Goranka Blagus Bartolec – WG1

Croatian Collocation Database (CCD)

5th general PARSEME meeting, 23-24 September 2015, Iași, Romania

About the CCD

- The *Croatian Collocation Database* (CCD) project (started in June 2014) has been conceived as a dynamic (upgradeable) dictionary of Croatian MWS that will be populated and processed in a relational database. The results of the project will be publicly available and searchable (through hints) on the website of the Institute of Croatian Language and Linguistics (www.ihjj.hr).
- The first (test) results of project (Croatian letters L, LJ, M and Š) are available on the following link: http://www-test.ihjj.hr/kolokacije/.

- The project is based on extensive data sources collected and processed for automatic detection by Dr. Stefan Rittgasser
- The current structure of the CCD project was constructed on the following sources:

(1) Croatian newspapers from 1998 to today
(2) various online sources
(3) contemporary Croatian lexicographical manuals
(4) *Narodne novine* online (official gazette of the Republic of Croatia)
(5) recent linguistic Croatian journals.

- Each MWE in the database will be marked with a special label according to its lexical and semantic features as a:
- **lexical or grammatical collocation**
- **phraseme / idiom**
- **free combination**
- **term**
- **pragmeme**
- **proverb**

and presented through examples of its use in everyday speech as a part of journalistic/marketing, conversational, literary style, or scientific style.

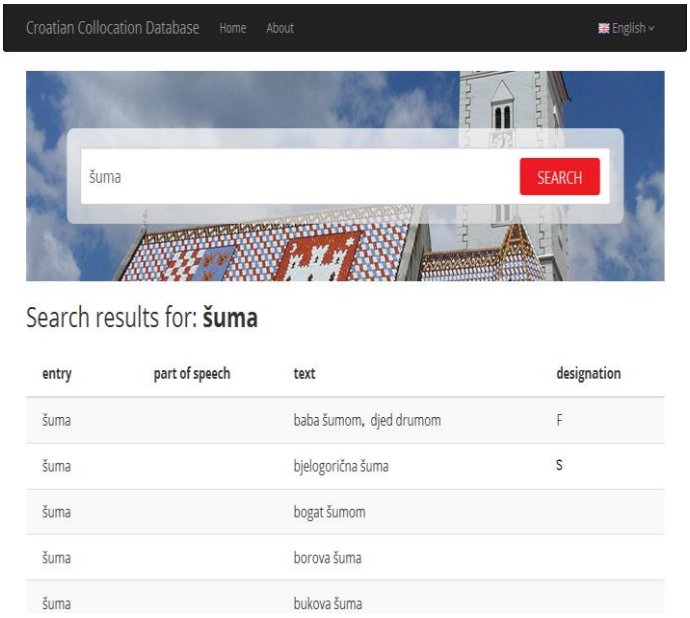
Methodology

- Data for the CCD was entered and processed in Microsoft Access under all parts of speech (noun, verb, adjective, and adverb) which form a single MWE, e.g. both the entries ***labudov*** ('swan') and ***vrat*** ('neck') contain the word combination ***labudov vrat*** ('swan neck').
- The CCD is based on the extended valence model – the basic form of word combination has been extended with words (verbs, adjectives, nouns) with which it forms the most common syntactic environment; e.g. both canonical idiom ***od vrata do vrata*** ('door to door') or canonical collocation ***lakša ozljeda*** ('minor injury') and extended form ***ići od vrata do vrata*** ('go door to door') or ***pretrpjeti lakše ozljede*** ('to suffer minor injuries´, etc.) are placed in the CCD.

- The working Access file contains nine columns: (1) **Entry** (the canonical form of the word); (2) **Part of speech** / Word class (only for homographic entries); (3) **Order of meaning** (if the word is polysemic); (4) **Text** (a key part of the database that contains the list of MWE); (5) **Synonym** (single- or multi-word synonym of the word combination in the Text column); (6) **Label** (symbol for types of MWE (the symbols are only applicable in a working Access file, not for a final public application) – phraseme / idiom (€), collocation / fixed phrase or term ($), proverb (@) and free combination (no symbol); (7) **Subject-field labels** for each word combination (e.g. Medicine, Physics, Architecture) or Usage labels (non-standard use, official or formal use, informal use, jargon/slang use, literary use, advertising use); (8) **Exclamation mark** (only in the first phase of the project if the combination is unclear or rare); (9) **Source** (if MWE is taken from the another lexicographic manuals).

The Relevance of the CCD

- At least the CCD is primarily based on traditional lexicographic and lexicological settings of multiword lexical units (Benson et al. 1997, Mel'čuk 2001), so that the main plan is to put together in one database the most common Croatian multiword lexical units by defining their semantic types and context of use.
- The database will be a useful source to be included in other more advanced MWE sources (Croatian and international) for the development of tools that permits the extraction MWEs on the basis of their semantic and lexical features (Sag et al. 2001).



References:

- Benson et al. 1997: Benson, M., E. Benson, R. Ilson. *The BBI dictionary of English word combinations*. Amsterdam – Philadelphia: John Benjamins Publishing Co.
- Blagus Bartolec 2014: Blagus Bartolec, G. *Riječi i njihovi susjedi: Kolokacijske sveze u hrvatskom jeziku*. Zagreb: Institut za hrvatski jezik i jezikoslovlje.
- Mel'čuk 2001: Mel'čuk, Igor. Collocations and Lexical Functions. – In: *Phraseology Theory, Analysis, and Applications*. Oxford – New York: Oxford University Press, pp. 23–53.
- Sag et al. 2001: Sag, I., T. Baldwin, F. Bond, A. Copestake, D. Flickinger. Multiword expressions: A pain in the neck for NLP. – In: Proceedings of CICLing-2002, Mexico City, pp. 1–15.

Institut za hrvatski jezik i jezikoslovlje

PARSEME