The Network Integrator of Croatian Lexicographical Resources

Marko Orešković, Juraj Benić, Mario Essert
National and University Library in Zagreb,
Faculty of Mechanical Engineering and Naval Architecture

e-mail: moreskovic@nsk.hr, juraj.benic@gmail.com, messert@fsb.hr

Abstract

In this paper we describe an online application which connects various Croatian lexicographical sources (the standard dictionary, professional terminology, and encyclopedias) into a new network framework, which solves the problems of the currently available network solutions and offers new features. Although developed for the formation and processing of the professional terminology of technical sciences (primarily engineering), this framework is easily extensible to all other science areas (humanities, medicine, architecture, etc.). From a word formation standpoint, this framework provides an automatic generation of a temporary dictionary of a given text (entered or taken from the web), comparison and updating of a permanent dictionary with a temporary one (detection of neologisms), an automatic display (standard, professional and encyclopedic) of definitions of the words from the main dictionary or external (network) sources of a given text. From the point of view of processing, the framework enables a classic text analysis (frequency, concordances, statistics), converting the collected data (e.g. technical words) into the linked open data (LOD), and storing it into the Virtuoso triplestore repository. On this data, a variety of professional ontologies for viewing and searching can be developed.

Keywords: network framework; automatic dictionary generation; integration of lexicographic web information; linked open data

1 Introduction

An electronic dictionary is a dictionary in a digital form, and if it is also accessible over a network (the Internet), it is called an online dictionary. There are several types of dictionaries with regard to their contents (general, terminological, multilingual, etc.), the way they were established (alphabetic dictionary, frequency dictionary, inverse dictionary, conceptual dictionary, etc.) and their technical performance (local, multi-user, integrative). The simplest one is that in which the software solution only searches the database and gives a definition and related content based on the search query (Burkhanov 1998, Chambers 1995). There are two types of dictionaries, language dictionaries and subject dictionaries, such as encyclopedias and lexicons. Language dictionaries initially deal with language or lexical terms and all their linguistic characteristics.

A language dictionary often tends, even implicitly, to respond to the need of establishing lexical norms of a given language. That is the same task the professional or terminological dictionaries have to accomplish for every profession (Fribley 2012, SaintDizier, Viegas 1995).

The best-known online language dictionary in Croatia is the "*Croatian Language Portal*" (HJP) that can be accessed at http://hjp.znanje.hr/ Its dictionary consists of 116,516 main entries, out of which 67,049 are nouns, 15,699 are verbs, 20,154 are adjectives, 7,017 adverbs, 111 prepositions, 72 conjunctions, 152 numerals, 102 pronouns, 98 particles and 302 exclamations. Apart from the main

entries and their definitions, the database also contains examples (~ 60,000), syntactic phrases (~ 18,000) and phrasal expressions (~ 10,000). The origin of a word is explained in the etymology zone and the origins of personal names and surnames (anthroponyms) and geographical names (toponyms) are explained in the onomastics zone. Therefore, each dictionary article is divided into six modules (basic grammatical information, definitions, collocations, phraseology, onomastics and etymology) which can be displayed or hidden by a menu on the left side of the screen (Anić, Jojić, Matasović 2003; Vrgoč, Fink-Arsovski 2008).

The best-known terminology repository in Croatia is the "*Struna",* which can be accessed at: http://struna.ihjj.hr. The building of such an infrastructure for the Croatian language is very important, especially since it has become an official language of the European Union. *Struna,* whose holder is the Croatian Institute for Language and Linguistics (IHJJ), currently contains about twenty different professions and professionals from these areas. It is a terminology database of the Croatian professional terminology which systematically collects, generates, processes and interprets the terminology of various professions (about 40,000 processed terms so far) to bring together and harmonize the terminology in the Croatian language. The project of the construction of the Croatian professional terminology was initiated by the Council for Standard Croatian Language Norm, and the project leader is the Institute of Croatian Language and Linguistics. This program builds the terminology infrastructure that almost all European countries already have.

Croatia has several encyclopedic portals whose holder is the "*Miroslav Krleža Institute of Lexicography"* (LZMK). The most famous among them is the "Croatian encyclopedia" network, http://www.enciklopedija.hr/, which collects information from various smaller network lexicons (e.g. Croatian family lexicon http://hol.lzmk.hr/, Movie lexicon, http://film.lzmk.hr/, Medical lexicon http://medicinski.lzmk.hr/, Istrian, Football, Biographical, and similar lexicons). All these network encyclopedias work on the same principle (Parker 2008): there is the form for the search query, which (if found) returns a description, content and explanation of a term with the field/area in which that terms appears (e.g. History, Literature, Sociology and the like). The network of Croatian Encyclopedia is based on the printed edition, which was published in 11 volumes from 1999 to 2009. In that edition there are 67,077 articles published on a total of 9,272 pages of a large encyclopedic format (with a total of 1,059,000 rows), and with 17,000 black-and-white additions and 504 pages of colour additions. That encyclopedia is the result of the work of about 1100 authors.

2 The Problems of Network Lexicographical Applications

These are the most important representatives of network repositories in Croatia, but, besides their significant popularity among many users, they also have certain common and individual limitations and problems. For example, users often cannot find a definition (in the *Struna* or *LZMK*) if they do not know the key word in advance (if they know it, then they mostly do not need the definition). This problem can partly be solved if the user guesses the first few letters of the word, and then waits for an auto-completer to offer a list of words starting with those letters. Whether the requested word will be found in the list or not, it is hard to say (it may start with some other initial letter). In the *Struna,* there is an indication that along with the words, their synonymic words or equivalents should be shown, but that is far from a real, high-quality implementation of what can already be seen in the WordNet (CroWN – the Croatian version of the Princeton WordNet is also far from actual usability). Although in the *Struna* and some *LZMK* encyclopedias it is provided that each word has its linguistic reference, in most cases it is absent, i.e. the basic grammatical types of professional / encyclopedic words are

unknown (noun, adjective, verb), just as their corresponding grammatical forms are not known (words made by morphological changes – declension or conjugation). This means that the information in the *Struna* cannot be used for the identification of professional terminology in professional texts, because in such texts, words appear in different grammatical forms. All other, more necessary, needs are linked to this deficiency (like the quantitative analysis of specialized texts, for example, the frequency of detection of terminology, collocations, usage of concordance etc.).
The *HJP* has the grammatical description of parts of speech, and their morphological forms are displayed on a separate page. However, although every natural language is dynamic and changeable, this portal is static and does not follow any changes or updates. It is similar to the *LZMK* encyclopedic portal. A major setback of these repositories is that there is no connection between the words in the definitions with words (entries) in the database. If the definition holds words that are already in the database, the user cannot see them. The only way to find them is to try to type in the search form. A common problem of a specialized vocabulary is that a definition can also be as incomprehensible as the term, and requires a number of iterations until the true meaning of the concept is revealed. Finally, constructing a centralized collection of information without a proper online access to enter and update this information and to enable parallel operations for numerous institutes, educational institutions and the academic community is a large and hard work – especially for the people in the *IHJJ* and *LZMK* who manage the database – and, on the other hand, there is the continuous growth of neologisms that cannot be properly solved without a distributed and parallelized approach. A common feature of this way of working is that repositories are hardly adaptable to rapid changes. Based on everything observed, the need appeared for considering new approaches and building an online solution which will successfully answer all the mentioned deficiencies. Moreover, the solution should give the user an opportunity to see the existing online information accessible for all the words in a text. It does not matter whether these words are from a standard dictionary, a terminology repository or an encyclopedia. Today's network technology makes such an opportunity possible.

3 From Conception to Realization

The basic idea was to build such a network framework where any text would be able to be uploaded (mainly by a professional), and processed in different ways:

1 An automatic creation of a dictionary of all the words it contains.

2 Results that should be interpretable according to the alphabet, frequency of occurrence, and parts of speech.

3 Allowing the user to get information about the words already known from the text (like term definitions) stored in the integrated dictionary by giving them a link to an explanation on a network destination.

4 Allowing an expert (registered / logged user) to tag unknown words and link them with the already known information from a lexicon or a network repository.

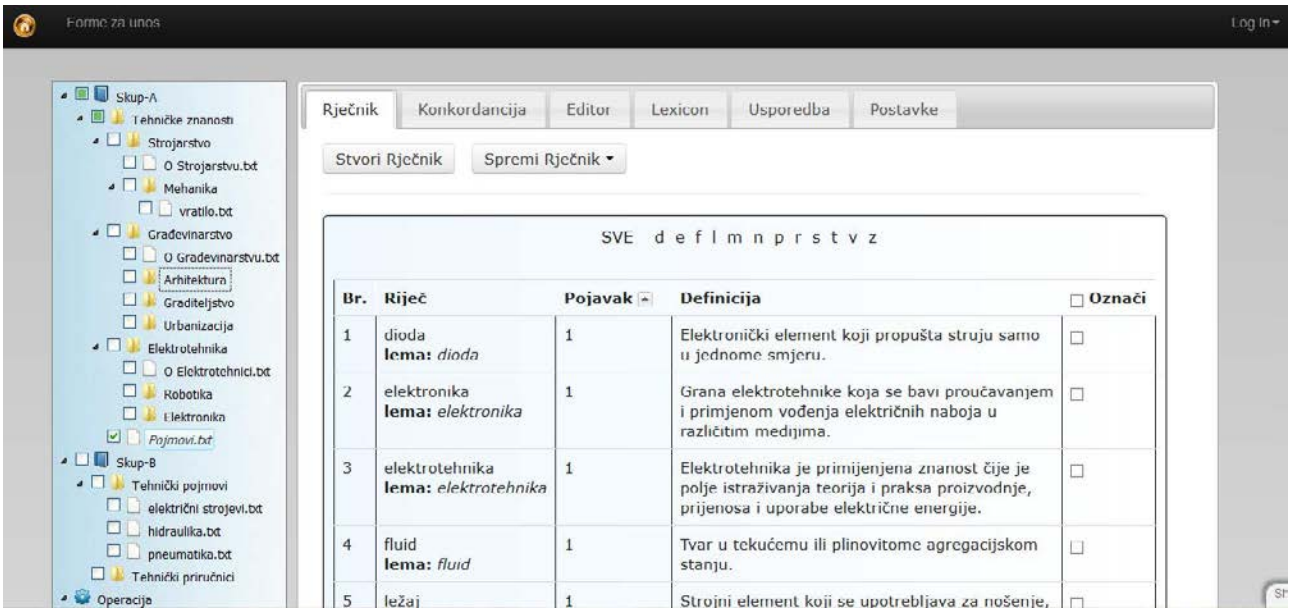

Figure 1: The network framework – realization.

This idea in its initial phase was realized as[1] a students' paper, which in 2015 received the Rector's award of the University of Zagreb. A computer program has been made in *web2py* technology in which the Python applications[2] are connected with HTML5/Bootstrap framework and LOD/ Virtuoso server technology (each data element is stored in triplestore). The program was later expanded (because of the many inflectional forms in the Croatian language, and the need for generating all grammatical forms of a word) with the program "*Morphology*" (https://jmarkucic.pythonanywhere.com/morf/default/imenice).[3] The network integrator of Croatian lexicographical resources was achieved with the program code of more than 3,500 lines. It is accessible at: https://jt195996.pythonanywhere.com/Test_FancyTree/default/index Figure 2 shows the scheme of the algorithm which links the desired text (from which the dictionary was made in Figure 1) with the information stored in the integrated dictionaries, and a link to the extended information on network destinations. Above every word that has stored information, the network framework displays numbers (e.g. [1] for HJP, [2] for Struna, [3] for LZMK). By clicking on them, a definition and a source link are shown. Such an integration of sources resulted in a significant increase in the organization of information and the speed of its retrieval (Geeraerts, 2010).

[1] J. Benić, J. Topić: "Mrežni program za tvorbu i obradbu tehničkih rječnika".
[2] The Python programs include the well known NLTK module (http://www.nltk.org/).
[3] Joško Markučič: "*Mrežni morfologijski program za hrvatski jezik*".

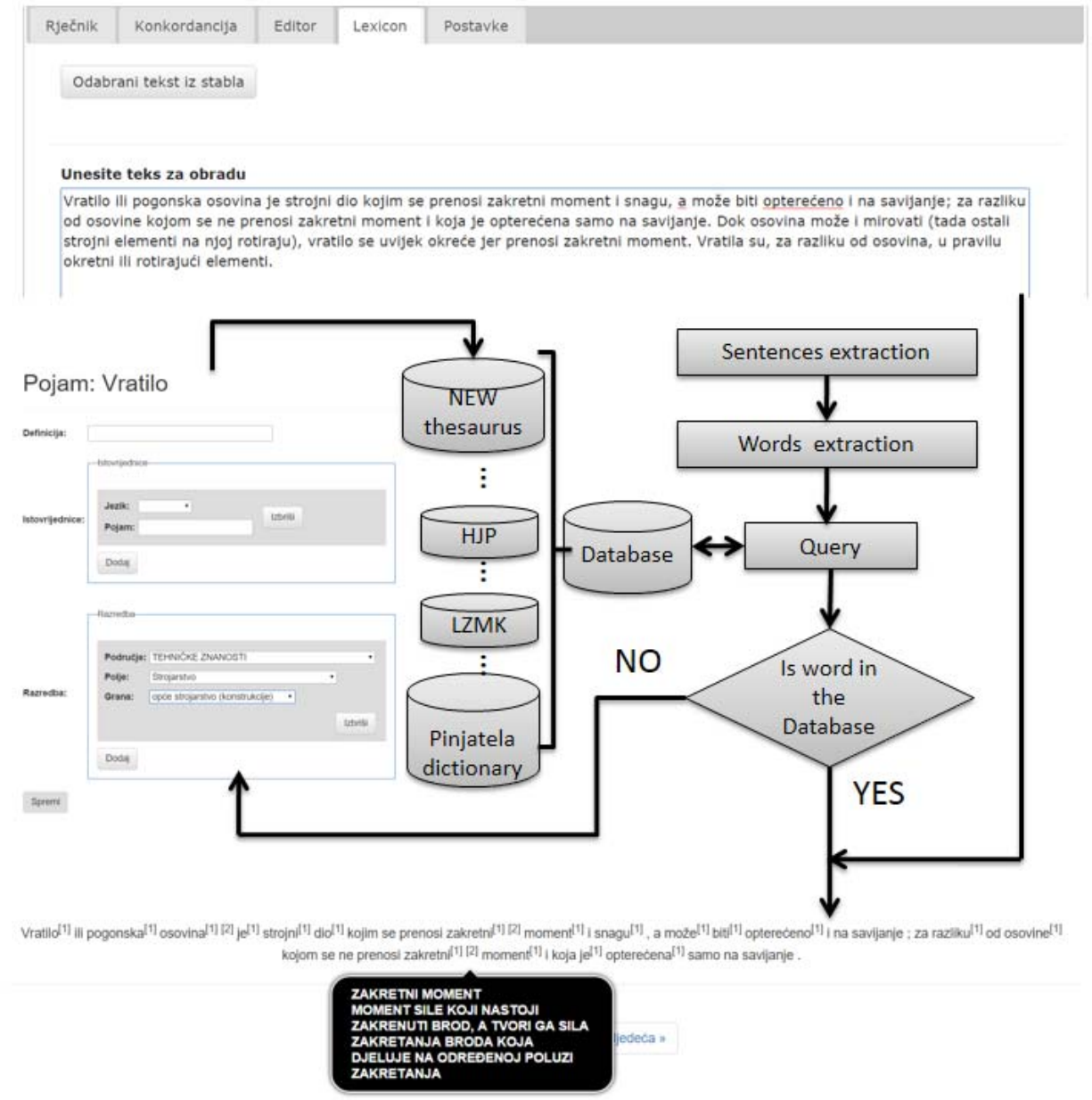

Figure 2: Algorithm scheme.

4 Conclusion

In this paper we gave a review of existing network lexical resources in Croatia (the standard language dictionary, terminology repository and network encyclopedias), their strengths and limitations. Then we explained the algorithm on which we based our network framework which links and integrates all of these resources and eliminates many restrictions. The program allows the user to load any text, and it is broken down into sentences and words with lemmas, which are then linked with online resources and automatically displayed with definitions. This allows the user to analyze the text, sentence by sentence, and to have a lexicographical definition for each word if it is already processed in some of the network repositories. The user does not need to waste time searching for words which could not be found by the program because they are not processed. In this way, the texts and lexicons are firmly linked, the processed words and their definitions are easily fetched, the new ones are easily

recognized and, if necessary, processed and permanently stored. The integrator dynamically updates its lexicon with new words from both the terminology and the standard language corpora.

5 References

Anić, V., Jojić, L., Matasović, R. (2003). *Hrvatski enciklopedijski rječnik*. Zagreb: Novi liber.

Burkhanov, I. Y. (1998). *Lexicography: a dictionary of basic terminology*. Rzeszów: Wydawn. Wyższej Szkoły Pedagogicznej w Rzeszowie.

Chambers, R. J. (1995). *An accounting thesaurus: 500 years of accounting*. Oxford, OX, UK; Tarrytown, N.Y., USA: Pergamon Press.

Fribley, K. (2012). *Find the right words with thesauruses*. Ann Arbor, Mich.: Cherry Lake Pub.

Geeraerts, D. (2010). *Theories of lexical semantics*. Oxford; New York: Oxford University Press.

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

Saint-Dizier, P., Viegas, E. (1995). *Computational lexical semantics*. New York: Cambridge University 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.