A project of a BNC comparable corpus of Polish

Rafał L. Górski
Institute of Polish Language
Cracow

UCCTS 2010
Omskirk
Parallel vs comparable

- **Parallel**
  - Same texts in L1 and L2 (translations)
  - Concordancer shows KWIC in L1 and its L2 translation

- **Comparable**
  - Two (or more) corpora of different languages with a same design
Parallel

• Advantages
  • A phenomenon \( Z \) in L1 can be directly compared to a phenomenon \( \Omega \) in L2 in the same context
Parallel

- **Advantages**
  - A phenomenon $Z$ in L1 can be directly compared to a phenomenon $\Omega$ in L2 in the **same context**

- **Disadvantages**
  - Small size
  - Bad balance
  - Influence of SL on TL
Why a comparable corpus?

• Practical reasons
  • No available parallel English-Polish Corpus
  • Feasibility
Why a comparable corpus?

- Practical reasons
  - No available parallel English-Polish Corpus
  - Feasibility

- Theoretical reasons
  - Size
  - Balance
  - No interference of SL
The possible use

- Comparative grammar
- Collocations
- Bilingual lexicography
- Cultural studies
- Etc.
BNC.pl

- Based on texts of the National Corpus of Polish NKJP
- www.nkjp.pl
National Corpus of Polish

• An ongoing project 2008-2010 founded by the Polish Ministry of Science and Higher Education

• Partners:
  • Institute of Computer Science
  • Institute of Polish Language
  • Chair for English, Lodz University
  • PWN Scientific Publishers
National Corpus of Polish

- 1 billion running words (opportunistic part)
- 300 million running words (balanced part)
- Structural, morphosyntactictic and shallow syntactic annotation
- WSD, named entity annotation
- Available on-line with two different concordancers
- Free unrestricted access
Compiling the corpus

- Take the texts from the National Corpus of Polish
- Label them according to text classification of the BNC
- Replicate the BNC
Compiling the corpus

Quit a straightforward task, isn't it?
Compiling the corpus

Quit a straightforward task, isn't it?

Er, not really...
Compromises...

- Mode
- Genre
- Medium
- Domain
- Keywords
- Auditorium (age, sex, level)
- Circulation status
- Author (age, sex, type)
Compromises...

- Hierarchy of features.
  - Mode
  - Genre
  - Domain
  - Medium
  - Keywords
  - Auditorium (age, sex, level)
  - Circulation status
  - Author (age, sex, type)
Size

• What is a word?

he has been watching the road
Size

• What is a word?

  he has been watching the road
  obserwował drogę

• Rybicki constant: an English text consists of ca 1,4 times more words than a Polish one
Size: a delicate question

- Speak Now or Forever Hold Your Peace!
The tagset

- Comparability of the tagsets
  - CLAWS
  - IPI PAN taset
- Comparing tagsets is doing contrastive grammar
- Some inconsistencies
Comparing tagsets

- *all* = Determiner-Pronoun
  *kazdy* = Adjective

- *Who* = Wh-pronoun
  *kto* = Noun
How comparable are they?

- Statistics estimating the similarity of labels
- Checking the homogeneity of the genres
- Number of texts making up each genre