Freely distributable subcorpora of the National Corpus of Polish

Łukasz Degórski
Institute of Computer Science, Polish Academy of Sciences

National Corpus of Polish

- $1.5 \cdot 10^9$ segments in total
- 250 million words in the balanced subcorpus
- 1 million words in the manually annotated subcorpus
- two corpus search engines
- **http://nkjp.pl**

Przepiórkowski, A., Górski, R. L., Łaziński, M. i Pęzik, P. (2010). *Recent developments in the National Corpus of Polish*. In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, LREC 2010, Valletta, Malta. ELRA.

Przepiórkowski, A., Bańko, M., Górski, R. L. i Lewandowska-Tomaszczyk, B., eds. (2012). *Narodowy Korpus Języka Polskiego* [Eng.: *National Corpus of Polish*]. Wydawnictwo Naukowe PWN, Warsaw.

Copyright-free subcorpus

- all texts from the whole corpus that are free from intelectual property constraints
- can be distributed without limitations
- almost 100 million words
- definitely unbalanced – most texts are parliament proceedings or law texts (acts)
- automatically annotated, levels: segmentation, morphosyntax, word senses, syntax words and syntax groups
- downloadable (15GB):

http://zil.ipipan.waw.pl/DistrNKJP

Manually annotated subcorpus

- does not contain full texts – just random samples
- as such can be distributed (GNU GPL v.3)
- balanced
- each sample: 40-70 words, logically consistent
- manually annotated (the most labour intensive subtask of the Corpus project), levels: segmentation, morphosyntax, word senses, syntax words and syntax groups, named entities
- downloadable (157MB):

http://clip.ipipan.waw.pl/LRT