Croatian Linguistic System Modules Overview

(Software Demonstration)

Marko Orešković, Jakov Topić, Mario Essert

National and University Library in Zagreb
Faculty of Mechanical Engineering and Naval Architecture
e-mail: moreskovic@nsk.hr, jakov.topic131@gmail.com, messert@fsb.hr

Abstract

In this paper we show several segments of program solutions which are a part of the Croatian Linguistic System (CLS) that is being developed in several ways and aims at achieving the final integration of all modules. Although the system aims at programmatically connecting all areas of linguistics (from phonetics to discourse), in the demonstration we will show only the segments that are related to general lexicon building (which includes a standard and terminological dictionary of the Croatian language) and will be connected with online repositories and encyclopedias. These program segments are searching for neologisms in documents (i.e. words that have not been marked in the general lexicon), generating grammatical forms of such words if they are changeable and saving them into a lexicon, adding semantic markups (like morphosyntactic) characteristics, and, finally, monitoring Croatian words in space and time.

Keywords: lexicon; Croatian Linguistics System; semantic markup; word evolution

1 Introduction

The Croatian Linguistic System is built by a group of linguistic and computer experts from the academic community that plan to implement it in the education system of the Republic of Croatia with the help of the CARNet (the Croatian Academic and Research Network). Hopefully, it will encourage students to get to know and respect the Croatian language. The subsystem for network dictionaries forming and processing is only a part of a general, integrated system that is shown in Figure 1. Users are marked with K1, K2, …Kn, each of them with their own web browser anywhere in the world. All of them can access the MG (Morphological Generator), the GL (General Lexicon of standard and professional terms) and the general corpus (standard and professional documents, syntax trees, ontologies etc.). According to the described retrieval, users cannot import or change anything in the shared storage. However, in the TEMP repository – user storage, every user can save or change anything that is offered (lexicon, syntax trees, ontologies etc.) and give others access to using their elements. The editorial board, which consists of the team of experts and administrators gathered around this project, decides whether the users' contribution will be publicly available. It should be noted, and as shown in Figure 1, that the link between each user and the TEMP repository is two-way, meaning that users can always use their data regardless of whether or not it is transferred to the 'common goods'.

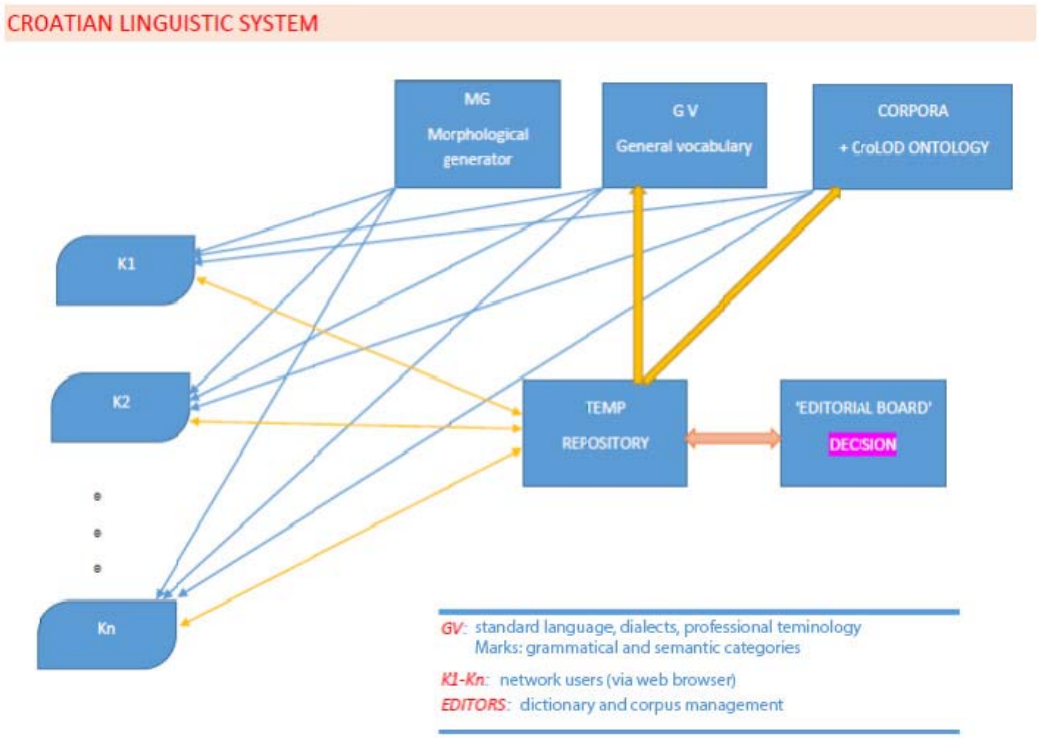

Figure 1: The Croatian Linguistic System.

2 Finding Neologisms

The general lexicon (GL) currently holds about a million morphosyntactically and (partly) semantically marked words from a private glossary.[1] New words, which are not in the database, are easily found by comparing with the GL from the user's text (corpus).

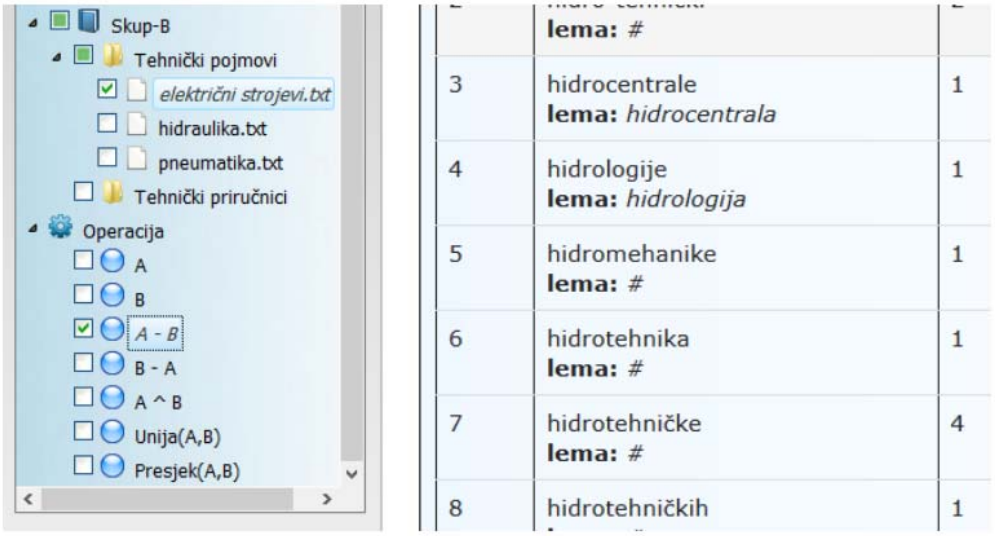

Figure 2: The vocabulary generator.

[1] Krešimir Pinjatela: "Hrvatska RIJEČ" database, Zadar, Croatia, 2001.

The program for the formation of the desired lexicons (vocabularies) from the documents stored in the network tree, with associated logic operations of such vocabularies for searching neologisms is available at: https://jt195996.pythonanywhere.com/Test_FancyTree/default/index.

3 The Morphological Generator

The morphological generator for the Croatian language ensures that all the grammatical forms of a word are found in the GL, which generates all possible word forms by applying the phonological changes and other morphological rules. The user only has to select the correct pattern, which is then (once and for all) stored in the database. In that way the GL is growing steadily (with continuous expert validation). The program can be accessed at the following address: https://jmarkucic.pythonanywhere.com/morf/default.

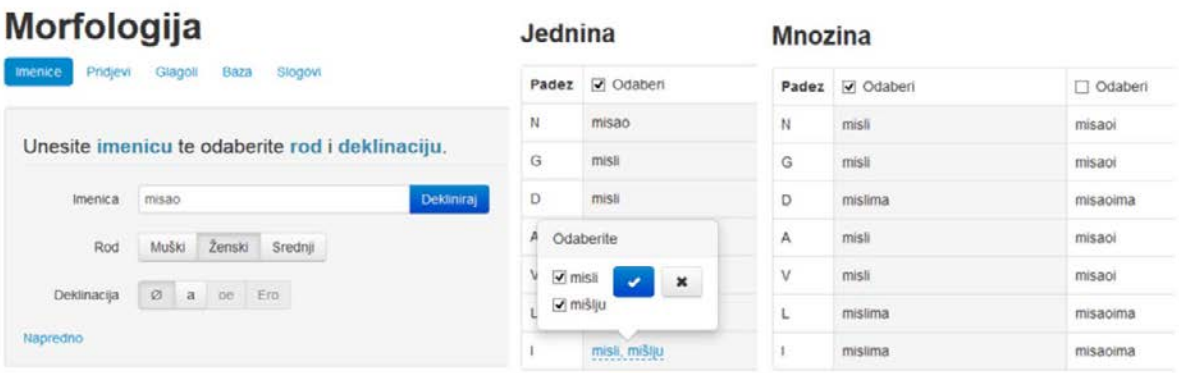

Figure 3: The Croatian morphological generator.

4 Semantic Word Markup

The morphological generator automatically assigns morphological characteristics to every word it generates. The semantic markup must be done by the user manually (and it is sufficient to do so only on the lemma of a word). The novelty of the approach is the T-structure tagging, which is in fact a hierarchy tree of different types of marks that are assigned to each word. The users have the ability to form their own T-structures or to use an already known scheme (e.g. semantic research by Jackendoff (2002), Pustejovsky (1998), Lieber (2009)). This program segment is available at: http://semantic.ene.li/#tstruct

5 Documents and Words in Space and Time

Finally, the Linguistic System has a module for storing documents and their diachronic monitoring in time and space. This enables historical tracking and evolution of Croatian words, as well as tracking of the violent (and often politically influenced) abolishment of words in order to unify close languages and nations.
This program segment is available at: http://bozoou.com/timeline/.

Figure 4. The module for space-time monitoring of the Croatian corpus and Croatian words.

6 Conclusion

This demonstration will show the basic programmatic modules which are useful in the formation and maintenance of lexicographic resources. It is a semi-automated extraction of words from the given documents in the corpus, their morphological analysis, semantic annotations, and finally their historical tracking in the time-space axis.

7 References

Aitchison, J., Gilchrist, A., Bawden, D. (1997) Thesaurus construction and use: a practical manual. London: Aslib.

Jackendoff, R. (2002). Foundations of language: brain, meaning, grammar, evolution. New York: Oxford University Press.

Lieber, R. (2009). Morphology and lexical semantics. Cambridge University Press.

Parker, P. M. (2008). Webster's Croatian-English thesaurus dictionary. San Diego, CA: ICON.

Pustejovsky, J. (1998). The generative lexicon. Cambridge, Massachusetts: MIT Press.

Vrgoč, D., Fink-Arsovski, Ž. (2008.) Hrvatsko-engleski frazeološki rječnik: kazalo engleskih i hrvatskih frazema = Croatian-English dictionary of idioms : index of English and Croatian idioms. Zagreb: Ljevak.

Acknowledgments: Part of the results presented in this paper was obtained through the HRZZ Research Project (under the UIP-11-2013 call) titled "Croatian Metaphor Repository" - sponsored by the Croatian Science Foundation.