A valency database as preparation for an online Valency Dictionary of Croatian Verbs

Matea Birtić
Institute of Croatian Language and Linguistics
mbirtic@ihjj.hr

1 Introduction

My paper attempts to identify the phases of the computer lexicographical process as proposed by Klosa (2013) in the creation of the online Valency Dictionary of Croatian Verbs. I have recognized five phases in our work as well as the beginning of a sixth phase. I cannot say anything about a seventh phase since our dictionary is still under construction. Acquaintance with Klosa's model helped me to realize some additional reasons for the failure of the first version of the Valency dictionary of Croatian verbs.

The Valency Dictionary of Croatian Verbs was an Institute of Croatian Language and Linguistics projects lasting from 2008 to 2013. The project was conducted by Institute's Croatian Standard Language Department. During this period, five associates worked on valency descriptions. The project work was not structured as a proper lexicographical or a computer- lexicographical process. The focus of the work was more on research than on planning the process of compiling a dictionary. It also lacked a computational support and a valency model was not chosen. The valency patterns of 2000 verbs were described and analyzed, but they were written in a text format (MS Word documents) and thus cannot serve as a basis for a paper dictionary or even less could be directly converted into a searchable database. There were some attempts to translate word documents in another format, all of them unsuccessful. There was not a single computational linguist or computer scientist to assist in the work. Also, the first project manager assumed a comprehensive dictionary of 24 400 verbs. As I look back at the situation yet, the project was lacking the proper lexicographic plan and it has an elusive goal. In 2013 the project was moved to the Department of General, Comparative and Computational Linguistics in order to obtain the proper computational support.

2 Lexicographical workflow

2.1 Preparation

The proper phase of preparation began only at the time of project's movement. Nevertheless, the years spent at the old project were kind of preparation for the new project, whose phase of preparation would last longer if research and acquaintance with a topic and a data were not done earlier. It was decided to start from the beginning and the first goal was to create Valency database for Croatian verbs which would serve as a basis for a paper dictionary and could be searchable online. In the spring of 2013 we had regular meetings where we discussed the conception of the dictionary. The team consists of two associates from the old project and four new members. The linguistic model has been selected and it was decided to adjust the German valency model as presented in VALBU 2004

for Croatian. We developed the analyses of valency patterns by ten groups of complements. In this stage the computational strategy has been selected too.

2.2 Data acquisition

In the phase of data acquisition we could not compile the special corpus for valency description due to insufficient tools. Therefore, our primary sources are the corpus *Croatian Language Repository*, the corpus *hrWaC* (Croatian Web) and word documents which contain already described verbs from previous project (which now serve as kind of corpus). Secondary sources include CroVallex, an already existing online valency dictionary of Croatian, German valency dictionary VALBU, Croatian grammar and syntax textbooks, and dictionaries of Croatian. We can characterize our dictionary as corpus-driven.

2.3 Computerisation

In the phase of computerisation TshwaneLex-based user interface for data entry was customized. The interface data is saved in a well-structured SQL database, which will enable us to develop a web-based query system for Valency database in the next project phase. The dictionary writing program is structured as three level user interface (first level consist of verb lemma, morphological specifications and collocations, the second level consists of verb definition(s), and the third level consists of valency patterns subordinated to the definitions).

In February 2014 the first version of a database was completed and first pilot entries were processed. The phase of computerisation lasted from May 2013 to April 2014 and further, but minor, revision are still possible due to new insights of lexicographers.

2.4 Data processing

In the phase of data processing the list of 897 verbs were selected as lemma entries on the basis of Croatian Frequency Dictionary (Moguš, M.; M. Bratanić; M. Tadić (1999) *Hrvatski čestotni rječnik*) and textbooks for studying Croatian as the second language. The both verb lists were automatically compared and final version was manually completed (f. e. by deleting some unnecessary forms originating from spelling doublets etc.). The verbs were grouped in semantic classes.

2.5 Data analysis

In the phase of data analysis we have chosen as a first module the semantic class of psychological verbs. Each associate has been given a list of ten verbs to start with. There are four lexicographers (one of the computational linguists also work as a lexicographer) and the project manager who also edits entries. As mentioned above, the analysis is performed in three layers. In this moment we are about to finish the analysis of psychological verbs and afterwards the project manager will begin with the checking and proofreading. Subsequently, the discussion on the equation of data analysis will follow with other associates.

2.6 Preparation for online release

We had preliminary talks with another computer scientist working in the Institute, and we asked him to open a subdomain *valencijski.ihjj.hr*. For web query we will follow the German model used in E-VALBU. We strive to have simple (by verbs) and advanced search (by complements).

So far, nothing has been published. We expect to have the semantic class of psychological verbs published online in October 2014. This group of verbs is our pilot project, which will enable us to get feedback from both users and experts. A given feedback will influence our further work; even change the model, which is in accordance with the thesis about overlapping of the computer lexicographical phases pointed out by Klosa (2013) and Tiberius & Schoonheim (2014).

3 Time span of the different phases

Duration	**2009.**	**2010.**	**2011.**	**2012.**	**2013.**	**2014.**	**2015.**	**2016.**	**2017.**	**2018.**	**2019.**	**2020.**
Phase												
Preparation												
Data acquisition												
Computerisation												
Data processing												
Data analysis												
Preparation for online release												
Afterlife												

Table 1 Process phases of dictionary project and their time span

4 References

Klosa, Annette (2013): The lexicographical process (with special focus on online dictionaries). In: Gouws, Rufus H. /Heid, Ulrich/Schweickard, Wolfgang/Wiegand, Herberst Ernst (Hgg.): *Dictionaries. An international Encyclopedia of Lexicography.* Supplement Volume: Recent Developments with Focus on Electronic and Computational Lexicography. Berlin, Boston: de Gruyter, S. 517-524. (Handbücher zur Sprach- und Kommunikationswissenschaft; 5.4).

Schumacher, Helmut; Jacqueline Kubczak, Renate Schmidt, Vera de Ruiter (2004): VALBU – Valenzwörterbuch deutscher Verben. Tübingen: Gunter Narr Verlag.

Tiberius, Carole; Tanneke Schoonheim (2014): The Algeemen Nederlands Woordenboek (ANW) and its Lexikographical Process, in: Vera Hildebrandt (Ed.): *Der lexikografische Prozess bei Internetbwörtebüchern. 4. Arbeitsbericht des wissenschaftliche Netzwerks "Internetlexikografie"*. Manheim: Institut für Deutsche Sprache. (Opal – Online publizierte Arbeiten zur Linguistik X/2014).

Electronic dictionaries:

E–VALBU, http://hypermedia.ids-mannheim.de/evalbu/index.html