The main developments are: 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.

1 Introduction

The aim of the paper is to present recent — as of July 2009 — 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.

The rest of this section presents the background of the project, the consortium, and envisaged applications of the results of the project in the humanities. The following section, §2, describes the expected results of the project in general terms.[6] Section 3 presents recent developments concerning corpus annotation and search tools, as well as text encoding standards deployed in the project. Section 4 concludes the paper.

1.1 Background

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[7] Polish corpora which are — to various extents —

[6] These two sections overlap to a large extent with Przepiórkowski *et al.* 2008.

[7] 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 1960. as the empirical basis of a frequency dictionary (Kurcz *et al.* 1990).

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.[8] 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.

1.2 Consortium

The project is unique in that it involves all major corpus developers for a given language, including the three developers of the three corpora mentioned above: the Institute of Computer Science at the Polish Academy of Sciences (ICS PAS; the Polish acronym is IPI PAN, hence the name of the corpus) in Warsaw, which coordinates the project, the PWN Publisher in Warsaw and the PELCRA group at the University of Łódź. The fourth partner is the Institute of Polish Language at the Polish Academy of Sciences (IPL PAS) in Cracow, the developer of an internal corpus, available only for research carried out at the Institute.

Another institution whose staff and students are involved in the project is the University of Warsaw. Some of the ideas which influenced the methodology of the project evolved in the Faculty of Polish Studies and in the Institute of Informatics, where two of the authors lecture on corpus linguistics, linguistic engineering and Polish grammar.

1.3 Practical applications in the humanities

The project is correlated with another national project, led by IPL PAS, aiming at the development of a new large dictionary of Polish, for which the National Corpus of Polish will serve as the empirical basis.

External users, such as lexicographers and linguists, will mainly be interested in searching the corpus for word and phrase concordances as well as collocations. The corpus can also serve as the treasure of well-known quotations from Polish and key words of Polish culture, with some emphasis on good representation of secondary school required readings in Polish literature and history. Therefore, quotations from the corpus will be crucial for new large dictionaries of Polish (including the new dictionary currently developed at IPL PAS, as well as dictionaries published by PWN), not only as a source of the typical uses of words, but also as a reference to cultural authorities rooted very well in the Polish literary tradition.

The corpus as a whole also enables creating a number of comparable corpora. The size of the corpus, exceeding the informal standard of 100 million, shall

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

guarantee that there will be sufficient number of texts of different genres to meet the selection criteria of target corpora. Technically, there are two possible ways of creating a comparable corpus: to create a separate subcorpus comparable to a corpus of a different language or to place in the header of selected texts an element stating that this very text is a part of a comparable corpus. Hence, the user will have the option of narrowing down their query to the texts marked as included in the comparable corpus, in the same way as they can narrow down the query to a certain author, genre or period.

A subproject of the NKJP consists in monitoring the words in daily and weekly newspapers and comparing word frequencies in two periods of time in order to test the saliency of words in every day public discourse. A demo version of this tool has recently been made available on the web site of the Association of Local Press (`http://www.gazetylokalne.pl/`). The project "Words of the Week" will be launched soon in cooperation with the weekly "Polityka" and later with Polish dailies. (A similar project was in progress on the website of the daily Rzeczpospolita from 2004 to 2007, see Łaziński and Szewczyk 2006).

2 Corpus

The intended size of the whole National Corpus of Polish is 1 billion words, of which at least 300-million word subcorpus will be carefully balanced.

2.1 Representativeness

In establishing the criteria of representativeness we build on our own experience, as well as on the experience of other national corpora, especially, the Czech National Corpus. We understand representativeness as representing the perception of language by a certain linguistic community (cf. Čermák *et al.* 1997), what in practice means reflecting the structure of readership in the structure of the entire corpus. There are theoretical reasons for this choice, besides practical considerations. Out of potential concepts of representativeness which may be applied to various corpora, the following two have the best methodological motivation: representing the population of texts or representing the structure of readership. If we adopted the model representing the production of text, around 90% of the corpus would consist of press. Hence the corpus would be representative but not balanced. The use of such a corpus in linguistic and lexicographic work is questionable (Górski 2008). The representativeness is based on several pools exploring the choices of reading of the average Polish reader as well as the circulation of the press. On the other hand we expect the corpus to be not only representative but also balanced, therefore the amount of press is lowered compared to the overall picture of the readership, so as not to let any text type dominate over the entire corpus (Górski 2009).

A first step towards establishing representativeness is determining the typology of text types. The typology is generally based on the bulk of work on Polish stylistics. There are however some text types which seem to be overseen by traditional stylistics, which had to be added. The classification is based mainly on intralingusitic features of texts. A small pilot study was conducted, so as to establish a set of purely linguistic factors differentiating the text types (Górski and Łaziński 2008). We are however aware of the fact, that the proposed classification is not the only possible.

As assigning a text to a certain type is not always a straightforward task, we decided that every text will be classified in a double-blind procedure by two linguists. In case of discrepancy they will have to discuss their decision.

To assure a wide coverage by topic we use classifications used by the libraries which will be encoded in the header of each text. To meet the needs of philologists and lexicographers we shall try as far as possible try to include in the corpus the most important works of modern Polish literature including poetry.

2.2 Spoken component

A 30 million word component of the NKJP will represent the spoken register of Polish. This part of the project is coordinated by the PELCRA team and it derives from the experience of compiling the 600 000 spoken-conversational component of the PELCRA corpus (Waliński and Pęzik 2007). Apart from transcripts of public speeches, parliamentary commission proceedings, televised debates, chat shows, radio interviews and news bulletins, a 3 million word subset of the spoken component will comprise natural, spontaneous conversations recorded by persons trained to preserve the natural character of the language data collected. Spoken NKJP data will be annotated with sociolinguistic metadata, including information on the age, gender, education and social background of the recorded speakers. Selected fragments of the spoken corpus will be aligned with the recordings and integrated in a relational database engine on top of which a publicly accessible web interface will be implemented (Pęzik *et al.* 2004).

2.3 Annotation

The entire corpus will be annotated linguistically, structurally and with bibliographic metadata. The basis of the linguistic annotation will be the full morphosyntactic annotation (not just parts of speech, but also values of cases, genders, etc., as appropriate). As in the IPI PAN Corpus, each segment (token) will contain not only the information about which morphosyntactic interpretation is correct in a given context, but also about all the other possible interpretations, rejected in the context (Przepiórkowski *et al.* 2004). Apart from the morphosyntactic annotation, the corpus will contain partial syntactic information, i.e., main types of syntactic groups will be identified, as well as named entities and various kinds of lexical constructions (so-called syntactic words). Also, forms of over 100 frequent semantically ambiguous lexemes will be disambiguated.

Because of the size of the corpus, it will not be possible to annotate the whole corpus manually. However, a 1-million word subcorpus of the representative 300-million subcorpus, reflecting its structure, is being annotated manually and it will be utilised for training and testing of linguistic tools which will subsequently be used for the automatic annotation of the whole corpus.

3 Recent developments

3.1 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 2009 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 standards proposed by ISO TC 37 / SC 4, 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 2009 (cf. also Bański and Przepiórkowski 2009), resulted in TEI P5 XML schemata encoding data models largely isomorphic with or at least mappable to those formats.

3.2 Annotation tools

As mentioned in §2.3 above, 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 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 2009b), if necessary.

Currently annotation is performed at the levels of sentence segmentation, word-level segmentation and morphosyntax, with word sense disambiguation expected to start in August 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.

3.3 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 is based on the combination of Apache Lucene (`http://lucene.apache.org/`) and relational database technologies, and it is partly inspired by the implementation of the PELCRA Corpus of Polish: we expect this approach to scale well with the size of the corpus. Apart from scalability, this search engine also focuses on providing convenient access to concordance and collocation search results in a variety of output formats, including downloadable spreadsheets, compressed URL-s, integrated browser plugins and web services. 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 not well formed queries, and an option to randomise search results.

Both search engines are successfully employed in the NKJP Demo (`http://nkjp.pl/index.php?page=6&lang=1`), which currently consists of about 500 million words.

4 Conclusion

As any "recent developments" 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. We hope this overview paper has whetted the reader's appetite for the final project results.

Bibliography

Bański, P. and Przepiórkowski, A. (2009). Stand-off TEI annotation: the case of the National Corpus of Polish. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III) at ACL-IJCNLP 2009*, Singapore.

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

Čermák, F., Králík, J., and Kučera, K. (1997). Recepce současné češtiny a reprezentativnost korpusu. *Slovo a slovesnost*, **58**, 117–124.

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.

Górski, R. L. (2008). Representativeness of the written component of a large reference corpus of Polish. Primary notes. Forthcoming.

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

Górski, R. L. and Łaziński, M. (2008). Wzór stylu i wzór na styl. Zróżnicowanie stylistyczne tekstów Narodowego Korpusu Języka Polskiego. Talk delivered at the *VII Forum Kultury Słowa*, Gdańsk, October 2008.

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.

Łaziński, M. and Szewczyk, M. (2006). Słowa klucze w semantyce i statystyce. słowa tygodnia „Rzeczpospolitej”. *Biuletyn Polskiego Towarzystwa Językoznawczego*, **LXII**, 57–68.

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.

Pęzik, P., Levin, E., and Uzar, R. (2004). Developing relational databases for corpus linguistics. In B. Lewandowska-Tomaszczyk, editor, *The proceedings* *of Practical Applications in Language and Computers PALC 2003*, Frankfurt am Main. Peter Lang.

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 *Proceedings of the Mondilex workshop in Warsaw, June 2009.*

Przepiórkowski, A. (2009b). Zasady znakowania morfosyntaktycznego w NKJP. Version 1.19 of 27 July 2009.

Przepiórkowski, A. and Bański, P. (2009). Which XML standards for multilevel corpus annotation? Unpublished manuscript.

Przepiórkowski, A. and Murzynowski, G. (2009). Manual annotation of the National Corpus of Polish with Anotatornia. Talk delivered at *Practical Applications in Language and Computers (PALC 2009)*, Łódź, April 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.

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.

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

Waliński, J. and Pęzik, P. (2007). Web access interface to the PELCRA referential corpus of Polish. In J. Waliński, K. Kredens, and S. Goźdź-Roszkowski, editors, *The proceedings of Practical Applications in Language and Computers PALC 2005*, pages 65–86, Frankfurt am Main. Peter Lang.

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.