Slovak National Corpus tools and resources

Radovan Garabík
Ľ. Štúr Institute of Linguistics
Slovak Academy of Sciences
Bratislava
Slovak National Corpus

- huge, representative corpus of modern written Slovak (1955+)
- 770 Mtokens
- specialised subcorpora (fiction, professional texts, journalistic texts, original Slovak fiction, balanced subcorpus, 55..89)
- metadata (bibliography, style/genre annotation)
- Manatee, Bonito, web interface – CQL, regular expressions, word/lemma/tag
. manually morphologically annotated corpus, 1.2 Mtokens
. automatically lemmatized, morphosyntactically annotated
Morphology

- tasks: given wordform, find (plausible) lemma
- given wordform, find grammar categories
- given lemma, enumerate all wordforms (+grammar categories)
- morphology database: 74000 lemmata, full paradigms
- MoinMoin, edit full paradigms
- disambiguation: morče
Corpus of Spoken Slovak

- sound samples of contemporary Slovak
- manual (±phonemic) transcription
- automatic lemmatisation and morphosyntactic analysis
- archive: FLAC, 48 kHz, 24-bit (200 GB)
- normalize
  - ➔ Ogg/Vorbis (aoTuV oggenc, mono, quality=0)
  - ➔ Ogg/Speex (quality=7, complexity=10)
- transcriber
- legality – copyright law, protection of personal information
- metadata: speaker, recording
- speaker: sex, age, profession, education, birthplace, liveplace, place, L1, languages, dialect, dialects, informed
- recording: topic, spontaneous, formal, familiar, equal, comment, date, place, speakers (nr. of)
- bonito, web interface – word/lemma/tag/pron
WordNet

- Slovak Online (Lifelong Learning Programme DG EAC/31/08) project
- grouping words into sets of synonyms (synsets)
- semantic relations between synsets
- hypernyms, hyponyms, holonyms, meronyms, antonyms
- automatic synset building
- parallels Princeton WordNet 3.0
- goal: 10000 words
Thank you for your attention