Building the Croatian Language Technologies Portal

Marko Tadić,[a] Ivana Simeon[b]
[a]Department of Linguistics, [b]Institute of Linguistics
Faculty of Philosophy, University of Zagreb
{marko.tadic, ivana.simeon}@ffzg.hr

Abstract

Designing a portal dedicated to the Language Technologies applicable to Croatian would make the linguistic research easier and more efficient, providing students, researchers and corporate users, as well as anyone interested in language technologies, with not only the better access to information, but also with a professional forum where they can interact with others. In November 2000, the Croatian Language Technologies Portal project initiated by the Institute of Linguistics was started under the auspices of the Croatian Ministry of Science and Technology. It is designed as the central reference point containing an extensive list of links to Croatian and international institutions and projects related to the natural language processing, as well as to computational linguistic tools. It also has interactive components: the search engine for the contemporary part of Croatian National Corpus. In final version it will include also discussion boards, news groups, etc.

1. Introduction

The Croatian linguistic community is lacking a single virtual meeting point which would provide the researchers and those interested in the Croatian language with relevant information, references and resources. Data on Croatian language on Web are dispersed across a number of institutionally and geographically heterogeneous and often unrelated sites. One would expect that the proclaimed central institution for Croatian language in Croatia — namely, the Institute of Croatian Language and Linguistics — would be the one where the efforts for making and maintaining the homepage of Croatian language should be situated. Unfortunately, it is not so.

The fact that, for a nation, it is very important to have the national language present on the Web is self-evident. That presence is not just the number of words on the web pages published in that language.[1] It is reflected even more in the existence and free web-access to the fundamental language resources such as national corpus and general (mono- and multi-lingual) and specialized dictionaries. That goal is impossible to achieve without developing the 'linguistic infrastructure for Information Society' — HLT for that language. In our case, Croatian.

The situation with Human Language Technologies for Croatian is a mere reflection of the situation about Croatian language on Web in general. In order to correct that — at least in the field of HLT — in November 2000 we launched the so-called i-project *Human Language Technologies for Croatian (web portal)* which was supported by Croatian Ministry of Science and Technology under grant No. 00-86.

[1] For the estimate about the quantity of web-published text in different languages see Grefenstette & Nioche (2000).

2. What are Human Language Technologies?

The field of HLT is defined in the EU Framework Programme 5 (with continuation in FP6) under the main research area Information Society Technologies (IST) which takes the largest cut of the whole programme (26.3 % of FP5 budget or 3,600 MEuro). The Key Action III of IST (which alone has a budget of 564 MEuro) is named Multimedia and Content Tools (MC&T). The largest part of MC&T is HLT.[2] In the light of EU accession, which we are certainly facing, we should develop extensive HLT for Croatian. And we are already far behind the schedule.

»Human Language Technologies contributes to enhancing usability and accessibility of digital content and services while supporting linguistic diversity in Europe. (...) It is designed to anticipate the needs of the converging telecommunications, computing and media industries, and related markets and technologies. (...) HLT actions initially addressed three intertwined areas centred around how people interact with information, with information services and with each other:

- *Multilingual communication,* aimed at building multilingual intelligence into business processes, communication services, information appliances, and public interest services.
- *Natural Interactivity,* with the aim of enhancing the naturalness of human-computer interactions and the effectiveness of interpersonal communications.
- *Cross-lingual information management,* with a view to improving the effectiveness of information access and the efficiency of information handling.«[3]

3. What is the Croatian Language Technologies portal?

What we wanted to do with this project is to make a catalogue of institutions, projects, language resources and tools as well as activities related to HLT and to Croatian LT in particular. By putting this catalogue in public (at http://www.hnk.ffzg.hr/jthj) we wanted to offer the starting point for those who want to find out more about it. We also hope that that would lead to the much needed spreading and development of the research and activity in the field.

4. Who is it intended for?

The Portal is primarily intended for everyone interested in the application of information technologies to the Croatian language. More specifically, the target user groups include:

- *students and researchers* (in humanities – linguistics, phonetics, Croatian language and literature studies, as well as in computer and telematic disciplines – information science, telematics, signal processing, artificial intelligence, natural language processing) whose focus of interest is the processing of language data;
- *IT companies* – e.g. developers of natural language interfaces, natural language-coded data-indexing, full-text search, text compression and data-mining systems;
- *telematic companies* providing systems using natural (Croatian) language (e.g. SMS)
- *translation offices and services* using machine translation and machine-aided translation tools.

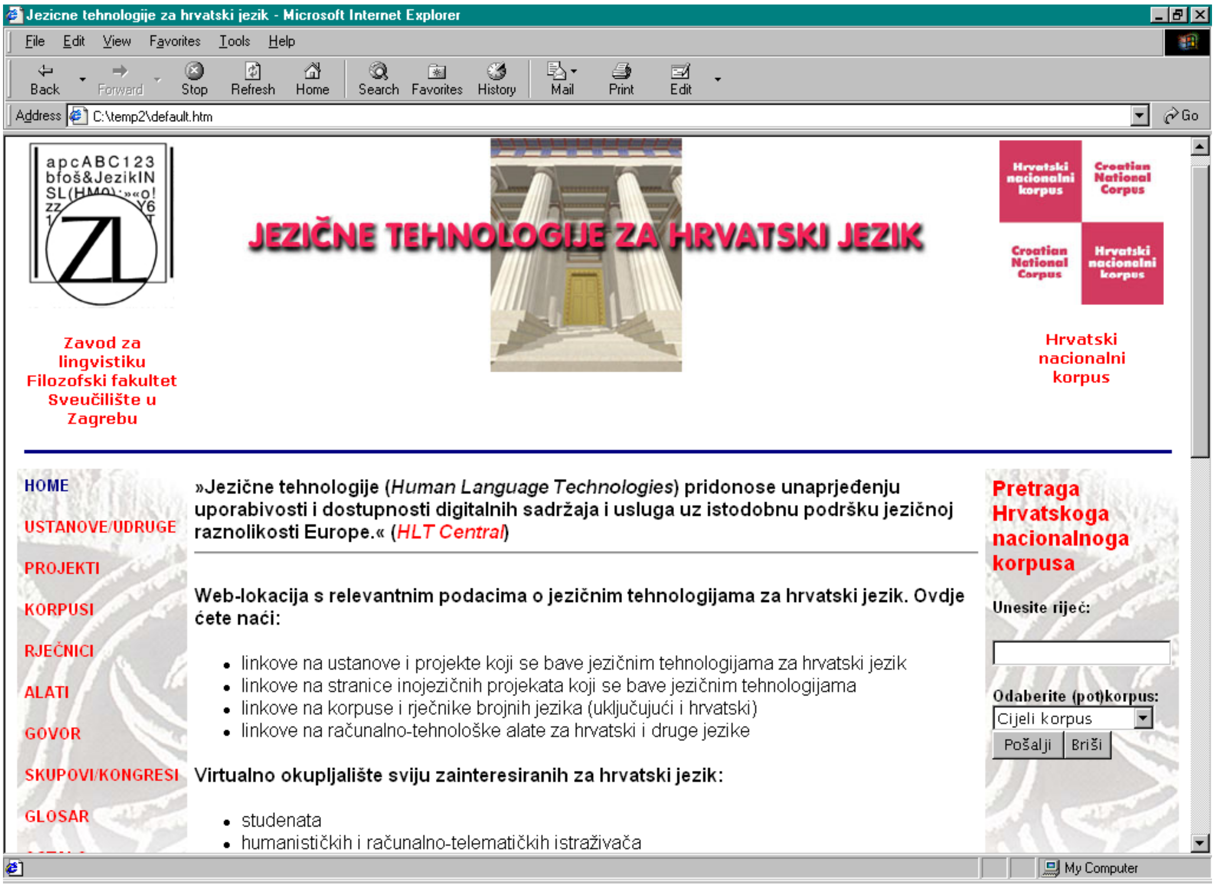
[2] Tadić (2000b) and see also in Petek (2000), p. 100.
[3] http://www.hltcentral.org/page-75.0.shtml.

5. The structure of the Portal

The portal is structured around two main areas:

- Informative component
- Interactive component

Each is subdivided in several sub-areas. The homepage reflects that division (figure 1).


Jezicne tehnologije za hrvatski jezik - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit
Address C:\temp2\default.htm
Go
apcABC123
bfoš&JezikIN
ZL
Hrvatski nacionalni korpus
Croatian National Corpus
Croatian National Corpus
Hrvatski nacionalni korpus
My Computer

Zavod za lingvistiku Filozofski fakultet Sveučilište u Zagrebu

JEZIČNE TEHNOLOGIJE ZA HRVATSKI JEZIK

Hrvatski nacionalni korpus

HOME
USTANOVE/UDRUGE
PROJEKTI
KORPUSI
RJEČNICI
ALATI
GOVOR
SKUPOVI/KONGRESI
GLOSAR

»Jezične tehnologije (*Human Language Technologies*) pridonose unaprjeđenju uporabivosti i dostupnosti digitalnih sadržaja i usluga uz istodobnu podršku jezičnoj raznolikosti Europe.« (*HLT Central*)

Web-lokacija s relevantnim podacima o jezičnim tehnologijama za hrvatski jezik. Ovdje ćete naći:

- linkove na ustanove i projekte koji se bave jezičnim tehnologijama za hrvatski jezik
- linkove na stranice inojezičnih projekata koji se bave jezičnim tehnologijama
- linkove na korpuse i rječnike brojnih jezika (uključujući i hrvatski)
- linkove na računalno-tehnološke alate za hrvatski i druge jezike

Virtualno okupljalište sviju zainteresiranih za hrvatski jezik:

- studenata
- humanističkih i računalno-telematičkih istraživača

Pretraga Hrvatskoga nacionalnoga korpusa

Unesite riječ:

Odaberite (pot)korpus: Cijeli korpus

Pošalji Briši

Figure 1. The homepage of Croatian Language Technologies portal

5.1. Informative component

5.1.1. Institutions and associations

This is a list of Croatian and foreign institutions, which are of relevance to the language technologies. The institutions included are the universities and, in particular, their departments that are involved in studying and developing language technologies, as well as state-operated and independent linguistic research groups, associations, institutions and labs. Each informative sub-component has the same structure. On the top of the page is the list of items relevant for Croatian, followed by items relevant for other languages in the middle. At the end the list of other similar web-based catalogues can be found.

5.1.2 Projects

Within the Croatian academic community, several projects are dealing, directly or indirectly, with the language technologies. Beside the project, which is being presented here, the *Computer Processing of* *the Croatian Language*, a project pursued at the Institute of Linguistics, Faculty of Philosophy, University of Zagreb is the most important project which is primarily concerned with language technologies and their implementation in Croatian linguistic studies. That project is the host of *Croatian National Corpus, which* is the most important Croatian language resource at the moment. This project hosts the *Croatian morphological lexicon*, which is also being developed and will be available on web soon in its full size (around 35.000 lexical entries with all possible word-forms).

However, a number of other institutions, namely the Department of Phonetics and the Department of Information Science at the Faculty of Philosophy, as well as the Faculty of Electrical Engineering and Computing and the Croatian Academic and Research Network also deal with language-related issues in their projects which are therefore listed under this link.

What we would like to see here is also a list of projects hosted by non-academic institutions which find their interest in Croatian language technologies either as investors and/or project partners/developers, or as (power, educated or test) users.

Several — already completed — European multi-lingual projects are also relevant for Croatian such as TELRI and TELRI II (Trans-European Language Resources Infrastructure) where the Institute of Linguistics was one of the associate partners and ELAN (European Language Activity Network) in which the Institute of Linguistic was acting as a full project member. Other relevant projects, starting with EU-funded mega-projects such as eContent (FP6), follow.

5.1.3. Corpora

Text corpora are the primary resource in language technology. Therefore, much effort is dedicated to their compiling and processing. The largest and the most advanced corpus of the Croatian language is the Croatian National Corpus *(HNK)*, compiled at the Institute of Linguistics, which currently consists of over 10 million tokens. It is composed of two constituents:

- 30-million *(30m)* representative sampled corpus of contemporary Croatian
- Croatian e-text Archive *(HETA)* which includes whole texts of contemporary or historical publications

The HNK is fully available and can be searched either through its homepage (http://www.hnk.ffzg.hr) or partially through this Portal.[4]

The corpora of foreign languages, which provide a good source of information and model for the compiling and processing methodology, are also listed.

In addition to monolingual corpora, bilingual and multilingual corpora are compiled for the purposes of cross-language research, translation studies and development of machine translation tools. Three bilingual corpora are being collected for Croatian: the Croatian-English Parallel Corpus, Croatian-Slovenian Parallel Corpus and Croatian-French Parallel Corpus all compiled by the Institute of linguistics and in different stages of completion.

5.1.4. On-line dictionaries

This link lists the on-line dictionaries currently available for Croatian. Unfortunately, the present number of such dictionaries is scarce. While many languages can boast general on-line dictionaries, there is only a limited number of specialized lexicons available for Croatian.

[4] More about Croatian National Corpus, its structure, size, time-span and accessibility see in Tadić (1996, 1998 and 1999).

5.1.5. Tools

The only Croatian-specific language tools that have been commercially designed so far are several spelling checkers (by Matica hrvatska/Sys,[5] Polar, TTE Inženjering, Dupor, etc.) and Softleks, a specialized text editor for writing dictionaries. However, there is an extensive list of language-independent tools and tools which can be adapted to suit the needs of a particular language. Some of these tools are still at the academic level while other are full-blown commercial products:

- *converters* (e.g. 2XML which converts RTF/HTML texts into XML)
- *mark-up tools* (dedicated editors for SGML/XML mark-up of language resources, SGML/XML parsers and validators)
- *concordancers* (which retrieve the text corpora and give results in its most common form for language researchers: concordances)
- *textual statistics and analysis tools*
- *taggers* (which mark the segments of the text on morphological and/or morphosyntactic level)
- *syntactic* tree-banks (which give the shallow or deep syntactic analysis of sentences)
- *semantic* networks (which give the semantic relations between words)
- *aligners* (which establish the connection between corresponding original and translated segments of bilingual and multilingual texts)
- *machine translation* and *machine aided translation tools* (which manage the complete machine translation or give so-called translation memories as aid in the process of massive translation)

5.1.6. Speech processing

This part of our portal has not been developed fully yet. The reason is that there are very scarce data about the research done in the field of Croatian speech processing. As far as we know there has been only one research project in that field — cooperation of Department of Phonetics, Faculty of Philosophy, Zagreb University with EU funded MBROLA project (based in TCTS Lab of the Faculté Polytechnique de Mons, Belgium) for speech synthesizer in 24 languages (including Croatian).

The US Army in cooperation with Carnegie Mellon University from Pittsburgh developed an English-Croatian and Croatian-English speech-to-speech translator for wearable computer.[6] They were testing the device in April 2001 in Zagreb and as far as we know no LT-aware team or person from Croatia were with them and not to mention the possibility of cooperation and research partnership. We hope that one of gains of this portal will be the feedback about such activities around Croatian in the field.

5.1.7. Conference calls

The Portal provides an extensive list of upcoming HLT-related conferences in Croatia, in Europe and around the world. The list will be constantly updated.

5.1.8 Language Technologies glossary

In order to simplify the browsing of the site and to make it more user-friendly, an 8-lingual (English, French, German, Spanish, Italian, Danish, Finnish and Croatian) glossary of computational linguistic terms with their definitions is included. In the future, the site will feature tutorials in computational and corpus linguistics.

[5] Silić-Ranilović-Batnožić (1997).

[6] The device was a product of combination of two Carnegie Mellon University projects: MT project DIPLOMAT (see http://www.lti.cs.cmu.edu/Research/Diplomat/) and speech interface for wearable computers project (see paper on *Speechwear* at http://www.speech.cs.cmu.edu/rspeech-1/air/papers/speechwear/speechwear.html).

5.1.9. Miscellaneous

This link points to LT-projects which are intended for people with special needs, several Croatian e-publishing web-sites, as well as the sites containing general information on the Croatian language.

5.2. Interactive component

5.2.1. Direct access and search of the Croatian National Corpus

The Croatian National Corpus is also accessible through the Portal. Only the 30m constituent can be searched. The user can submit a query with or without joker character and the results are given in the form of concordances with frequency information.

5.2.2. Discussion area (projected)

The discussion area is to be one of the main interactive features of the site. It should provide the users from all language research areas and backgrounds with a forum where they can submit their opinions, interact and exchange experiences with other users and obtain advice or assistance.

5.2.3. Direct access and search of the parallel corpora (projected)

At the Institute of Linguistics, Faculty of Philosophy, University of Zagreb several parallel corpora are being collected.[7] All of them are XML-encoded and aligned at the sentence level. They will be available for parallel concordance search at this Portal upon their completion.

The Croatian-English parallel corpus[8] containing 113 issues of the Croatia Weekly newspaper from 1998 to 2000, consisting of 1.6 million tokens for Croatian and 1.9 million tokens for English.

The Croatian-Slovenian parallel corpus, consisting of legal and scientific documents, as well as fiction, is being compiled together with the research partner from Ljubljana: Slavistic Department, Faculty of Philosophy, University of Ljubljana.[9]

The Croatian-French parallel corpus consists primarily of fiction texts. This corpus is at the stage of planning and basic text collecting.[10]

6. Further improvements and prospects

The Portal will be frequently updated and expanded to include new projects, newly established Croatian and foreign institutions/projects for language technologies. New resources will be added and, as mentioned above, didactic materials, such as computational and corpus linguistics tutorials. The Portal will also further develop its interactive component in order to obtain more input from the users and to provide them with advanced web-services.

Its ultimate goal is to help build awareness of the available language technologies for Croatian, within both Croatian and foreign research communities and institutions.

[7] See 5.1.3.

[8] Tadić (2000a) and Tadić (2001).

[9] Požgaj-Hadži & Tadić (2000).

[10] Tadić & Raffaelli (forthcoming).

It will also contribute to the efforts to achieve recognition of Croatian as an independent language and to include it in the global IT trends — i.e. digital communication channels of the 21st century.

7. Conclusion: Welcoming the feedback

The current version of the Portal offers as much language resources and tools, as we were capable to collect at the moment. However, we are aware that there is always room for improvement and will welcome any feedback from users. Therefore, we invite you to send us your input, comments and suggestions which would help us make this Portal meet your needs better and offer a greater contribution to the Croatian language studies both home and abroad.

8. References

Grefenstette, G., Nioche, J. (2000) Estimation of English and non-English Language Use on the WWW. In RIAO 2000 Proceedings.

HLT Central (2000) (http://www.hltcentral.org).

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. (2000) Slovensko-hrvatski paralelni korpus, Proceeding of the conference Jezikovne tehnologije za slovenski jezik, Ljubljana, 17-19. 10. 2000.

Rudnicky, A. I., Reed, S. D., Thayer, E. II. (1997) SpeechWear: A mobile speech system. (http://www.speech.cs.cmu.edu/rspeech-1/air/papers/speechwear/speechwear.html).

Silić, J., Ranilović, B., Batnožić, S. (1997) Hrvatski računalni pravopis, Matica hrvatska and Sys print, 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. (2000a) Building the Croatian-English Parallel Corpus. In Proceedings of LREC2000, ELRA, Paris-Athens, pp. 523-530.

Tadić, M. (2000b) Information Retrieval Meets Human Language Technology. In Quest for Information, CUC2000 Proceedings CD, CARNet, Zagreb, ISBN 953-6802-01-5.

Tadić, M. (2001) Procedures in Building the Croatian-English Parallel Corpus. In Special Issue of the International Journal of Corpus Linguistics, Vol 0(0), pp. 1-17.

Tadić, M., Raffaelli, I. (forthcoming) Croatian-French Parallel Corpus. In Suvremena lingvistika, 51.