

The TEI and the NCP: the Model and its Application

Piotr Bański

Institute of English Studies,
University of Warsaw
pkbanski@uw.edu.pl

Adam Przepiórkowski

Institute of Computer Science,
Polish Academy of Sciences
adamp@ipipan.waw.pl

**We wish to acknowledge support from grant #R1700303
from the Polish Ministry of Science and Higher Education.**

National Corpus of Polish: basic information

- <http://nkjp.pl/>
- 2007–2010
- text corpus with a sizable spoken transcript part
- slightly over 10^9 segments, available for searching
- 300 million segments balanced
- 1 million hand-annotated/validated
- TEI XML, stand-off markup
- annotation layers designed to be isomorphic with the current best practices and emerging standards
- source + several layers of annotation (see the poster)
- tools under the GNU GPL