Beyond one language – language technologies in expanding Europe

Radovan Garabík

Ľudovít Štúr Institute of Linguistics

Slovak Academy of Sciences

Bratislava

http://korpus.juls.savba.sk

Linguistic technologies

- everything is expanding
- amount of information is rapidly increasing
- necessity to process the information, but also:
- previously unprecedent possibilities to get a lot of empirical data
- modern computer systems: enough power to process complex linguistics data
- harvest the information produced in modern society &
- provide new ways of dealing with traditional linguistics theories

What is trendy in modern linguistics

- morfological dictionaries
- thesauri
- semantic networks
- syntactic dictionaries
- encyclopædical dictionaries
- terminological databases
- speech recognition & generation
- natural language analysis
- huge language corpora

Slovak National Corpus

- we want to be trendy as well!
- taking a stab at almost everything mentioned above
- but no audio examples :-)

Karl-Franzens-Universität, Institut für Slawistik, Graz

- Quantitative Text Analysis
- Word Length (Frequencies) and their Distribution in Slavic Texts

$$P_x = \frac{\binom{M+x-1}{x} \cdot \binom{K-M+n-x}{n-x+1}}{\binom{K+n-1}{n}}$$

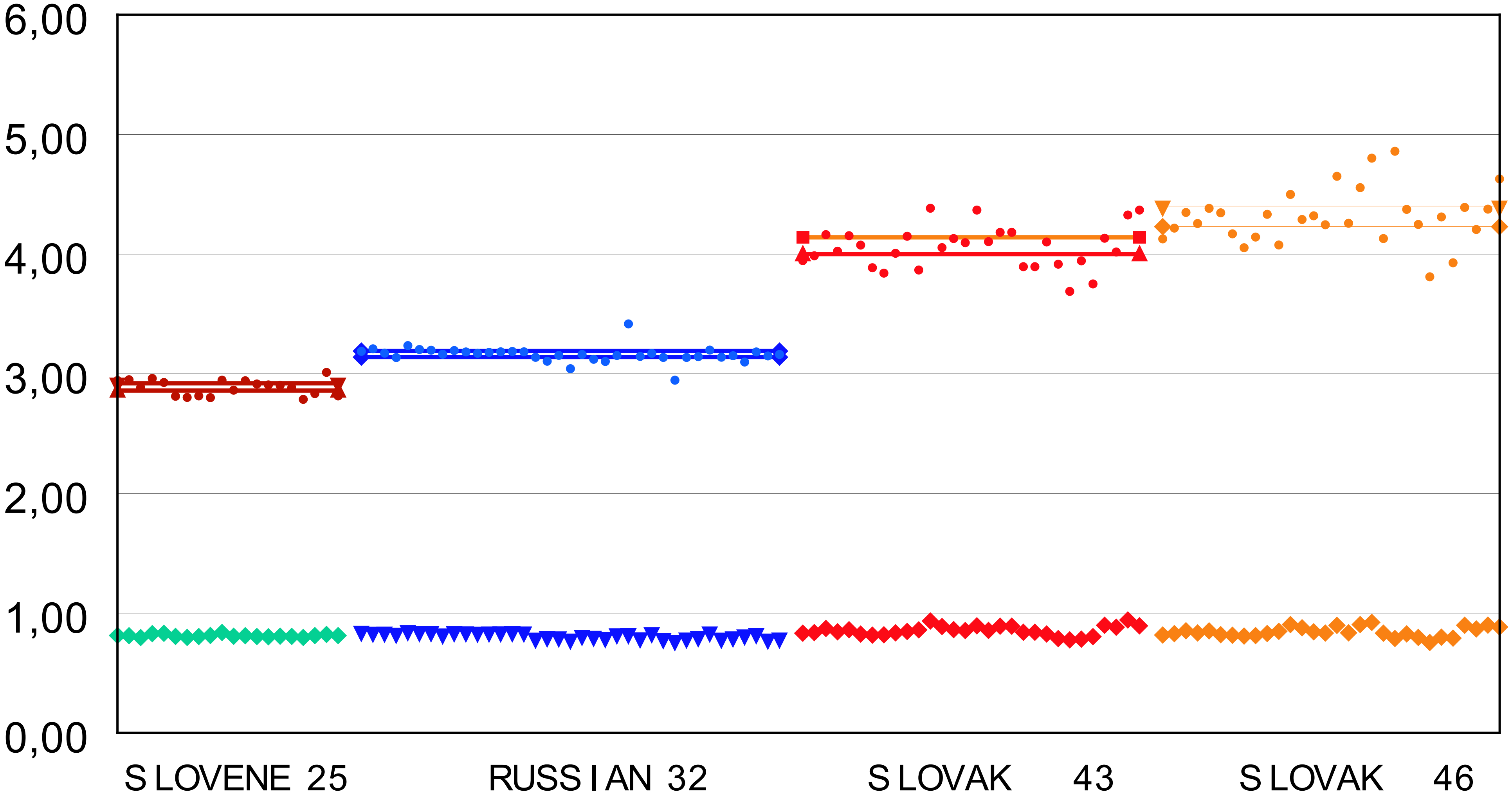

Illustration 1: Parameters K and M for 4 Slavic languages

Universität Regensburg, Institut für Slawistik

- parallel corpora (German, Russian, Slovak, English, Czech, Polish, Ukrainian)
- automatic text alignment
- lemmatized texts