The comprehensive Slovenian-Hungarian dictionary: bilingual lexicography meets monolingual lexicography

Iztok Kosem
iztok.kosem@cjvt.si
Centre for Language Resources and Technologies
Faculty of Arts, University of Ljubljana
NO DICTIONARY IS AN ISLAND
Dictionary situation in Slovenia

- Slovenian-Hungarian situation:
  - existing dictionaries very outdated
  - not corpus-based
  - so far mainly compiled by individuals
  - few available in electronic form (desktop application only)
  - all under copyright
  - but still regularly used
Slovenian-Hungarian language resources (N=142)

- Slovensko-madžarski slovar (Jože Hradil)
- Szlovén-magyar szótár (Hradil Jože)
- Ma-slo slo-ma moderni slovar (E. Bernjak)
- Magyar–szlovén szlovén–magyar modern szótár (Bernjak E.)
- e-zbirke izrazov (http://eslovar.datadev.si/)
- online szószedetek (http://eslovar.datadev.si/)
- Slovensko-madžarski navzkrižni slovar (Jože Hradil)
- Szlovén-magyar magyar–szlovén szótár (Hradil Jože)
- Linguee: http://www.linguee.com/
- EUR-LEX: http://eur-lex.europa.eu/
- Glosbe: https://sl.glosbe.com/
- Drugo:
- Egyéb:

I am familiar with it
I use it regularly
Dictionary situation in Slovenia

- Bilingual dictionary situation
  - Most dictionaries outdated
  - Dictionaries under copyright (some are being bought out as part of the Development of the Slovene Language in Digital Environment project)

- Monolingual dictionary situation
  - More up-to-date (*)
  - Corpus-based
  - Good variety of types (general, specialized etc.)
  - Some available under open licence (but not the main ones)
Data model (for all types of lexicographic data)

Digital Dictionary Database (for Slovene)

- Syntactic Structures
- Lemma-Sense = Concept
- Definitions Translations
- Morphology Pronunciation
- Semantic Frames / Types
- Collocations
- Attestations Corpus Data
Lexical unit + Sense = Concept (stable id)

- Lexicographic Data as a Service (LDaaS)
- Available through API service
- Open Access (CC BY-SA)
- In CLARIN.SI repository

- Development of Slovene Language in Digital Environment (2020-2022)
Comprehensive Slovenian-Hungarian Dictionary

- Funded by Slovenian Research Agency (SRA) since 2018
- Based on a concept (the project also funded by SRA)
- Entries compiled completely from scratch
- version 1.0 at https://viri.cjvt.si/slovensko-madzarski/eng/:  
  - 10,946 entries
  - 15,265 collocations
  - 2,416 examples
  - 33,298 translations
- Date of publication: 20. 10. 2021
- Database available in CLARIN.SI repository under 4.0 CC-BY-SA license (http://hdl.handle.net/11356/1453)
Team

- Over 20 colleagues
  - lexicographers
  - lexicographers-translators
  - proofreaders
  - terminologists
  - linguists
  - computational linguists, IT staff
- International advisors:
  - dr. Jelena Kallas (Institute for the Estonian Language)
  - dr. Katalin P. Márkus (Károli Gáspár University of the Reformed Church, Budapest)
  - dr. Attila Mártonfi
  - Michael Rundell (Lexicography Masterclass)
  - dr. Carole Tiberius (Dutch Language Institute)
  - dr. Tamás Váradi (Research Institute for Linguistics, Hungarian Academy of Sciences)
Sketch Engine API adapted for our purposes.

Collocations statistics.

Longest commonest match.

Examples search: cerkven

1. Variante
2. Homonomije
3. Povezano
4. Oznake
5. Opombe

Show examples
Show clusters

Collocations

Vstavi v: 1 o zaradbi

OK

SKE

cerkven

Examples
Collocations
Dictionary content

• Novelties:
  • Slovene-centred sense division
  • Multiword units as entries
  • New system of labels
  • Speech synthesis for all parts of the dictionary (Slovenian and Hungarian)
  • Each sense has an indicator
  • Each sense attributed (at least one) semantic type
All the content from the database available, but not all is translated
Conclusions and plans

• Bilingual and monolingual lexicography exploiting mutual interests
• Dictionaries and other projects “helping” each other
  • A project of updating Thesaurus and Collocations dictionary underway
  • Semantic data for/from ELEXIS Word Sense Disambiguation task
  • New corpora, e.g. monitor corpus (*presentation tomorrow at 9am!*)
• Future
  • More bilingual data to be added (recently approved project, Ministry of Culture)
  • Customized editor for the Digital dictionary database
  • Including state-of-the-art technologies (definition extraction etc.)
  • Linking dictionaries (Dictionary Matrix, ELEXIS)
Acknowledgments

This project was funded by ARRS (Slovenian Research Agency) through the research programme P6-0411 "Language Resources and Technologies for Slovene“, and the infrastructure programme I0-0022 "Network of Research Infrastructure Centres (MRIC)“ at the University of Ljubljana.