

# Information Retrieval Meets Human Language Technology

Marko Tadić

Department of linguistics Faculty of Philosophy, University of Zagreb  
Ivana Lučića 3, 10000 Zagreb  
[marko.tadic@ffzg.hr](mailto:marko.tadic@ffzg.hr)

## Abstract

The paper tries to highlight several points where informational retrieval could benefit from human language technology. Although there is an increasing quantity of multimedia content on the web, the majority of information is still coded in natural language(s). Since the global trend in web language usage is coding in native language, it leads to the situation in which web pages in English do not represent the majority of web-text anymore. To retrieve information from web pages requires several tools, which have to deal with particular human language. These tools can be borrowed from the field of human language technology (HLT) and applied to web-text. Some possible areas of HLT which could be used in information retrieval from web-text are: 1) morphological processing: should be able to cope with different word-forms in particular language in order to make language specific full-text search available; 2) named entities recognition: should give the possibility to get information on concepts which have fixed names/ titles/formulas (such as personal/institutional/geographical names, temporal expressions etc.); 3) semantic thesauruses: should give the ability to retrieve information on the basis of language synonymy/proximity. The situation in the field of HLT for Croatian is discussed at the end of article.

## 1. Information retrieval

Today we are facing the situation where WWW became the largest open depository of information. It is more-or-less open, more-or-less freely browsable, and retrievable. Having in mind the amount of existing web-pages — I wouldn't even dare to give the exact number — we could also make a plausible assumption that the majority of information retrieval (in number of queries) today is Web-based.

What is the content of web pages? Essentially it is:

1. text
2. pictures (still or moving, bitmapped or vectors)
3. audio
4. any possible combination of items above

Although there is enormous quantity of multi-media information we wouldn't be very wrong if we say that the majority of structured information is still coded in textual form. Therefore, the information extraction from text is needed.

There are many techniques to retrieve chunks of texts regarding the structure of document (headings, titles, subtitles, captions, leads, etc.) but however efficient we could be in retrieving e.g. all the headlines of HRT news from their web-page it seems that we are missing something. We could dig a bit more information if we go deeper and try to get information, which is coded in natural language. The natural language is the system, which is responsible organizing the structures that are more-or-less under paragraph level. Tools for retrieval should be "sensitive to" or "aware of" human (natural) language.

Existing retrieval tools are sophisticated web-search engines, which allow complex (Boolean) query formation; exact or fuzzy matching; proximity searches and similar techniques.

These techniques are applicable for English language and are very efficient thanks to its linguistic structure. Linguistically speaking English language is:

- analytical language (Eng. *by the sword*  $\diamond$  Cro. *mačem*)
- has simple (almost none) morphology (Eng. *wolf/wolves/wolf's*  $\diamond$  Cro. *vuk/vuka/vuku/vuče/vukom/vuci/vukovi/vukove/vucima/vukovima*)
- fixed sentence order (Eng. *John loves Mary/Mary loves John*  $\diamond$  Cro. *Ivica voli Maricu/Maricu voli Ivica/Marica voli Ivicu/Ivicu voli Marica*)

What about other languages with different structures? At this point the field of Human language technologies could help us.

## 2. Human language technologies (HLT)

The name of the field could rise some brows. Let's explain it. If the technology is "science of technical procedures, which are used to process materials into products"<sup>1</sup> then in this case the material is natural language and the products are systems, which allow the user to use his/her natural language in computerised context easily.

Human language technologies are language specific i.e. they depend directly on the structure of processed natural language and should be found out separately of other languages.<sup>2</sup>

### 2.1 HLT according to EU Framework Programme 5

The field of HLT is defined in the European Union Framework Programme 5 under the main research area Information Society Technologies (IST) which takes the largest cut of the whole programme (26.3 % of FP5 budget). The Key Action III of IST (which alone has a budget of 564 millions of Euro) is named Multimedia and Content Tools (MC&T). The largest part of MC&T is

<sup>1</sup> Leksikon JLZ (1974), str. 974 (authors' translation).

<sup>2</sup> This doesn't mean that we should know nothing of other languages HLT. On the contrary — the experience of others should help us from wandering around and making the same mistakes they did.

HLT.<sup>3</sup> In the light of EU accession, which we are certainly facing, we should develop extensive HLT for Croatian.

(evidence and statistics) for developing other resources or building language tools.

## 2.2 Components of HLT

There are three main components of HLT:

1. language resources
  - 1.1. corpora
  - 1.2. dictionaries
2. language tools
  - 2.1. morphological level
    - 2.1.1. generators/analysers
    - 2.1.2. POS taggers
    - 2.1.3. ...
  - 2.2. syntax level
    - 2.2.1. shallow/deep/robust parsers
    - 2.2.2. sentence parts recognition (noun phrases)
    - 2.2.3. ...
  - 2.3. semantics
    - 2.3.1. detecting lexical meaning (synonymy, antonymy...)
    - 2.3.2. detecting sentence meaning (agent, patient...)
  - 2.4. pragmatics
    - 2.4.1. detecting and resolving deictics (today, this, we...)
    - 2.4.2. contextualizing utterances
    - 2.4.3. ...
  - 2.5. machine (aided) translation systems
3. commercial products

Language resources are corpora (collections of texts) and dictionaries stored in digital form (as e-text). Language tools are applications, which process or use the language resources. Commercial products are applications, which are results of research/treatment of language resources with language tools and are usually applied to new texts. Language tools are specific for each language and their building starts from the basis. Language resources serve as that basis and provide fundamental language data

## 3. Several HLT sub-fields interesting for information retrieval

Several techniques or language tools are applicable to the task of information retrieval. Here we shall concentrate only on three of them, which could give the acceptable results:

1. morphological processing
2. named entity recognition
3. (semantic) thesauruses

### 3.1 Morphological processing

Morphological processing is needed for morphologically rich languages like Finnish, Turkish, Russian, Croatian, Hindi etc. Words in texts appear in several word-forms. The simple solution for text information retrieval, which is "morphologically sensitive", is to enable search engines to generate different word-forms for the word(s) in query (see Fig. 1)

This could be done in at least two ways

1. real-time morphological generator which generates word-forms on the fly
2. morphological lexicon which has all the possible word-forms of at least 10.000 headwords

Either way includes morphological layer in search engines, which should enable language specific full-text search. For Croatian the morphological generator was theoretically defined and modeled in Tadić (1994).

### 3.2 Named entity recognition

Looking into the linguistic structure of text one can easily see that there are practically no texts without named entities (NE). NE usually carries extensive information because they connect the text with non-textual world with

## Morphologically sensitive query

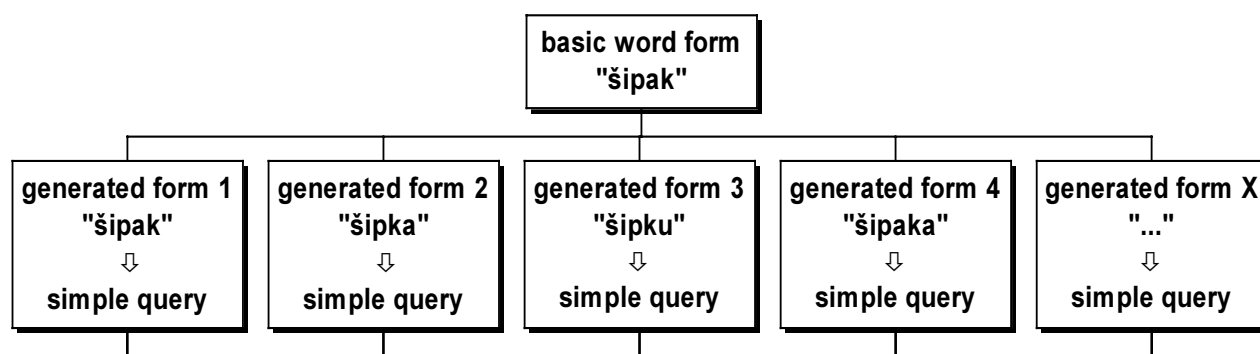


Fig 1 Morphologically sensitive query

<sup>3</sup> <http://www.cordis.lu/ist/home.html>. See also in Petek (2000),

their ability to ostensibly point to referents.<sup>4</sup> The common questions (*Who? When? What? Where?*) which are being asked when basic information about certain event are needed to situate it in time and space, and to tie it with responsible agents can be applied in this case. Named entity (NE) recognition is essentially the procedure of identifying and categorising names in text. The task of NE recognition was introduced by DARPA as a part of message understanding process. It has also been included in Message Understanding Conferences/ Competitions (MUC 6 (1995) and MUC 7 (1998)). In the task of NE recognition during MUC7 seven types of NE were introduced:

1. person
2. organisation
3. location
4. date
5. time
6. money
7. percent

Named entity recognition looks simple and straightforward but it is far from being trivial. Performance measures show how demanding this task can be. Human NE recognition yields result of 98-99% and best automatic systems can not get over 94%. NE recognition consists of two steps:

1. name identification (less problematic but the problem of different word-forms i.e. usage of morphological layer should be considered)
2. name categorisation (more complex, needs co-textual clues...)

The first step is usage of name lists, which enable the system to recognise boundaries of NE in text. Once the boundaries are detected the categorisation can take place. The Fig. 2 shows one simple newspaper article with NE marked just to demonstrate how frequent NEs are and how much information they are carrying.

```
<XML>
<BODY>
<DIV0 type="MAIN">
<HEAD type="NA">Nagrada zagrebačkim
gitaristima</HEAD>
<P><ENAMEX TYPE="ORGANIZATION">Zagrebački
gitaristički kvartet</ENAMEX> osvojio je
prvu nagradu na <ENAMEX
TYPE="ORGANIZATION">Međunarodnome
gitarističkom natjecanju Simone
Salmaso</ENAMEX> u <ENAMEX
TYPE="LOCATION">Viareggiu</ENAMEX> u
konkurenciji 14 komornih sastava (u
kategoriji D). Prvo mjesto je kao solist
osvojio i član toga renomiranoga
zagrebačkog sastava <ENAMEX
TYPE="PERSON">Darko Pelužan</ENAMEX> u
konkurenciji 30 gitarista (u kategoriji C).
Članovi <ENAMEX
TYPE="ORGANIZATION">Zagrebačkoga
gitarističkog kvarteta</ENAMEX> (koji je
1990. osnovao profesor <ENAMEX
TYPE="PERSON">Ante Čagalj</ENAMEX>,
pretežno od studenata gitare) sada su još
<ENAMEX TYPE="PERSON">Mihaela
Pažulinec</ENAMEX>, <ENAMEX
TYPE="PERSON">Krunoslav Pehar</ENAMEX> i
<ENAMEX TYPE="PERSON">Melita
Ivković</ENAMEX>. To nije prvi put da
<ENAMEX TYPE="ORGANIZATION">Zagrebački
gitaristički kvartet</ENAMEX> osvaja prvu
nagradu na nekome međunarodnom natjecanju u
<ENAMEX TYPE="LOCATION">Italiji</ENAMEX>:
pobijedio je i prije dvije godine u <ENAMEX
TYPE="LOCATION">Tarantu</ENAMEX> na 6.
međunarodnom natjecanju <ENAMEX
TYPE="ORGANIZATION">Trofeo
Kawai</ENAMEX>.</P>
<BYLINE><ENAMEX
TYPE="ORGANIZATION">Večernji
list</ENAMEX></BYLINE>
</DIV0>
</BODY>
</XML>
```

Fig. 2 Newspaper article with marked NEs

## Semantically sensitive query

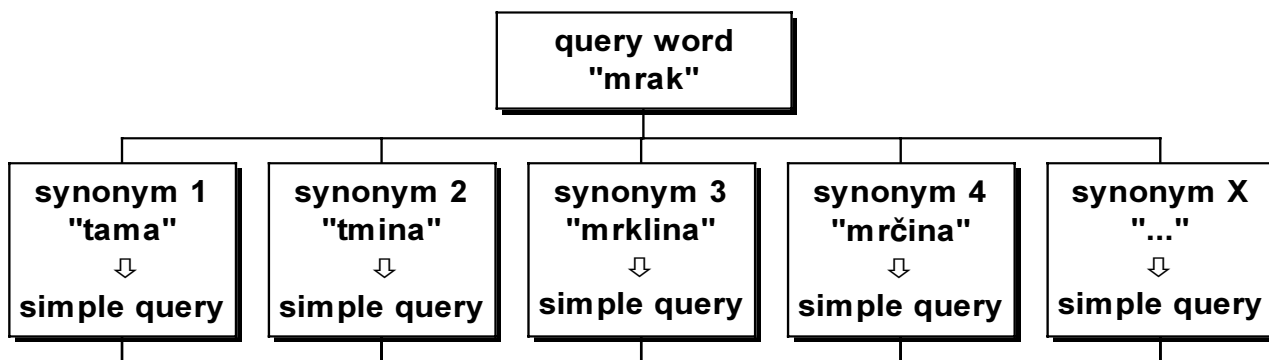


Fig. 3 Semantically sensitive query

<sup>4</sup> There are serious semantic theories which argue that names are not part of natural language(s) at all. Unfortunately this paper is not the place to discuss them.

The problematic NEs can be noticed in the example from Fig. 2 where the name of organisation is composed of common nouns accompanied by proper noun i.e. name of the person ("Međunarodnome gitarističkom natjecanju Simone Salmaso"). The NE recognition systems should be able to resolve and correctly recognise proper boundaries of NE as well as its category (in this case organisation instead of person).

The lists of NEs, which belong to the subfield of language resources, are of utmost importance for this task which usage in information retrieval need no particular explanation.

### 3.3 (Semantic) thesauruses

The tools, which are being considered here until now, function only on the plane of linguistic expression. Could it be possible to have a tool, which would enable the user to retrieve terms/words (and adjacent documents) not by variation of their expression but by variation of their content?

Semantic thesauruses should be that kind of tool. Essentially thesauruses are dictionaries where words are grouped by the similitude of their meanings (or difference of meanings or other kind of relation with them). Thesauruses should enable the user to retrieve information on the basis of linguistic:

1. synonymy: words with similar meaning ('mrak' & 'tama')
2. antonymy: words with different meanings ('mrak' <> 'svjetlo')
3. hyponymy: words which denote subordinated concepts ('alat' > 'čekić')
4. hypernymy: words which denote superordinated concepts ('kliješta' < 'alat')
5. meronymy: words which denote part-of ('glava' <- 'tijelo')

To achieve this goal the search engines should be able to generate several simple queries on the basis of the semantic relations coded in thesaurus (Fig. 3).

Such thesauruses can be produced as the result of building global semantic nets, which are depositories of words, organised according to semantic relations between them. Such famous semantic nets are WordNet<sup>5</sup> and EuroWordNet<sup>6</sup>.

## 4. State-of-art in HLT for Croatian

How does the situation in the field of HLT stand for Croatian? Just the quick glance will give us very disappointing picture:

### 4.1 Resources

In the subfield of language resources some activity can be detected.

After more than 30 years of tradition in Croatian corpora processing, at the Institute of linguistics, Faculty of philosophy, Univ. of Zagreb (which actually became the

referent institution for Croatian language processing), the Croatian National Corpus<sup>7</sup> is being collected. It is about to reach 30 million words by the end of 2000 and is expected to be expanded to 100 million words later. This corpus should give the basic evidence and statistics about Croatian language.

The second corpus compiled in the same institution is Croatian-English Parallel Corpus<sup>8</sup> (3.5 million of words, aligned on sentence level, coded in XML<sup>9</sup> according to XCES<sup>10</sup> standard)

The third corpus is Croatian-Slovene Parallel Corpus<sup>11</sup> (1 million words, aligned on the sentence level) which will be completed by middle of 2001.

The parallel corpora represent the inevitable basis for all kinds of translation studies as well as for machine (aided) translation systems.

Regarding electronic dictionaries the Croatian Frequency Dictionary<sup>12</sup> has been published in the basis of Moguš's 1 million corpus of Croatian. Its digital form will be available soon at the web site of the Institute of linguistics. Croatian Morphological Lexicon is being generated (10.000 headwords with accompanied word-forms by end of 2000) within the aforementioned project 130718.

### 4.2 Tools

With language tools the situation for Croatian is much more problematic.

On morphological level there is a generator<sup>13</sup> but no analyser. There is no Part-of-Speech tagger although some research on its building has been done at the Department of Informatics at the Faculty of Philosophy in Zagreb.<sup>14</sup>

On syntactic level there is no system for sentence parts detection and no parsers.

On semantic level there are no thesauruses yet (some are being collected at the moment), there is no WordNet and no systems for lexical and/or sentence meaning detection. I wouldn't even try to mention machine (aided) translation systems for Croatian.

### 4.3 Commercial products

For Croatian there are several spelling checkers but no grammar checkers. There is one commercial morphological generator called 'Morphological thesaurus'<sup>15</sup> and, unfortunately, no natural-language aware systems for information extraction/retrieval.

<sup>7</sup> Tadić (1996, 1998, 1999). See <http://www.hnk.ffzg.hr> where the freely retrievable test version is accessible. The research is financed by the Croatian Ministry of Science and Technology under project name *Computational processing of Croatian language* (130718).

<sup>8</sup> Tadić (2000).

<sup>9</sup> Bray-Paoli-Sperberg-McQueen (1998). See also <http://www.w3.org/>.

<sup>10</sup> Ide-Bonhomme-Romary (2000).

<sup>11</sup> Požgaj-Hadži – Tadić (forthcoming in 2000). The research is financed by Ministries of Science and Technology of Slovenia (project J6-7802-0581-99) and Croatia (project 130821).

<sup>12</sup> Moguš-Bratanić-Tadić (1999).

<sup>13</sup> Tadić (1994).

<sup>14</sup> Žubrinić (1995).

<sup>15</sup> Silić-Ranilović-Batnožić (1997).

<sup>5</sup> Miller (1990). See also <http://www.cogsci.princeton.edu/~wn/>.

<sup>6</sup> See <http://www.hum.uva.nl/~ewn/>.

#### 4.4 Warnings

For the conclusion, some warnings should be stated:

Regarding Croatian language, no one will develop HLT for Croatian beside us. This is the fact, which should give a thorough push to our efforts in that field because we are already late!

If HLT for Croatian won't be developed soon (within next few years) Croatian will become functionally illiterate language because of inability to participate in digital communication channels of 21st century (Internet, GSM...) Users of these channels will not cease to have the need of communication and in the lack of the ability of their own natural language to fulfil their needs, they will start to use another language which is more equipped for that usage. That will lead to limited usage of Croatian in communication and that situation could become even worse.

In order to stop that possible direction of development (or deterioration) some organised steps should be taken. The organised and well-funded research field should be established. Even more, Human Language Technology for Croatian should have the status of fundamental research in humanities. It should also have the status of strategic research for the Republic of Croatia and should become important scientific part of the Strategy of Croatian Development.

#### 5. Literature

- Bray, T., Paoli, J., Sperberg-McQueen, C. M. (eds.) (1998) Extensible Markup Language (XML) Version 1.0. W3C Recommendation. (<http://www.w3.org/TR/1998/REC-xml-19980210>).
- Ide, N., Bonhomme, P., Romary, L. (2000) XCES: An XML-based Encoding Standard for Linguistic Corpora. In Proceedings of LREC2000, ELRA, Paris-Athens, pp. 825-830.
- Leksikon JLZ (1974).
- Miller, G. (1990) Five papers on WordNet. In Special Issue of International Journal of Lexicography 3(4), revised August 1993, pp. 1-86.
- Moguš, M., Bratanić, M., Tadić, M. (1999) Hrvatski čestotni rječnik, Institute of linguistics, Faculty of Philosophy, University of Zagreb and Školska knjiga, Zagreb.
- MUC 6 (1995) Proceedings of the Sixth Message Understanding Conference (MUC-6), Morgan Kaufmann, San Mateo, CA..
- MUC 7 (1998) Proceedings of the Seventh Message Understanding Conference (MUC-7). (<http://www.muc.saic.com>)
- Petek, B. (2000) Funding for Research into Human Language Technologies for Less Prevalent Languages. In Developing Language Resources for Minority Languages: Reusability and Strategic Priorities, LREC2000 Workshop Proceedings, ELRA, Paris-Athens, pp. 100-105.
- Požgaj-Hadži, V., Tadić, M. (forthcoming) Slovensko-hrvatski paralelni korpus, Proceeding of the conference Jezikovne tehnologije za slovenski jezik, Ljubljana, 17-19. 10. 2000.
- Silić, J., Ranilović, B., Batnožić, S. (1997) Hrvatski računalni pravopis, Matica hrvatska and Sys print, Zagreb.
- Tadić, M. (1994) Računalna obradba morfologije hrvatskoga jezika, PhD thesis, Faculty of Philosophy, University of Zagreb, Zagreb.
- Tadić, M. (1996) Računalna obradba hrvatskoga i nacionalni korpus. In *Suvremena lingvistika* 41-42, pp. 603-612.
- Tadić, M. (1998) Raspon, opseg i sastav korpusa suvremenog hrvatskoga jezika. In *Filologija* 30-31, pp. 337-347.
- Tadić, M. (1999) Hrvatski nacionalni korpus na Internetu. In *Jezik* 46, 5, p. 200.
- Tadić, M. (2000) Building the Croatian-English Parallel Corpus. In Proceedings of LREC2000, ELRA, Paris-Athens, pp. 523-530.
- Žubrinić, T. (1995) Mogućnosti strojnoga označavanja i lematiziranja korpusa tekstova hrvatskoga jezika, MA thesis, Faculty of Philosophy, University of Zagreb, Zagreb.