Croatian Collocation Database

Goranka Blagus Bartolec
Institute of Croatian Language and Linguistics, Zagreb
WG 1

1. About the Project

The *Croatian Collocation Database* (CCD*)* project (started in June 2014) has been conceived as a dynamic (upgradeable) dictionary of Croatian word combinations that will be populated and processed in a relational database. The project will last until September 2017.

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 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, who worked at the University of Heidelberg and the University of Mannheim. (His sample of database has been published on *lingua-hr.de*). The current structure of the CCD project was constructed on the following sources: (1) Croatian daily, weekly and monthly newspapers from 1998 to today; (2) various online sources; (3) contemporary Croatian lexicographical manuals (dictionaries, lexicons, encyclopaedias); (4) *Narodne novine* online (official gazette of the Republic of Croatia); (5) recent linguistic journals with articles on the topics of word combinations in Croatian.

The CCD will enable different kinds of research in the field of semantic relationships between words (synonymy, antonymy, homonymy, hyperonymy etc.), as well as exploring the possibilities of combining words to the syntagmatic level. Each word combination in the database shall be marked with a special label according to its lexical and semantic features (as a lexical or grammatical collocation, as a phraseme / idiom, as a free combination, as a term, as a pragmeme, as a proverb). Each collocation will be presented through examples of its use in everyday speech as a part of journalistic/marketing style, conversational style, literary style or scientific style. The CCD will be the basis for the creation of multilingual dictionaries of word combinations – Croatian and other European languages (English, German, French, and Russian etc.).

2. Methodology

In the first phase of the project (first year), data for the collocation database was entered and processed in Microsoft Access. A list of word combinations was entered in a separate file for each letter of the Croatian alphabet. Combinations of words were entered into the database under all parts of speech (noun, verb, adjective, and adverb) which form a combination, e.g. both the headwords *labudov* (‘swan’) and *vrat* (‘neck’) contain the word combination *labudov vrat* (‘swan neck’). The collocation database is based on the extended valence model – the basic form of word combination has been extended with words (verbs, pronouns, 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 database.

The working Access file contains nine columns: (1) Headword/Entry (the canonical form of the word); (2) Part of speech / Word class (only for homographic headwords/entries); (3) Order of meaning (if the word is polysemic); (4) Text (a key part of the database that contains the list of word combinations); (5) Synonym (single- or multi-word synonym of the word combination in the Text column); (6) Label (symbol for types of word combinations (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 for each word combination (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 the word combination is taken from the another lexicographic manuals).

Prepared data from working Access files (for each letter) will be transferred in the publicly available database as Hypertext. The publicly available database will initially have four columns (1, 2, 4, 5), and will be later upgraded with new columns as previously described. The Croatian letters L, LJ, M and Š are available in that form on the link: http://www-test.ihjj.hr/kolokacije/.

3. Relevance of the project

At least this project is primarily based on traditional lexicographic and lexicological settings of multiword lexical units (Benson et al. 1997, Blagus Bartolec 2014, 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).

4. 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.