A BNC-comparable Corpus of Polish – Applications in Pedagogy

Rafał L. GÓRSKI
Department of General and Indo-European Linguistics, Jagiellonian University
Institute of Polish Language, Polish Academy of Sciences
Cracow, Poland
rafalg@ijp-pan.krakow.pl

This poster introduces an ongoing project of an English-Polish comparable corpus, or strictly speaking, a monolingual Polish corpus directly comparable to British National Corpus (BNC). It shall overcome some well known shortcomings of a parallel corpus such as small size, influence of the source language on the target language, lack of texts translated from Polish to English or shortage of certain text types. It is not intended to replace a parallel corpus, but rather to complement it. In the talk we shall discuss the degree of comparability at various levels: design, size and tagset, as well as possible diachronic discrepancy.

The Polish counterpart of the BNC will be drawn from the resources of the National Corpus of Polish (NCP), a project which was completed in 2011 (see Przepiórkowski *et al.* 2008; forthcoming). Thus, it shall relay also on the structural and linguistic annotation of the latter. One should bear in mind that such a task is more feasible than compiling a parallel Polish-English or a PNC-comparable corpus of English texts from scratch.

We reclassify all texts according to the guidelines of classification of BNC so as to replicate its structure as far as it is possible. Although the NCP is much larger than the BNC, there are some texts which are underrepresented in the former compared to the latter. The last step of the project is an estimation of the comparability of the two corpora.

Although the main objective of the corpus is comparative linguistics, it is also compiled with the eye on pedagogy.

Pęzik (2011) reports the use of NCP in the training of translators, especially for achieving naturalness in the use of collocations. We hope that the planned corpus will be ahead of the original NCP for this purpose. Since specific collocations belong to different registers of the language, lists of collocations in two languages are more reliable when extracted from corpora with the same design. The same holds for lexis.

Students are expected to recognize the diversification of registers of language not only of the language of their study, but also of their mother-tongue, especially when they are taught translation. The instruction should also raise their awareness of the differences of the features of a particular register in L1 and L2, in our case English and Polish. A best tool to acquire this knowledge is a corpus. However, as well known, the typology of text-types varies from corpus to corpus, so it is essential to assure that one compares really homogeneous types. Worse still, similarly labeled genres in fact can be different to some extent in two corpora, especially if each text is described by multiple categories (genre, level, topic etc).

A thorough study of the variation among registers both in a monolingual and bilingual perspective can improve teaching materials, especially for keeping naturalness, by drawing the attention on differences and similarities. Last but not least comparative (corpus) linguistics in turn helps to improve the curriculum.

Certainly all these tasks can be achieved with two different monolingual corpora, nevertheless the results obtained from them will be either less reliable or will demand extra work on normalizing the data. Bearing in mind that the work needed for the creation of such a corpus is considerably smaller then compiling a corpus from scratch, the undertaking is worth the effort.

References

Pęzik, P. (2011). 'Providing corpus feedback for translators with the PELCRA search engine for NKJP'. In Góźdź-Roszkowski, S. (ed.) *Explorations across Languages and Corpora. PALC 2009*. Frankfurt am Main: Peter Lang. 135-134.

Przepiórkowski A., Górski R. L., Lewandowska-Tomaszczyk B. & Łaziński M. (2008). 'Towards the National Corpus of Polish.' In: *Proceedings of the 6th Language Resources and Evaluation Conference (LREC 2008)*. Marrakesh, Morocco.

Przepiórkowski, A., Górski R. L., Lewandowska-Tomaszczyk B. & Bańko, M. (forthcoming). *Narodowy Korpus Języka Polskiego* [The National Corpus of Polish]. Warszawa: Wydawnictwo Naukowe PWN.