PACO – a free online parallel concordancer for English–Polish Parallel Corpus

Grzegorz Krynicki (Adam Mickiewicz University, Poznań)

Parallel corpora are becoming an important source of reference for many language-related domains including machine and human translation, contrastive language study and Foreign Language Teaching (Brown et al. 1993, Malmkjaer 1998, Ebeling 1998, Barlow 1996a, b). The successful application of parallel texts in any of these domains depends on the presence of parallel concordancing programs that enable search and analysis of large quantities of aligned bitexts.

The paper will present the English-Polish Parallel Corpus (EPPC) collected by the author and an free online parallel concordance program (PACO) for search and analysis of EPPC.

EPPC is a 52-million-token parallel bi-directional corpus aligned at the sentence level. It is composed of mainly literary, legal and computer-science texts. EPPC is lemmatised and annotated for part of speech and inflection. Each sentence in EPPC is also annotated for the frequency of its component words as verified against monolingual reference corpora.

PACO is a parallel concordancer with a Web interface for search and analysis of XML-annotated parallel corpora. PACO applied to the EPPC provides boolean search of its content and of its annotation layers. The search may be limited to selected sections of EPPC if necessary. PACO also provides descriptive statistics about the corpus text and annotation tags.

An important feature of PACO is that it can be used to add user-defined texts to EPPC. Newly added texts are automatically aligned at the sentence level by means of the *hunalign* module (Halácsy et al. 2005), indexed and added to the corpus.

PACO additionally enables the analysis of bilingual terminology automatically induced from the EPPC content by means of word-alignment IBM model 1 algorithm (Brown et al. 1993).

The paper will present how PACO can be used in Machine Translation, Computer Assisted Translation, bilingual Polish-English lexicography and teaching English as a foreign language.

References

Barlow, Michael. 1996a. “Parallel texts in language teaching”, in: Simon Botley – Julia Glass – Tony McEnery – Andrew Wilson (eds), 45–56.

Barlow, Michael. 1996b. “Analysing parallel texts with ParaConc”, in ALLC/ACH '96 Proceedings, Universidad de Bergen, Noruega. http://gandalf.aksis.uib.no/allc/barlow.pdf (last access: May 18, 2007).

Brown, F. Peter – Stephen A. Della Pietra – Vincent J. Della Pietra – Robert L. Mercer. 1993. “The mathematics of Statistical Machine Translation: Parameter estimation”, *Computational Linguistics* 19, 2: 263–311.

Ebeling, Jarle. 1998. “Contrastive linguistics, translation, and parallel corpora”, *Meta* 43(4).

Halácsy, Péter – Dániel Varga – András Kornai – Viktor Nagy – László Németh – Viktor Trón. 2005. “Parallel corpora for medium density languages”, *Proceedings of RANLP 2005*. Borovets, Bulgaria. Pp. 590–596.

Halverson, Sandra. 1998. “Translation studies and representative corpora: Establishing links between translation corpora, theoretical/descriptive categories and a conception of the object of study”, *Meta* 43, 4: 494–514.

Malmkjaer, Kristen. 1998. “Love thy neighbour: Will parallel corpora endear linguists to translators”, *Meta* 43, 4: 534–541.