Towards the National Corpus of Polish

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

[1]Institute of Computer Science
Polish Academy of Sciences
ul. Ordona 21
01-237 Warszawa
Poland

[2]Institute of Polish Language
Polish Academy of Sciences
al. Mickiewicza 31
31-120 Kraków
Poland

[3]Polish Scientific Publishers PWN
ul. Miodowa 10
00-251 Warszawa
Poland

[4]University of Łódź
Al. Kościuszki 65
90-514 Łódź
Poland

[5]University of Warsaw
Krakowskie Przedmieście 26/28
00-927 Warszawa
Poland

adamp@ipipan.waw.pl, RafalG@ijp-pan.krakow.pl, blt@uni.lodz.pl, M.Lazinski@uw.edu.pl

Abstract

This paper presents a new corpus project, aiming at building a national corpus of Polish. What makes it different from a typical YACP (Yet Another Corpus Project) is 1) the fact that all four partners in the project have in the past constructed corpora of Polish, sometimes in the spirit of collaboration, at other times — in the spirit of competition, 2) the partners bring into the project varying areas of expertise and experience, so the synergy effect is anticipated, 3) the corpus will be built with an eye on specific applications in various fields, including lexicography (the corpus will be the empirical basis of a new large general dictionary of Polish) and natural language processing (a number of NLP tools will be constructed within the project).

1. Introduction

The aim of this paper is to introduce a new large corpus project, *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.

What makes it different from a typical YACP (Yet Another Corpus Project) is 1) the fact that all four partners in the project have in the past constructed corpora of Polish, sometimes in the spirit of collaboration, at other times — in the spirit of competition, 2) the partners bring into the project varying areas of expertise and experience, so the synergy effect is anticipated, 3) the corpus will be built with specific applications in various fields in mind, including lexicography (the corpus will be the empirical basis of a new large general dictionary of Polish) and natural language processing (a number of NLP tools will be constructed within the project).

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[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.

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. The authors of this note, representing each of these four partners, constitute the Project Management Board.

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 In-

---

[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 1960. 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.

formatics, where two of the members of the Project Management Board lecture on corpus linguistics, linguistic engineering and Polish grammar.

1.3. Lexicographic Aims

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. A strict collaboration between corpus compilers and lexicographers will ensure that the compilers of the corpus will receive feedback from the lexicographers, so as to design better tools for use in lexicography. A similar feedback will be given by the experience of PWN which is the biggest publisher of Polish dictionaries.

In what follows we describe the main characteristics of the planned corpus and the methodology involved in its development (§2.), linguistic tools and search engines which will be constructed in the process (§3.), as well as possible practical applications of the NKJP, outside of linguistics and computer science (§4.).

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 intend to 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 structure of the readership of a linguistic community" (cf. Čermák et al. (1997)). There are theoretical reasons for this choice, besides the 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 of representing the production of text, around 90% of the corpus would consist of the press. Hence, the corpus would be representative, but not balanced. The use of such a corpus in lexicographic work is questionable (Górski, 2008).

Although we do not intend to carry out a survey answering the question of preferences for reading in Poland, there are sources which enable us to reconstruct the picture of readership. Before we draw up the design principles for the corpus we have to establish an objective text and genre typology. As there is no common consent about this topic, a pilot study will be conducted, so as to establish a set of intralinguistic factors differentiating genres one from another. This is possible, because there is a considerable number of texts already available from the four existing corpora.

Finally, as we envisage the use of the corpus in large lexicographic enterprises, not only the representativeness of the corpus is important, but also its maximal thematic and stylistic diversity. Reaching a reasonable compromise between representativeness and diversity will be one of the main objectives in the composition of the NKJP.

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 the 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 at least partial syntactic information, i.e., main types of syntactic groups will be identified. Moreover, we are planning to perform some Word Sense Disambiguation.

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, will be 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. For any type of manual annotation, there will be at least two annotators working in parallel, so that the inherent difficulty of the task may be measured (via the inter-annotator agreement) and so that the annotation may be improved (by solving any discrepancies and appropriately modifying the annotator guidelines, if necessary).

In various kinds of annotation we will follow standards and best practices, to the extent to which such standards are compatible with the linguistic system of Polish. In particular, we plan to carefully consider the recommendations of the ISO/TC 37/SC 4 subcommittee, the TEI guidelines, any future recommendations of the CLARIN project (http://www.clarin.eu/), and build on the previous experience of the IPI PAN Corpus with the XCES format (Ide et al., 2000; Suderman and Ide, 2006).

3. Tools

Two sets of tools will be developed within this project: linguistic annotation tools and efficient corpus search tools.

3.1. Linguistic Tools

Two taggers were developed within earlier projects carried out at ICS PAS, cf. Dębowski (2004) and Piasecki and Godlewski (2006), but their accuracy is only about 92.5%. This is partially due to the very large tagset of Polish (around 1150 naturally occurring tags out of 4179 potential tags; cf. Przepiórkowski (2005)), but also due to the lack of a large carefully cross-checked training corpus. Substantial increase of the tagger accuracy, on the basis of the 1-million word manually annotated subcorpus mentioned above, is one of the aims of this project.

The corpus will also be annotated syntactically. This will be achieved with the help of improved versions of (one or the combination of) the syntactic parsers developed at ICS PAS: the shallow parser Spejd (Przepiórkowski and Buczyński, 2007) and the deep parser Świgra (Woliński, 2005).

The third tool will be a Named Entity Recognition (NER) tool, which should be able to recognise out-of-vocabulary (unseen) named entities. Previous attempts at NER for Polish exist (e.g., Piskorski (2004)), and we expect to build on them.

Finally, a tool which will have to be created from scratch, so its quality may be considerably lower than for the morphosyntactic and syntactic annotation, is a Word Sense Disambiguation program, distinguishing between the basic senses of verbal, nominal, adjectival and adverbial forms. To the best of our knowledge, no such tools exist for Polish.

3.2. Indexing and Search Tools

Since developing an efficient search tool able to manage a 1-billion corpus is a potentially high-risk task, we plan to pursue two approaches in parallel. The first approach is based on the standard techniques of relational databases, as implemented for the PELCRA Corpus of Polish (`http://korpus.ia.uni.lodz.pl/`): we expect this approach to scale well with the size of the corpus, although it is not clear to what extent it may accommodate more complex types of linguistic search at various levels of annotation. The second approach is based on Poliqarp (Janus and Przepiórkowski, 2007a; Janus and Przepiórkowski, 2007b), 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.[3]

4. Practical Applications in the Humanities

The main aim of the external corpus users — lexicographers and linguists — will be searching for word and phrase concordances. 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 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 newspapers and comparing word frequencies in two periods of time. On the basis of the difference in the distribution of words, the system will select the most frequent and most imporatant proper and foreign names, which should be paid special attention in spelling dictionaries and search tools. The monitoring can also serve as a test of popularity of politicians and celebrities (see Łaziński and Szewczyk (2006)).

5. Conclusion

Building a very large corpus requires solving many scientifically and technologically interesting problems, such as the problem of what it means for a corpus to be representative or well balanced, and how to substantially improve the accuracy of a tagger, but also many mundane issues concerning the copyright law, text acquisition, data formats, etc. Probably corpus developers often do not fully realise the complexity of the project when they embark on it.

The current corpus project is perhaps unique in that all partners have considerable previous experience in constructing large corpora, which have become well established in Polish linguistics and computational linguistics. We hope to have learned from our and each other's errors and we are looking forward to the synergy effect that will doubtlessly take place, given the varied areas of expertise and foci of interest of the partners of this project.

6. References

František Čermák, Jan Králík, and Karel Kučera. 1997. Recepce současné češtiny a reprezentativnost korpusu. *Slovo a slovesnost*, 58:117–124.

Łukasz Dębowski. 2004. Trigram morphosyntactic tagger for Polish. In Mieczysław A. Kłopotek, Sławomir T. Wierzchoń, and Krzysztof Trojanowski, editors, *Intelligent Information Processing and Web Mining*, Advances in Soft Computing, pages 409–413. Springer-Verlag, Berlin.

---

[3]Experiments with other publicly available open source search engines are also planned, including Lucene (`http://lucene.apache.org/`), Xaira (`http://www.oucs.ox.ac.uk/rts/xaira/`) and Manatee/Bonito (`http://www.textforge.cz/`; Rychlý and Smrž (2004)).

Rafał L. Górski. 2008. Representativeness of the written component of a large reference corpus of Polish. Primary notes. Forthcoming.

Nancy Ide, Patrice Bonhomme, and Laurent Romary. 2000. XCES: An XML-based standard for linguistic corpora. In *Proceedings of the Third International Conference on Language Resources and Evaluation, LREC 2000*, pages 825–830, Athens. ELRA.

Daniel Janus and Adam Przepiórkowski. 2007a. Poliqarp 1.0: Some technical aspects of a linguistic search engine for large corpora. In Waliński et al. (Waliński et al., 2007).

Daniel Janus and Adam Przepiórkowski. 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.

Ida Kurcz, Andrzej Lewicki, Jadwiga Sambor, Krzysztof Szafran, and Jerzy Woronczak. 1990. *Słownik frekwencyjny polszczyzny współczesnej.* Wydawnictwo Instytutu Języka Polskiego PAN, Cracow.

LREC. 2004. *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004*, Lisbon.

Marek Łaziński and Monika Szewczyk. 2006. Słowa klucze w semantyce i statystyce. słowa tygodnia „Rzeczpospolitej". *Biuletyn Polskiego Towarzystwa Językoznawczego*, LXII:57–68.

Maciej Piasecki and Grzegorz Godlewski. 2006. Reductionistic, tree and rule based tagger for Polish. In Mieczysław A. Kłopotek, Sławomir T. Wierzchoń, and Krzysztof Trojanowski, editors, *Intelligent Information Processing and Web Mining*, Advances in Soft Computing, pages 531–540. Springer-Verlag, Berlin.

Jakub Piskorski. 2004. Extraction of Polish named-entities. In LREC (LREC, 2004), pages 313–316.

Adam Przepiórkowski and Aleksander Buczyński. 2007. ♠: Shallow Parsing and Disambiguation Engine. In Zygmunt Vetulani, editor, *Proceedings of the 3rd Language & Technology Conference*, pages 340–344, Poznań, Poland.

Adam Przepiórkowski, Zygmunt Krynicki, Łukasz Dębowski, Marcin Woliński, Daniel Janus, and Piotr Bański. 2004. A search tool for corpora with positional tagsets and ambiguities. In LREC (LREC, 2004), pages 1235–1238.

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

Adam Przepiórkowski. 2005. The IPI PAN Corpus in numbers. In Vetulani (Vetulani, 2005), pages 27–31.

Piotr Pęzik, Eric Levin, and Rafał Uzar. 2004. Developing relational databases for corpus linguistics. In Barbara Lewandowska-Tomaszczyk, editor, *The proceedings of Practical Applications in Language and Computers PALC 2003*, Frankfurt am Main. Peter Lang.

Pavel Rychlý and Pavel Smrž. 2004. Manatee, Bonito and Word Sketches for Czech. In *Proceedings of the International Conference «Corpus Linguistics – 2004»*, pages 124–132, St. Petersburg. St. Petersburg State University Press.

Keith Suderman and Nancy Ide. 2006. Layering and merging linguistic annotations. In *Proceedings of the 5th Workshop on NLP and XML (NLPXML-2006): Multi-Dimensional Markup in Natural Language Processing*, pages 89–92, Trento. ACL.

Zygmunt Vetulani, editor. 2005. *Proceedings of the 2nd Language & Technology Conference*, Poznań, Poland.

Jacek Waliński and Piotr Pęzik. 2007. Web access interface to the PELCRA referential corpus of Polish. In Waliński et al. (Waliński et al., 2007), pages 65–86.

Jacek Waliński, Krzysztof Kredens, and Stanisław Goźdź-Roszkowski, editors. 2007. *The proceedings of Practical Applications in Language and Computers PALC 2005*, Frankfurt am Main. Peter Lang.

Marcin Woliński. 2005. An efficient implementation of a large grammar of Polish. In Vetulani (Vetulani, 2005), pages 343–347.