The Development of a Network Thesaurus with Morpho-semantic Word Markups

Marko Orešković, Mirko Čubrilo, Mario Essert
National and University Library in Zagreb,
Faculty of Mechanical Engineering and Naval Architecture
e-mail: moreskovic@nsk.hr, mcubrilo@foi.hr, messert@fsb.hr

Abstract

This paper presents a part of the network frame of Croatian linguistics which focuses on a new kind of thesaurus, based on morpho-semantic features of words. Instead of the classic (e.g. MULTEXT-EAST) POS tagging of words for grammatical and some semantic categories (e.g. animate), in this paper every word has its hierarchical T-structure which can hold various data types in its branches (string, integer, link, word list, ordered word list etc.), and in that way words and their various occurrence possibilities in a text can be described even better. Moreover, the known WordNet or other semantic structures (e.g. the Croatian Language Portal, terminology repository or a network encyclopedia) can be presented as T-structure nodes in the same way. During this process each word in the definition of an entry is linked to a lexicon, which results in increasing the semantic connectivity of words by at least one order of magnitude (about ten times more semantic relations). Searching through and browsing such a network dictionary brings a new dimension, and words in the dictionary, beside the paradigmatic, also possess all the syntagmatic properties, because the computer processes their appearance in any utterance or sentence as a series of connected nodes (LOD objects). This provides the possibility of storing all data in triplestore (e.g. on the Virtuoso server).
Keywords: lexical markup framework; morpho-semantic tree-structures; semantic networks

1 Introduction

Despite the favourable reviews, for instance that the "Croatian Wordnet will at the same time be a thesaurus, a dictionary of synonyms and a valency lexicon of Croatian verbs in digital form"[1], together with a significant number of contributors (Raffaelli 2012, Šojat 2009) from 2007 until today, the Croatian Wordnet (also known as CroWN) is not included in other WordNet development programs (EuroNet, BalkaNet) and is not connected to any other languages like many others are (e.g. Slovenian Wordnet).

The main problem of this type of translation and concordance with the PWN was in choosing *VisiDic* as the software editor, since it uses its own indexes, which made linking it to other languages problematic. This is probably the reason why the CroWN is not a part of any other publicly available web-portal with international WordNets. Our ambition is to implement it taking a different approach: instead of making the Croatian Wordnet a thesaurus, we have laid the foundations for and started developing a network framework CLW (Croatian Linguistic Web), which will include the CroWN as only one among many other linguistic attributes of Croatian words from a given corpus. The CLW links modules for the Croatian word formation processes (Morphology portal[2]), the HJP linguistic

[1] http://hnk.ffzg.hr/rmjt/p3.html
[2] https://jmarkucic.pythonanywhere.com/morf/default/

portal[3] and many other lexical encyclopedias of the Miroslav Krleža Institute of Lexicography [4]. The central part of the CLW is a thesaurus with morpho-syntactic word markups, which is intended to be constantly improved. It is expected to process both the paradigmatic and syntagmatic characteristics of the language.

2 Related Works

Although many lexical frameworks have been developed since the early 1980's (e.g. Acquilex, Multilex, Genelex, Eagles, Isle, Mile and others), nowadays everybody tends to raise the standard, known as the LMF – Lexical Markup Framework (Francopoulo 2013). The first step in developing the LMF was to design an overall framework based on the general features of existing lexicons and to develop a consistent terminology to describe the components of those lexicons. The standards are fundamental to exchange, preserve, maintain and integrate the data and language resources (LRs), to achieve interoperability in general, and they are an essential foundation for any LR infrastructure. The CLW implements the LMF infrastructure to achieve linkage between Croatian and other language WordNets to be included in a global grid.[5] However, the CLW developed its own annotation structure, which replaces the standard morphological tags, and at the same time enables the definition of the semantics tags for test systems built by different authors (e.g. Szymanek, Jackendoff, Pustejovsky). As we know, Ray Jackendoff has developed a decomposition system of semantic representation or Lexical Conceptual Structure (LCSs), as he calls it, which stands for hierarchical arrangements of functions and arguments (Jackendoff 2002). The primitives of the system are semantic functions and smaller atoms of meaning represented as features (e.g. [bounded], [internal structure]) which allow for the discussion of aspectual characteristics of verbs and quantificational characteristics of nouns (Lieber 2009). Similarly, the decomposition framework of semantic description that has been developed in the work of Anna Wierzbicka is also admirably comprehensive and, unlike Jackendoff's, it is broadly cross-categorial. She set the number of indefinable primitives at fifty-six (until today). Rochelle Lieber relies on the work of Jackendoff and Szymanek and defines his own basic categories for derivational affixes. What needs to be pointed out is that within the CLW framework, it is possible to use and test any of the semantic categories or to build one's own through the Tree or T-structures, which are similar to the WordNet hierarchical structures and are used for morphological and semantical word markup.

3 The Morpho-syntactic Markup

A morphological tag is a symbol encoding (morphological) properties of a word. The size of a tagset depends on a particular application as well as on language properties: for inflectional languages it is necessarily large. For example: the Penn tagset (without any formal internal structure) for American English: 36 tags; The Lancaster-Oslo-Bergen Corpus: 132 tags; the Czech structural and positional tagset: about 4,000 tags. The Czech tagset mixes the morpho-syntactic annotation, and it combines several morphological categories into one, which explains the huge amount of tags. A natural way of making tags manageable is to use a structured system where a tag is a composition of tags, each coming from a much smaller and simpler atomic tagset tagging a particular morpho-syntactic

[3] http://hjp.znanje.hr/
[4] http://enciklopedija.lzmk.hr/
[5] http://globalwordnet.org/global-wordnet-grid/

property (e.g. gender or tense). In that way MULTEXT-East Tagset V.4[6] was made, which includes 13 languages: English, Romanian, Russian, Czech, Slovene, Resian, Croatian, Serbian, Macedonian, Bulgarian, Persian, Estonian, Hungarian. Harmonized tagsets make it easier to develop multilingual applications or to evaluate the language technology tools across several languages, which is interesting from the perspective of linguistic typology as well, because the standardized tagsets allow for a quick and efficient comparison of language properties. However, that approach also has serious problems, especially when it is used on a corpus which represents a mixture of multiple languages (e.g. *SETimes* articles) or the *Apertium*[7] lexicon, which is not Croatian. Various grammatical categories and their values might have different interpretations in different languages. For example, definiteness is expressed differently in various languages: determiners in English, clitics in Romanian; only pronominal adjectives in Lithuanian, adjectives in Croatian etc.
Further development of these tagsets[8]does not contribute to the conservation or to the development of a language. In the same MULTEXT-East tagset, the morphological and semantic properties are mixed. That is the main reason why it seemed necessary to introduce the T-structures that would be able to make a distinction between the simple atomic subsets of any morphological category marks, as well as semantics. It is a realization of what Rochelle Lieber calls "an anatomical metaphor", while for James Pustejovsky it is "Qualia structure" (Pustejovsky 1998). A good side of this implementation is that for every annotation of corpus there is an easy way to use a self-developed annotation system or one developed by other authors.

4 The T-structures

The T-structure is a recursive tree-based structure which can contain morphological attributes or syntactic categories of a word in any language. One of these structures is the commonly known POS, which is not the same for every language, so it is possible to define it like any other grammatical or syntactic subcategory. This enables defining the subcategory "number" in the Russian language that will have only "plural" and in Croatian and Slovenian also "dual", and the sub-branches *pluralia tantum* (trousers, scissors) and *singularia tantum* (fruit, bread, milk) for some nouns. Apart from that, words from the same POS category (e.g. category "number") can have the subcategories of gender and case, and then, in their own languages (e.g. Croatian), we call them numerical nouns or adjectives or even numeral adverbs, instead of having a T-structure number tag in POS and a T-structure accompanying tag in GENDER (male, female, neuter) or any other T-structure. This provides the maximal flexibility of word tagging and the finest granulation of properties of any language and its peculiarities.

[6] Erjavec 2010. http://nl.ijs.si/ME/
[7] http://wiki.apertium.org/
[8] Ljubešić 2013. http://nlp.ffzg.hr/data/tagging/msd-hr.html

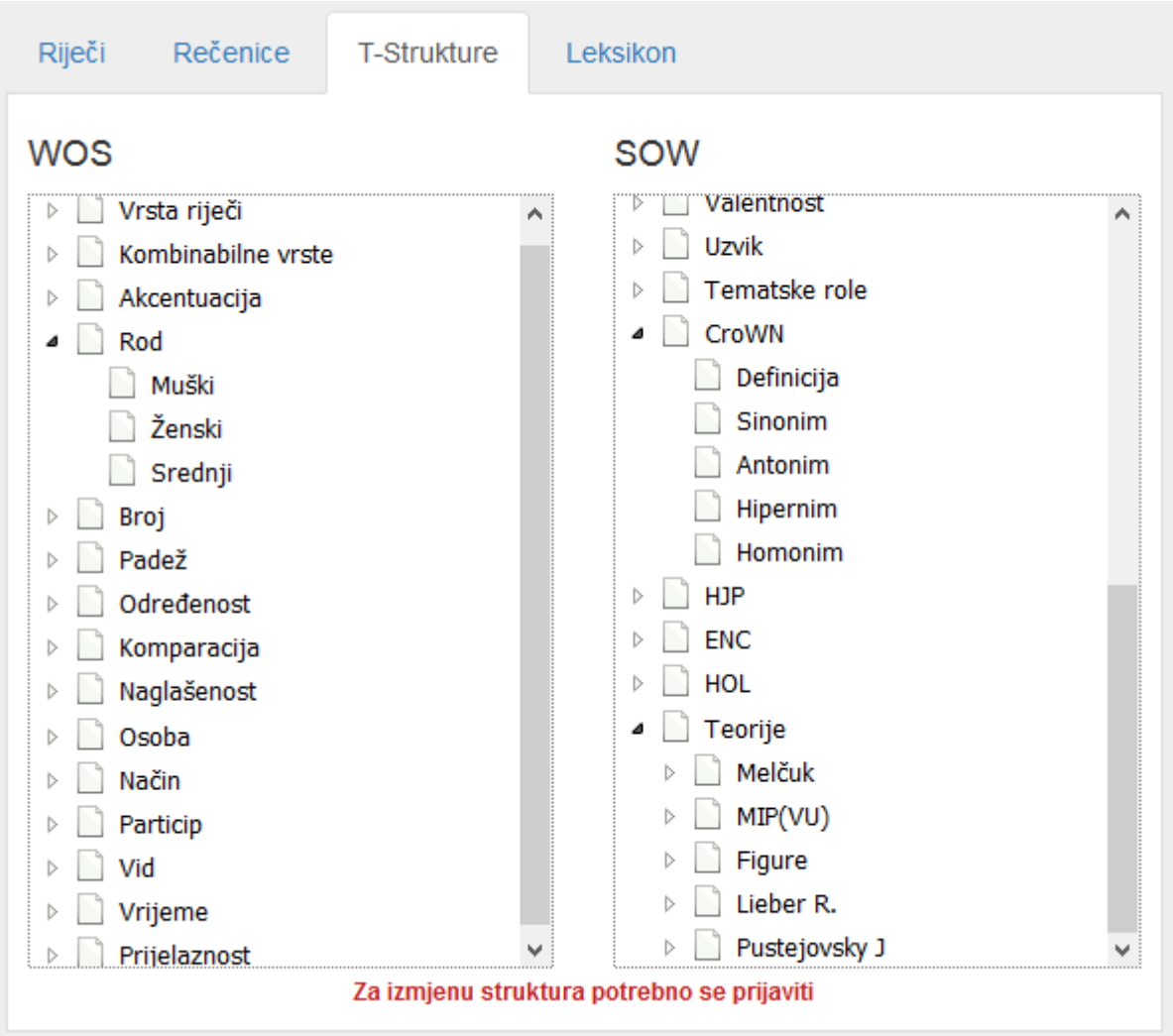

Figure 1: The T-Structures for the Croatian language.

Filling the T-structures with morphological attributes can be done manually, but also automatically (via API). In our case, automatic filling was done by transforming tags from the *Pinjatela* (private) glossary[9], which contains about 100,000 Croatian words with the accompanying lemmas from which all other grammatical forms are developed by the generator of the Morphology portal. Of course, the morphological attributes and their associated structures, which are opposed to SOW (semantics of words), due to the classical understanding of language classified as WOS (words of sentence), hold the information about the semantic categories. These categories can include the meaning of morphemes and the way they are combined to form meanings of complex words, including derived words, compounds, and words formed by conversion. Although it is still not implemented, it will be possible to use Melčuk's collocation functions in the structures (Gelbukh, Kolesnikova 2014), such as the already built-in structures like the CroWN (or any other Wordnet), the HJP (Croatian linguistic portal) or the LZMK encyclopedia. The user assigns semantic attributes to a word in a corpus from the SOW category in the same way as in the WOS. In the CroWN T-structure case, it means that for every word in a sentence we choose the semantic property from many other definitions listed inside the structure.

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

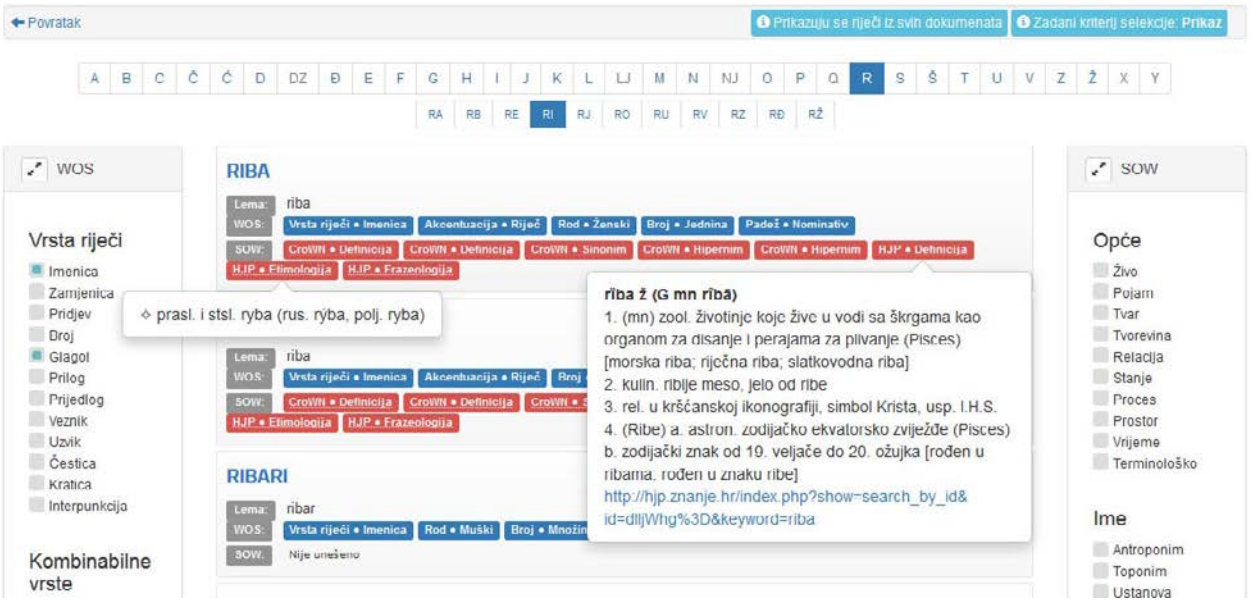

Figure 2: Thesaurus.

As shown in Figure 2, word filtering in the thesaurus is done by letters, their pairs and any other attribute from the WOS/SOW categories. Among the SOW categories, there are also the CroWN (with attributes like definition, synonyms, antonyms, hypernyms etc.), the HJP linguistic portal (with its own attributes, for example phraseology, that correspond with the WordNet definitions), the lexicographic data from the LZMK encyclopedias, and there is also a network version of the printed synonyms dictionary in preparation (Šarić, Wittschen 2008). The network thesaurus has no limit for any further extensions and is not connected with language, but is conceptually designed for the language with any corpus and marks in the native language and in English. The system includes different user levels, from administrator to user groups and regular view-only users.

5 A Good Foundation for Future Work

T-structures could be built for every language separately, and in further research it will be possible to join the T-structures from one language to another through sentence patterns, which will pave the way for automated translation. Thanks to the described approach, the words in sentences do not carry only grammatical, but also semantic features (any of them). It is important to note that one and the same word that occurs multiple times in the same sentence does not have to have the same meaning. For example, in Croatian the word *'put'* can have two meanings: 'path', and 'skin tone', and every occurrence of such words in the global dictionary has been specially tagged with the word's own semantic categories or different semantic attributes from the same category. Such solutions, from the programmers' point of view, were a real challenge, equal to linking all the relevant language resources on the Internet which are publicly accessible and joined to form a whole. This is, among other things, related to building links in definitions (glosses) to other words in the dictionary, and creating domains / sources of the words in an automatic way. The direct application of this approach is a utilization of such word attributes as groups of domains or co-domains of language collocation functions. The thing missing in the WordNet structure (Fontenelle 2012) is found here, the functions are not related to one, but several variables or even their groups. For example, in the classical

interpretation, “young girl” ('*djevojka*’ in Croatian) could be the antonym for “young man” ('*mladić*' in Croatian), but the word “old lady” (*‘starica’* in Croatian) could also be used as the antonym for the same word. Instead of *'sex*', we take ‘*age*’ as the criterion for expressing oppositeness, which creates a semantic ambiguity in their record. The CLW system will enable an easy way of defining a word in the database with the function ‘*antonym()*’ like ‘*antonym(young_girl, sex)=young_man*’ (Croatian *antonym(djevojka, spol)=mladić*) or ‘*antonym(young_girl, age)=old_lady*’ (Croatian *antonim(djevojka, starost)=starica*). Besides the T-Structures, the F-structures are also implemented, which, according to the MT theory of I. Melčuk, enables the creation of a set of rules for the translation of word groups. Each F-structure is defined by its name and the accompanying input arguments. The mapping of words is based on the given attributes. The Figure 3 shows how F-structures are stored in the database.

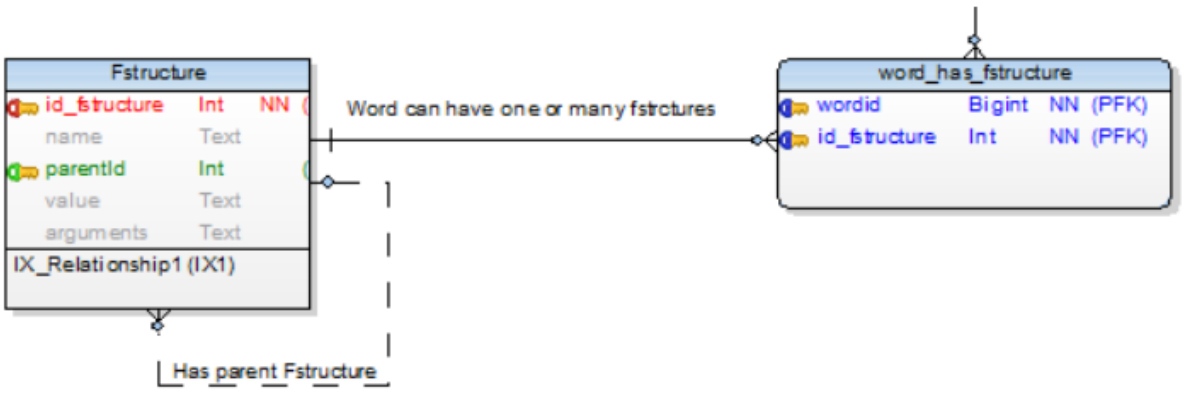

Figure 3: The F-Structures ERA model.

6 Conclusion

The thesaurus (Figure 4) is conceived and designed as a part of the computer linguistic system (CLW) that includes all the important components of linguistics: generative morphology, corpus, syntax, and semantic marking, searching and formation of patterns, generation of lexicons and other. The system has a multi-user interface with administrative, group and user privileges, which enables collaborative work on shared linguistics projects over the Internet. As a semantic hierarchical structure, the WordNet has influenced the building of new data structures that we call T-structures for the morpho-syntactic and semantic markup of words inside the thesaurus framework. In semantic terms, the T-structure includes vertical (paradigmatic) components (WordNet, linguistic portals, encyclopedias etc.), as well as horizontal (syntagmatic) values for the future building of collaborative databases. In that way all definitions (glosses) of words in the WordNet (or any other encyclopedia linked to framework) get their link property, which results in a significant expansion of a number of linked semantic nodes.

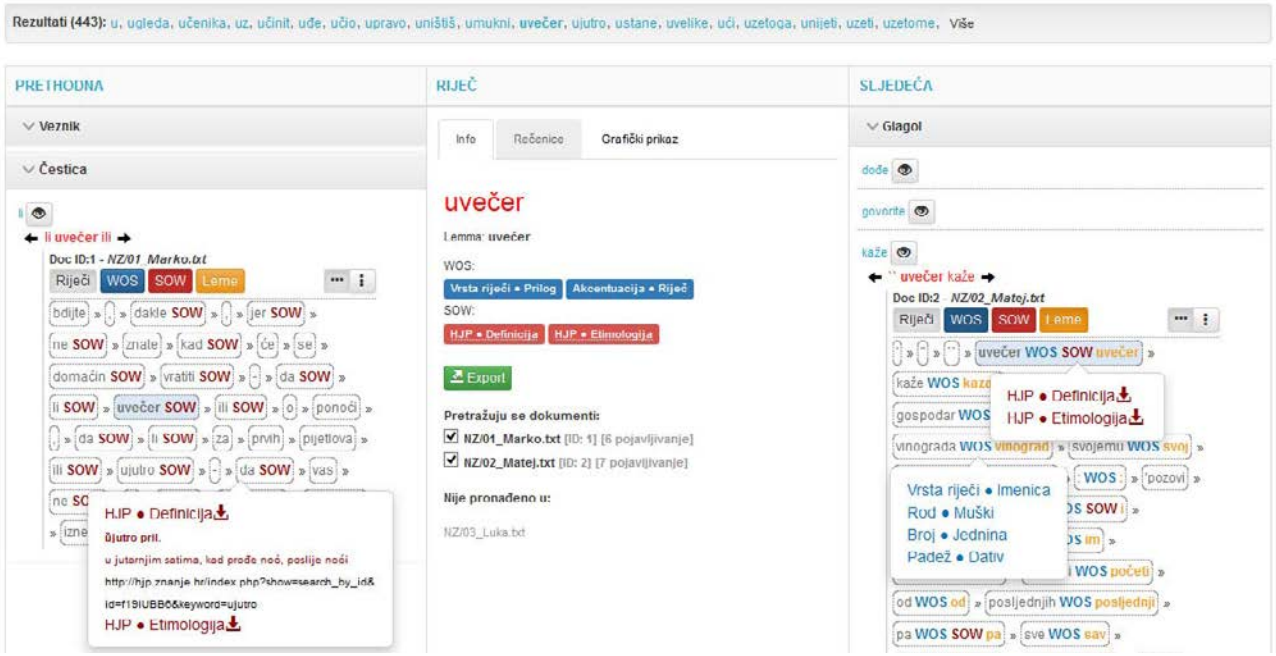

Figure 4: The CLW network thesaurus.

7 References

Fontenelle, T. (2012). Wordnet, Framenet and Other Semantic Networks in the International Journal of Lexicography – the Net Result? *International Journal of Lexicography*, Vol. 25, pp. 437-449.

Francopoulo, G. (2013). *LMF Lexical Markup Framework*, John Wiley & Sons, ISTE.

Gelbukh, A., Kolesnikova, O. (2014). *Semantic analysis of verbal collocations with lexical functions*. Berlin Heidelberg: Springer-Verlag.

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.

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

Raffaelli, I., Katunar, D., (2012). Lexical-Semantic Structures in Croatian WordNet. *Filologija*, Vol. No. 59.

Šarić, Lj. / Wittschen,W. (2008). *Rječnik sinonima hrvatskoga jezika*. Čakovec: Naklada Jesenski i Turk.

Šojat, K. (2009). Morphosyntactic annotation in the Croatian Wordnet. *Suvremena lingvistika*. Vol. 35, No. 68.

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.