Describing polysemy of Croatian language more precisely

MAJA ŽIVKO
Department of Information and Communication Sciences
Faculty of Humanities and Social Sciences
Ivana Lucića 3
CROATIA
mazivko@ffzg.hr

Abstract: Polysemy is the coexistence of several meanings in one word. Some words develop many meanings and their each new meaning can often form another definitions. Dictionary entries have numbered or lettered sub-parts, and these are informing the user of meaning differences. Unfortunately dictionaries can not describe polysemy entirely. If we look at the web corpus we can see that in corpus we find much more sense distinctions than in electronic or printed dictionaries. This calls for finding the way of getting right division of word meanings. Even for lexicographers there are no objective criteria for systematically extracting from corpus data the kinds of information useful to dictionary users. FrameNet and Mindnet are the projects that find a way to solve this problem. The aim of this paper is to accent that there is a need for constructing this project also for Croatian language.

Key-Words: Polysemy, FrameNet, Mindnet, semantic network, web corpus

1 Introduction
Corpus data can be a source for describing polysemy words in more precisely way than dictionaries. But the problem is that even extracting data from web corpora cannot give a full information about polysemy. The aim of this paper is to show that there is a need of constructing a natural language understanding system that can derive understandings from text samples based on the occurrence of a particular word in a particular context. Framenet [4] and Mindnet [5] are projects devoted to this problem, but there is a need for realisation of this kind of projects in Croatian language also.

2 Problem Formulation
In this paper I will compare polysemy of adjective lonely in Croatian and English language to show how information given in Croatian language portal [1] are not enough for understanding this language for foreigners. Source for my research was Croatian Web Corpora [2] and English web Corpora. [3] I will show here only examples which deffers from Croatian to English language.

1...u slučaju da ostane politički usamljen... 2. usamljen u ostvarivanju dijaloga u regiji .koji je usamljen branio prostor gola ......3. Napadačima je lakše dolaziti u prilike kad nisu usamljeni u 16-tercu. 4. Rood van Nisterlooy trebao bi istražiti usamljen u napadu... 5. U napadu je bio usamljen. 6. lopta je preskočila tri braniča Slavena i usamljenom Rushfeldtu nije bilo teško zabiti...7. istina i da su gotovo svaku pucali potpuno usamljeni..8. Varteks je izbio na peto mjesto, a Zadar je usamljen na drugu 9. no Dinamo je i dalje usamljen na vrhu s 12 bodova više od Šahmara. 10. Dejan Borovnjak je s 32 koša bio usamljen kod gostiju..11. gražanje javnosti dokaz da je Kežman samo usamljeni reprezentant starih običaja...ima usamljen život.12. sada usamljen na vrhu ljubavi strijelaca. …13. hoće li ovakvo ubojstvo ostati usamljen slučaj...14. u tome su među poznatima u Hrvatskoj usamljeni...15. Modrulj zna napasti ljude, pogotovo usamljene manje čamce ili ronioce..16. S novom pobjedom je Svindal ostao usamljen na vrhu poretka s 285 bodova..19.WW više nije usamljen s ponudom takvih mjenjača..20. modul zna napasti ljude, pogotovo usamljene manje čamce ili ronioce. 21. Dimitrije Pejanović nije usamljen u ovim razmišljanjima..22...pa je Agüero ostao usamljen u vrhu napada u borbi s Portovim braničima...23. no bio je usamljen u svom trudu..24. je ponovio da će Slovenija ustrajati u blokadi Hrvatske te da u tome neće biti usamljena , prenose slovenski mediji..25. ...takav usamljen u vrhu Dinamova napada..26. Nije Robertino usamljen slučaj neopraviteljne boravka izan poznatima u Hrvatskoj usamljeni...15. Modrulj zna napasti ljude, pogotovo usamljene manje čamce ili ronioce..22. Dimitrije Pejanović nije usamljen u ovim razmišljanjima..22..pa je Agüero ostao usamljen u vrhu napada u borbi s Portovim braničima...23. no bio je usamljen u svom trudu..24. je ponovio da će Slovenija ustrajati u blokadi Hrvatske te da u tome neće biti usamljena , prenose slovenski mediji..25. ...takav usamljen u vrhu Dinamova napada..26. Nije Robertino usamljen slučaj neopraviteljne boravka izan poznatima u Hrvatskoj usamljeni...
zatvorskih rešetaka...27. Tako je Klasnić često usamljen u napadu ili sjedi na klupi...28. A ni Massachusetts nije bio usamljen u istjerivanju vraga ognjem i sumporom...29. U klasi 4 najbrž je bio Roberto Rauš, a usamljen u klasi 3 bio je Božo Prnić.

The Croatian examples that I showed here can not be translated into English language using adjective lonely, e.g. we cannot say 1. he becomes politically lonely or 5. he was lonely in fighting 14. they were...lonely example or 15. it looks like it was a lonely anger...

In Anić’s dictionary adjective lonely is described like this: 1. separated from other 2. who is alone, without anybody. This definition is correct but here are not given all information which are important for translating one language into another.

3 Problem Solution
We can see, that there is a real need of decribing polysemy words of Croatian language more precisely. This problem can be solved by constructing FrameNet [4] or Mindnet [5] projects for Croatian language.

3.1. FrameNet
The FrameNet project is building a lexical database of English based on annotating examples of how words are used in actual texts. It provides a unique training dataset for semantic role labeling, used in applications such as information extraction, machine translation, event recognition, sentiment analysis, etc. [4] FrameNet is based on a theory of meaning called Frame Semantics. It relates linguistic semantics to encyclopaedic knowledge. The basic idea of Frame Semantics is that one cannot understand the meaning of a single word without access to all the essential knowledge that relates to that word [6] e.g. the person can not understand the word cure without knowing anything about the situation of curing, which also involves a doctor, patient, medicaments, hospital and relationships between curing and doctor, doctor and patient and so on...Word activates a frame of semantic knowledge relating to the specific concept it refers. .A semantic frame is a collection of facts that specify characteristic features, attributes, and functions of a denotatum, and its characteristic interactions with things necessarily or typically associated with it.[6] A semantic frame can also be defined as a coherent structure of related concepts that are related such that without knowledge of all of them, one does not have complete knowledge of any one; [6]Frames are based on recurring experiences. So the curing frame is based on recurring experiences of curing. Words specify a certain perspective from which the frame is viewed.[6] For example cure views the situation from the perspective of the doctor and sickness from the perspective of the patient. as the main semantic principle.

Fillmore [7] says that the classification of frame types can be used in analyzing the vocabulary of a language. “This analysis can begin at the one end with lexical items that have relatively simple word-to word mappings, such as the names of colors or the names of natural kind, and can continue on to the highly elaborated conceptual frameworks that preposse subtle sorts of knowledge about the intellectual and institutional use”.

The basic idea of FrameNet is that the meanings of most words can best be understood on the basis of a semantic frame. In the FrameNet project there are frame lements and lexical units. Here is the example when lexical unit is noun [4]:

- [Punishment This attack was conducted] [Support in] RETALIATION [Injury for the U.S. bombing raid on Tripoli... or adjectives such as asleep in the Sleep frame:

- [Sleeper They] [Copula were] ASLEEP [Duration for hours]

3.2. Mindnet
MindNet is the product of a large collaborative effort within the NLP group in Microsoft Research. The automatic extraction of semantic relations (or semrels) from a definition or example sentence for MindNet produces a hierarchical structure of these relations, representing the entire definition or sentence from which they came. Such structures are
stored in their entirety in MindNet. About 25 semantic relation types are currently identified during parsing and LF construction, including **Hypernym**, **Logical Subject**, **Logical Object**, **Synonym**, **Goal**, **Source**, **Attribute**, **Part**, **Subclass** and **Purpose**. Relationships, consisting of one or more semantic relations connected together, constitute paths between two words, e.g. one path linking *car* and *person* is:

```
```

An extended path is a path created from subpaths in two different inverted LF structures. For example, *car* and *truck* are not related directly by a semantic relation or by a LF path from any single LF structure. If the two paths *car*-**Hypernym**-*vehicle* and *vehicle*-**Hypernym**-*truck*, each from a different LF structure, are joined on shared word *vehicle*, the resulting path is *car*-**Hypernym**-*vehicle*-**Hypernym**-*truck*. [5]

A MindNet database is an example base which stores information about the linguistic context in which word tokens were encountered in a corpus; a word’s meaning is defined by the pattern of its contextualized associations with other words.[5]

### 4 Conclusion

Finding natural language understanding system for understanding text samples based on the occurrence of a particular word in a particular context would allow computer lexicons and dictionaries to be more comprehensive and internally consistent. There is a need for building a lexicon of Croatian lexical units described in terms of frame semantics.

**References:**