Parallel Corpora compiled at the Slovak National Corpus
Workshop on parallel corpora, Mainz, 2012-09-10

Radovan Garabík

Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava
Overview of PC

- Slovak-Russian (100 k sentences)
- Slovak-French (1100 k sentences)
- Slovak-Czech “fiction” (740 k sentences)
- Slovak-Czech combined (6.4 M sentences)
- Slovak-English fiction\(^1\) (4.3 M sentences)
- Slovak-English combined (10 M sentences)
- Slovak-Bulgarian (20 k sentences)
- Slovak-Latin (22 k sentences)
- Slovak-Chinese
- Czech-English-Russian UMC 0.1

- hunalign, manatee backend, web interface (Karrigell, old & new)

\(^1\)supported by the grant FP7-ICT-2007-3-231720 (EuroMatrixPlus)
Slovak-Russian parallel corpus

- Faculty of Philology, St. Petersburg State University
- mostly fiction
- Russian→Slovak only, 100 k sentences
- 2006
- 100 k sentences, 1.9 M words
- http://korpus.juls.savba.sk/parus/ (old interface)
- lemmatised, annotated (Dialing)

- new version
- TreeTagger
Slovak-French parallel corpus

- fiction
- Acquis communautaire
- French→Slovak only
- 2006
- 1.1 M sentences, 66 M words fr, 59 M words sk
- http://korpus.juls.savba.sk/frask/ (old interface)
- lemmatised, annotated (TreeTagger)
- double treatment of the ligature œ
Slovak-Czech parallel corpus

- fiction, popular science
- Czech↔Slovak, 3rd language\(^2\)→Czech+Slovak
- 2010
- 740 ksentences, 10.5 Mwords
- 15 ksentences manually aligned
- [http://korpus.juls.savba.sk/skcs.html](http://korpus.juls.savba.sk/skcs.html) (new interface)
- lemmatised, annotated (ČNK tagset)
- [http://slovnicky.korpus.sk/?d=pskcs](http://slovnicky.korpus.sk/?d=pskcs)

\(^2\)75 % English, the rest German, Polish, Italian, French and Ancient Greek
Slovak-Czech free corpora

- Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University in Prague
- texts unencumbered by copyright
- EU, EC (europarl, JRC-Acquis, EC journal, OPUS EMEA, EUconst, Official Journal of the EU)
- OpenSource manuals (OPUS KDE4, PHP)
- English→Czech+Slovak
- 2012
- downloadable – lemmatised, morphologically annotated
- combined with the “fiction” corpus –
  - http://korpus.juls.savba.sk/skcs.html
- 6.4 M sentences, 120 M words per language
Slovak-English parallel corpus

- mostly fiction (≈ 564 books)
- English→Slovak, small part Slovak↔English
- 2011, enlarged in 2012
- 4.3 M sentences, 63.3 M words en, 53.9 M words sk
- http://korpus.juls.savba.sk/sken.html (new interface)
- lemmatised, POS tagged (TreeTagger, Penn TreeBank tagset)
- http://slovniky.korpus.sk/?d=psken
Slovak-English free corpora

- Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University in Prague
- texts unencumbered by copyright
- EU, EC (europarl, JRC-Acquis, EC journal, OPUS EMEA, EUconst, Official Journal of the EU)
- OpenSource manuals (OPUS KDE4, PHP)
- English→Slovak
- 2012
- downloadable – lemmatised, morphologically annotated
- combined with the “fiction” corpus –
  http://korpus.juls.savba.sk/sken.html
- 10 M sentences, 196 M words en, 173 M words sk
Slovak-Bulgarian parallel corpus

- Institute of Mathematics and Informatics, BAN
- fiction
- Bulgarian↔Slovak, 3rd language→Bulgarian+Slovak
- 2011
- 20 ksentences, 291 kwords bg, 268 kwords sk
- http://korpus.juls.savba.sk:8090/ (new interface)
- *not* lemmatised\(^3\)
- currently stalled

\(^3\)but promised
Slovak-Latin parallel corpus

- classical, medieval & modern language
- Latin → Slovak exclusively
- 2011–
- http://korpus.juls.savba.sk/skla.html (new interface)
- lemmatised, POS tagged (TreeTagger)
- 22 ksentences (580 kwords la, 510 kwords sk)
- currently expanding
Slovak-Chinese parallel corpus

- Faculty of Philosophy/Faculty of Arts, Comenius University, Bratislava
- a different aim – learners’ corpus
- Chinese↔Slovak
- pǔtōnghuà, simplified script
- 1:1 translations
- manually segmented, aligned, morphologically & POS tagged, lemmatised
- pīnyīn, numerical pin1yin1, bopomofo
- vapourware
- goal – several ksentences
File formats

- it’s a mess
- internally: vertical file, tab-separated attributes, links to aligned sentences
- OPUS, europarl, UMC sentence aligned files
- Intercorp - alignment in a separate file
Corpus interface

- web based
- python, Karrigell
- manatee (1) backend
- CQP syntax
- regular expressions
- shortcuts: no square brackets →
  
  \[\text{lemma}="(?i)x" \mid \text{word}="(?i)x"\]
Dictionary interface

- **dictd (RFC 2229)**
- **web interface**
- **exact, prefix, suffix, substring, re**
- **index:**
  - exact form (dictd will take care of case)
  - ASCII-only separate index
- **MOSES phrase tables – factored corpora word|lemma**
- **filter better alignments**
- **4 word long**
- **lemmas and words as indices**
- **sk-cs: 2.4M pairs, 20M index words, 459k unique**
- **sk-en: 550k pairs, 4.5M index words, 155k unique**
Thank you for the attention