Building the Croatian National Corpus

Marko Tadić

Department of Linguistics, Faculty of Philosophy, University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
marko.tadic@ffzg.hr

Abstract

The paper presents the work being done so far on the building of the Croatian National Corpus (HNK). It's being collected since 1998 at the Institute of Linguistics, Faculty of Philosophy, University of Zagreb. The size, time-span, its composition and criteria for text selection are being presented. The HNK consists of two parts: 1) 30-million corpus of contemporary Croatian language, 2) Croatian Electronic Textual Archive. The procedures of the corpus mark-up and processing are being discussed. One of the most interesting features of this corpus since its launch in 1998 is its availability for querying through the WWW. The future directions of 30m corpus enlargement to 100m in next few years, enhanced corpus management and querying as well as annotation and processing are being discussed at the end.

1. Introduction or predecessors

Since 1967 when the first text has been treated and computer-processed like a corpus (Bujas, 1975) the tradition of building Croatian monolingual corpora never ceased. It was entirely situated in the Institute of Linguistics at the Faculty of Philosophy, University of Zagreb. That tradition diversified from the very start in two directions regarding sources of texts, but mainly using the same methodology. The first direction, being extensively developed in '70 and '80, was processing of diachronical texts: notably old Croatian writers (Mediaeval, Renaissance, Baroque etc.). The second direction was processing of synchronical texts i.e. contemporary texts in Croatian. It was established in 1976 with the launch of the project *Computer Processing of Croatian Literary Language* in which the *One-million Corpus of Croatian Literary Language* was compiled. It was composed of five subcorpora (drama, poetry, prose, newspapers and textbooks) and with texts dating from 1935 to 1978. Due to different organizational, technical and partly war conditions, that corpus has not been completed until 1996.[1] It also served as the ground material for the first *Croatian Frequency Dictionary* (Moguš-Bratanić-Tadić, 1999).

Political changes in Croatia (first free elections in 1990 and independence from Yugoslavia in 1991) definitely had influence on the status of Croatian language which was transferred from formally official in one of six federal units of former Yugoslavia — but in reality in oppressed position in Federation — to a fully acknowledged state language.[2] These changes certainly had an impact on usage of Croatian as well, but there were no language data on which the research could be done. The only objective starting point for that was the development of language resources i.e. corpus which could be representative for Croatian in general. The results of its processing could then be compared with the results of processing of already existing corpora. Besides, the development of large-scale language resources for other languages became important for development of language technologies for each of them. So the idea of collecting *Croatian National Corpus* (henceforth HNK) was presented in (Tadić, 1996) and further elaborated in (Tadić, 1998). It also coincided with the launch of other corpora projects like Czech National Corpus (CNC) in 1994, Hungarian National Corpus (HNC) in 1997, FIDA corpus of Slovene in 1997 etc.

In the next part of this paper the general architecture of HNK and corpus parameters will be presented. In the third part the processing and the tools used are discussed. At the end of paper the conclusion gives also an overview of perspectives.

2. Architecture and parameters of HNK

The initial idea was to build a multimillion representative corpus of contemporary language, but it soon came up that there are needs to build corpora for old texts as well because previous research needed the revision and new readings in accordance with new material discovered. The decision to organize HNK in two components was also made having in mind the blurring of the distinct borderline between large corpus and text archive in so-called 3rd generation of corpora. Besides, having a larger corpus or text archive, as the background source of text from which a representative corpus can be derived is not a new concept.[3] The two-component architecture also respected existing tradition and already established research needs.

2.1. 30-million corpus of contemporary Croatian

First component of HNK (henceforth 30m) is collected from the written contemporary Croatian texts. The borderline of contemporarity was put in the year of 1990.

[1] See more about the history of Croatian corpora building in (Tadić, 1997).

[2] It doesn't mean that Croatian language didn't exist before 1990. On the contrary: it existed and survived in Austro-Ungarian Empire, in Kingdom of Yugoslavia between two World Wars as well as under communist authorities of post-WW2 Yugoslavia with similar tendencies of unification with Serbian language under name Serbo-Croatian, Croato-Serbian or any other version of two-folded name anchored in already outdated 19th century Slavistic tradition from Šafarik on. Yet its very name in Yugoslavia was sometimes officially recognized (like in WWII and 1974 Constitution) and sometimes completely banned (between two wars, '50-'70). The international recognition of Croatian as the distinct Slavic language remains to be done in years to come but that is well beyond the scope of this article. See more in (Kačić 1997).

[3] Sinclair (1987) about the building of COBUILD corpus and Čermák (1997) about the building of CNC.

Thus only the texts from 1990 on are being used.[4] The size of 30 millions was regarded as time/cost achievable at the time the whole project was launched. It turned out that it was underestimated in both, time/cost and size since the frist projected deadline for completion of 30m was by the end of 2001 and the 100m corpus would be preferred target today for a representative general-language corpus.

2.1.1. Structure of 30m and text typology

The proper corpus design should be made in the respect of the research done by social anthropologists or cultural sociologists on text production and reception. That kind of research, which should present corpus designers with numbers of titles lent in libraries or weeks spent on bestseller-lists, overall circulation figures etc., would give a more profound insight in text flow in society. Since no finance for that kind of targeted research was available, we had to use existing results from several commercially oriented marketing researches on newspaper and magazines reception (e.g. Gral, 1998; Podgornik, 1998) as well as literary critics on contemporary Croatian fiction (Oraić Tolić, 2001).

Also, when we started in 1998 we also consulted the recommended standards for text typologies (EAGLES, 1996) and structures of other large corpora (BNC and CNC).[5] That resulted in overall structure of 30m:

Text type	%
Faction (informative texts)	**74**
Newspapers	*37*
daily	22
weekly	9
bi-weekly	6
Magazines, journals	*16*
weekly	9
monthly	4
bi-, tri-monthly	3
Books, brochures, correspondence...	*21*
publicistics	4
popular texts	3.5
correspondence, ephemera	0.5
arts and sciences	13
Fiction (imaginative texts): prose	**23**
novels	*13*
stories	*5*
essays	*4*
diaries, (auto)biographies...	*1*
Mixed texts	**3**

Table 1: Structure of 30m

The text typology of 30m looks like this:

```
HNK.M         medium
HNK.M.G        spoken
HNK.M.E        electronic...
HNK.M.P        written
HNK.M.P.O       published
HNK.M.P.O.B      brochure
HNK.M.P.O.K      book
HNK.M.P.O.N      newspapers
HNK.M.P.O.N.D     daily
HNK.M.P.O.N.T     weekly
HNK.M.P.O.N.2     bi-weekly
HNK.M.P.O.R      magazines, journals
HNK.M.P.O.R.T     weekly
HNK.M.P.O.R.M     monthly
HNK.M.P.O.R.V     bi-, tri-monthly
HNK.M.P.N       not published
HNK.M.P.N.J      public
HNK.M.P.N.I      internal
HNK.M.P.N.O      personal

HNK.V       type, genre
HNK.V.N      faction
HNK.V.N.Z     sciences
HNK.V.N.Z.P    natural sciences...
HNK.V.N.Z.T    technical sciences...
HNK.V.N.Z.M    biomedical sciences...
HNK.V.N.Z.B    biotechnical sciences...
HNK.V.N.Z.D    social sciences...
HNK.V.N.Z.H    humanities...
HNK.V.N.S     expert texts
HNK.V.N.S.Z    travels
HNK.V.N.S.K    criticism
HNK.V.N.S.M    media
HNK.V.N.S.C    criminalistics
HNK.V.N.S.S    sports
HNK.V.N.S.P    politics
HNK.V.N.S.E    ecology, bioethics etc.
...
HNK.V.N.N     non-expert texts
HNK.V.N.N.P    publicistics
HNK.V.N.N.O    popular texts
HNK.V.N.N.E    ephemera
HNK.V.U      fiction
HNK.V.U.P      prose...
HNK.V.U.D      drama...
HNK.V.U.S      poetry...
HNK.V.M      mixed texts...
```

Figure 1. Text typology of HNK with respect to EAGLES recommendations

2.1.2. Availability of source texts

In the process of collecting texts some types/genres are easier to get while some are really difficult. Since at the beginning the decision has been made to avoid any typing or OCR input, we were forced to use texts available only in electronic form. As it was expected, there were no problems with newspapers, fiction, textbooks from social sciences or humanities, but natural and technical sciences still demonstrate severe lack of texts. This still results in

[4] In order to design a corpus the decision about time-span had to be done. Knowing the sociological background for Croatian, one could argue that that decision is not linguistically motivated and that is true. But one has to draw a line somewhere and 1990 seems to be important year for Croatian language as the whole and particularly from today's perspective. Besides, the developers of the Czech National Corpus had the same problem and decided for 1989, the year Václav Havel was elected for president and the year of political changes in former Czechoslovakia.

[5] Čermák (1997).

corpus disproportion. The 30m corpus is now at the size of 17 million tokens.

2.2. Croatian Electronic Textual Archive

The second component of HNK (henceforth HETA) is composed as the text archive without any size, time-span or text-type limitations. Currently in HETA there are texts collected from various sources, which are produced either before 1990 or after, but their complete insertion in 30m would result in its disbalance. Each of collections of texts is represented as a separate corpus: e.g. *Complete works of Ivan Gundulić*, *Complete works of Marko Marulić in Croatian*, *Croatian Literary Classics* (collection of 65 prose pieces from 15th to early 20th century, 3 million tokens), *Complete works of Miroslav Krleža* (being collected now, estimated on 7 million tokens), *Corpus of Croatian Chatrooms and Usenet Groups* (1.5 million tokens) or several newspapers collections (around 3 million tokens each). Contemporary texts are used in 30m if they fulfill criteria for insertion. Currently for HNK there are more than 100 million tokens collected in raw text format, but only about 28% are already converted into corpus format.

2.3. Corpus format

At the end of 1998 the decision has been made to enconde the corpus with XML although in that time it looked odd because of plethora of already existing SGML tools and standards. The further development turned that to a good decision. The HNK is encoded according to the XCES standard for corpus encoding (Ide, Bonhomme, Romary, 2000). It uses UNICODE standard for language specific characters (č,ć,đ,š,ž...). For the next processing phase (POS tagging/lemmatization) we would like to test the stand-off annotation which is easily achieved with XML.

3. Corpus processing and tools

The collecting of texts started in summer of 1998 when some, although limited, financing was provided by the government.[6] The first and cheapest source was newspapers and their web editions, which covered more than 60% of paper version. The DTP sources followed soon, providing us with other text types.

3.1. 2XML tool

In order to facilitate tedious process of text conversion from different formats, we developed a tool for conversion in XML. The tool, named 2XML, converts HTML/RTF files to XML in two steps. The first one produces the intermediary "dirty" XML with HTML/RTF text attributes preserved. In the second phase, the user-defined script is applied to intermediary XML format resulting in "full-blown" XML file. The tool has the capability of processing whole directories with the same script thus automating the conversion of large quantities of texts from the same source.[7]

3.2. Tokenizer

Simple tokenizer that converts XML files in two formats was built also:

- tabbed file format convenient for importing texts to databases (Figure 2)
- tokenized XML file (Figure 3).

The tabbed file format looks like this:

```
<DIV type="article">vl990301gr01   7      X
<HEAD type="nn">   vl990301gr01   28     X
U                  vl990301gr01   44     R
GORICI             vl990301gr01   46     R
SVETOJANSKOJ       vl990301gr01   53     R
ODR&#381;AN        vl990301gr01   66     R
12                 vl990301gr01   78     B
.                  vl990301gr01   80     I
FESTIVAL           vl990301gr01   82     R
PJEVA&#268;A       vl990301gr01   91     R
AMATERA            vl990301gr01   104    R
</HEAD>            vl990301gr01   111    X
<HEAD type="na">   vl990301gr01   118    X
Ivana              vl990301gr01   134    R
osvojila           vl990301gr01   140    R
&#382;upanijski    vl990301gr01   149    R
Sanremo            vl990301gr01   165    R
</HEAD>            vl990301gr01   172    X
<HEAD type="pn">   vl990301gr01   179    X
*                  vl990301gr01   195    I
Od                 vl990301gr01   197    R
20                 vl990301gr01   200    B
natjecatelja       vl990301gr01   203    R
&#382;iri          vl990301gr01   216    R
```

Figure 2. Tabbed file with token in 1st column, filename in 2nd, byte-offset in 3rd and token type in 4th where R = token, I = punctuation, X = XML tag, B = numeral

The tokenized XML file looks like this:

```
<DIV type="article" n="vl990301gr01">
<HEAD type="nn">
<W type="R">U</W>
<W type="R">GORICI</W>
<W type="R">SVETOJANSKOJ</W>
<W type="R">ODRŽAN</W>
<W type="B">12</W>
<W type="I">.</W>
<W type="R">FESTIVAL</W>
<W type="R">PJEVACA</W>
<W type="R">AMATERA</W>
</HEAD>
<HEAD type="na">
<W type="R">Ivana</W>
<W type="R">osvojila</W>
<W type="R">županijski</W>
<W type="R">Sanremo</W>
</HEAD>
<HEAD type="pn">
<W type="I">*</W>
<W type="R">Od</W>
<W type="B">20</W>
<W type="R">natjecatelja</W>
<W type="R">žiri</W>
```

Figure 3. Tokenized XML file

3.3. Sentence delimiting

Sentence boundary detection is being done with algorithm which inserts </S><S>-boundary between punctuation and capital letter. It is then filtered for known exceptions (mr., dr.,...). We have quite a lot of problems with manual correction of order-numbers which are by Croatian orthography written with a period after an Arabic or Roman numeral and that period is also a full-stop in 24% of cases.

[6] Grant for research project 130718 of Ministry of Science of Technology of the Republic of Croatia.

[7] About 2XML see in Tadić (2000).

3.4. POS tagging/lemmatization

Croatian, like all Slavic languages, is morphologically rich. A quick glance on flective system would give an inventory of seven cases, two numbers and three genders in declension (with peculiar intervening of semantics of animate/inanimate type in accusative singular at some declensional patterns), all that in two more categories (definite and indefinite) for adjectives, as well as comparison which spreads to adverbs, accompanied with declension of pronouns and cardinal numbers from one to four. In verbal system there are three simple indicative and three periphrastic tenses (with difference in genders and numbers in participles), two additional participles, two imperatives, two conditionals, all that in passive and very complex system of aspects (verbs can be perfective or imperfective which can be also iterative in several different ways...) which is historically substratum of tense system.

From that simple survey one would expect to see a lot of syntactic relations in sentences encoded with morphological categories and that is true. So, the POS annotation and lemmatization is more important for this type of language than, for instance English.

Unfortunately, that task for HNK is at beginning. The Croatian Morphological Lexicon (HML) has been generated by morphological generator for Croatian (Tadić, 1994). For each of 36.000 headwords/lemmas a complete list of all possible word-forms is generated and they are accompanied with MSD tags according to MULTEXT-East project specification.[8] Although it did not participate in that project from the scratch, beside initial six CEE languages, in 1998 the categories/values were defined for Croatian and they were included in revision of the MULTEXT-East MSD recommendation (Erjavec, 2001b).

```
= abeceda Ncfsn
abecede abeceda Ncfsg
abecedi abeceda Ncfsd
abecedu abeceda Ncfsa
abecedo abeceda Ncfsv
abecedi abeceda Ncfsl
abecedom abeceda Ncfsi
abecede abeceda Ncfpn
abeceda abeceda Ncfpg
abecedama abeceda Ncfpd
abecede abeceda Ncfpa
abecede abeceda Ncfpv
abecedama abeceda Ncfpl
abecedama abeceda Ncfpi
= abolicija Ncfsn
abolicije abolicija Ncfsg
aboliciji abolicija Ncfsd
aboliciju abolicija Ncfsa
abolicijo abolicija Ncfsv
aboliciji abolicija Ncfsl
abolicijom abolicija Ncfsi
abolicije abolicija Ncfpn
abolicija abolicija Ncfpg
abolicijama abolicija Ncfpd
abolicije abolicija Ncfpa
abolicije abolicija Ncfpv
abolicijama abolicija Ncfpl
abolicijama abolicija Ncfpi
```

Figure 4. Sample from Croatian Morphological Lexicon where the word-form is at the beginning of a row, lemma in the middle and MSD at the end

We are planning to match the HML with corpus to get the annotated corpus, which should in the final output look like this:

```
<DIV type="article" n="vl990311ck01">
<HEAD type="nn">
<TOK type="R"><ORTH>POLICIJA</ORTH>
  <LEX><BASE>policija</BASE>
       <MSD>Ncfsn</MSD></LEX></TOK>
<TOK type="R"><ORTH>O</ORTH>
  <LEX><BASE>o</BASE>
       <MSD>Spsl</MSD></LEX></TOK>
<TOK type="R"><ORTH>DETALJIMA</ORTH>
  <LEX><BASE>detalj</BASE>
       <MSD>Ncmpl</MSD></LEX></TOK>
<TOK type="R"><ORTH>VEZANIM</ORTH>
  <LEX><BASE>vezan</BASE>
       <MSD>Afpmpl-</MSD></LEX></TOK>
<TOK type="R"><ORTH>UZ</ORTH>
  <LEX><BASE>uz</BASE>
       <MSD>Spsa</MSD></LEX></TOK>
```

Figure 5. Sample of POS/MSD disambiguated and annotated (lemmatized) text from HNK

But since there is no statistical data about the possibilities of homographic forms in Croatian, our intermediate format would have to look like this:

```
<DIV type="article" n="vl990311ck01">
<HEAD type="nn">
<TOK type="R"><ORTH>POLICIJA</ORTH>
  <LEX><BASE>policija</BASE>
       <MSD>Ncfsn</MSD>
       <MSD>Ncfpg</MSD></LEX></TOK>
<TOK type="R"><ORTH>O</ORTH>
  <LEX><BASE>o</BASE>
       <MSD>Spsl</MSD>
       <MSD>I-s</MSD>
       <MSD>Ncnsn</MSD></LEX></TOK>
<TOK type="R"><ORTH>DETALJIMA</ORTH>
  <LEX><BASE>detalj</BASE>
       <MSD>Ncmpd</MSD>
       <MSD>Ncmpl</MSD>
       <MSD>Ncmpi</MSD></LEX></TOK>
<TOK type="R"><ORTH>VEZANIM</ORTH>
  <LEX><BASE>vezan</BASE>
       <MSD>Afpmsin</MSD>
       <MSD>Afpmpdn</MSD>
       <MSD>Afpmpln</MSD>
       <MSD>Afpmpin</MSD>
       <MSD>Afpfpdn</MSD>
       <MSD>Afpfpln</MSD>
       <MSD>Afpfpin</MSD>
       <MSD>Afpnsin</MSD>
       <MSD>Afpnpdn</MSD>
       <MSD>Afpnpln</MSD>
       <MSD>Afpnpin</MSD>
       <MSD>Afpmsiy</MSD>
       <MSD>Afpmpdy</MSD>
       <MSD>Afpmply</MSD>
       <MSD>Afpmpiy</MSD>
       <MSD>Afpfpdy</MSD>
       <MSD>Afpfply</MSD>
       <MSD>Afpfpiy</MSD>
       <MSD>Afpnsiy</MSD>
       <MSD>Afpnpdy</MSD>
       <MSD>Afpnply</MSD>
       <MSD>Afpnpiy</MSD></LEX></TOK>
<TOK type="R"><ORTH>UZ</ORTH>
  <LEX><BASE>uz</BASE>
    <MSD>Spsa</MSD></LEX></TOK>
```

Figure 6. Sample of POS/MSD non-disambiguated text

[8] Erjavec, Ide (1998) and Erjavec, Lawson, Romary (1998)

The homographical word-forms from different lemmas are not so common (particularly, overlapping between nouns and verbs is quite rare) but in Croatian texts in general a lot of "internal" homography between different word-forms belonging to the same lemma can be found. The syncretism of cases is very frequent (see *detaljima* and *vezanim* in Figure 6). So the process of disambiguation is somewhat different than in English and includes more co-textual data but usually theya re adjacent or in very near distance.

We are planning to automatically annotate and manually correct a 1 million tokens of Croatian and then use them as the training corpus for a tagger. Knowing the results of evaluating different taggers for tagging similar Slavic languages — e.g. Erjavec, Džeroski, Zavrel (2000) on Slovene — we'll try to use the TNT tagger.

3.5. Collocation detecting

Up until now only one test-research of collocation detecting has been done on HNK data. It consisted of finding adjacent bi-grams from the test version of 30m encompassing 7.6 million tokens. The method used was the statistical measure of Mutual Information[9] and it has been applied on non-lemmatized texts.

Further testing on smaller lemmatized samples yielded better results but the question whether to lemmatize collocations in order to treat them statistically remains to be answered — at least for Croatian with its morphological complexity. By lemmatizing a collocation, it looses its "collocational strength" or "pattern" because characteristic combinations of MSDs attached to collocational constituents are being mapped to MSD of lemmas of these constituents. In that way, something what is not collocation at all, but a mere coincidence, may be included in statistics resulting in its decreased accuracy. On the other hand, lemmatization definitely helps to collect flective variations of word-forms of collocational constituents and normalize them under same lemma. In that way statistics becomes more accurate. The exact ratio between gain and loss in statistical accuracy between lemmatized and non-lemmtized collocations remains to be investigated.

3.6. Availability of HNK

Knowing that Croatian is lesser spread language, one of the main tasks from the launch of the project was to make HNK available for querying over WWW service. The homepage of HNK can be found at the address http://www.hnk.ffzg.hr.

At the end of 1998 free web access to the first test version of 30m corpus (3 million tokens) was enabled. Incremental test versions (7.6, 9 and 12 million tokens) were realized in next three years. HETA was available since 1998 also, while *Croatian Classics* (ca 3 million tokens) were added in 1999 and other contemporary text collections in 2001 and 2002.

The rationale behind putting test (non-representative and disproportional) versions of corpora to web was that it is better to have any kind of data available, be it even so inconsistent, than no data at all. The visiting statistics, as well as number of projects and/or individual researchers using HNK data proved that the initial decision to go public was correct.

Domain name	Visits	%
Croatia (.hr)	119542	63.01
Commercial (.com)	22371	11.79
Germany (.de)	18312	9.65
Network (.net)	11400	6.01
Austria (.at)	2848	1.50
Educational (.edu)	2037	1.07
Slovenia (.si)	1378	0.73
Netherlands (.nl)	990	0.52
Australia (.au)	870	0.46
Czech Republic (.cz)	782	0.41
Italy (.it)	770	0.41
Canada (.ca)	747	0.39
Yugoslavia (.yu)	698	0.37
France (.fr)	682	0.36
Sweden (.se)	669	0.35
Poland (.pl)	555	0.29
Bosnia and Herzegowina (.ba)	502	0.26
Russian Federation (.ru)	396	0.21
United Kingdom (.uk)	396	0.21
Japan (.jp)	373	0.20
Switzerland (.ch)	326	0.17
New Zealanad (.nz)	294	0.15
Denmark (.dk)	276	0.15
Slovakia (Slovak Republic) (.sk)	224	0.12
Hungary (.hu)	205	0.11

Table 2: Visiting statistics of HNK site in 1999 and 2000 by domain names (top 25)

Figures in Table 2 show that almost 37% of all traffic is outside Croatia (.com Croatian companies being relatively rare) what assured us that HNK overcame the national role and became every-day source of data for researchers and lecturers from many Slavic departments of universities abroad (particularly German ones).

The only service, which is available for current test-version of HNK, is web-concordance. It is given in KWIC format with option to turn to KWIL format with larger co-text for individual tokens. Absolute and relative frequency data accompany the concordance.

Somewhat different from other projects, the whole corpus is maintained on WindowsNT platform with MS SQL server, as the DBMS because we had free academic licenses of all software needed. The web-based search interface was programmed entirely by HNK developing team, featuring ASP with ODBC connectivity. Since only the test versions of corpus were made available on web, and those versions were not XML encoded, the search interface was not developed any further. It allows queries on single words only (but enables joker characters from any side).

At the moment we are experimenting with new corpus manager called Manatee[10] coupled with Bonito search client. That manager was developed by Pavel Rychlý at the University of Brno and it runs on Unix machines but the Bonito client has both, Unix and Windows versions.

[9] In fact Pointwise MI, as is defined in Manning, Schütze (1999).

[10] Discussion on corpus managers can be seen in Rychlý (2000).

4. Conclusion or perspectives

The paper presented the work done so far in developing Croatian National Corpus. The two-constituent structure of HNK will remain in the future because it has proved to be useful concept of collecting texts-candidates for corpus and selecting texts, which are being inserted in corpus.

The first and most important task after completion of 30m corpus would be its enlargement to 100m. This size is today's standard for general language representative balanced corpora and Institute of Linguistics will soon have enough texts to complete this task. The spoken subcorpus (e.g. discussions from Croatian Parliament will be included).

The work on POS tagging/lemmatization will go further with the goal of POS tagging/lemmatization of complete 30m corpus after training the tagger. It remains to be seen if the original MULTEXT-East tagset should be kept in its developed form or some kind of its derivation or downsizing would be necessary in order to increase precision/recall of tagger.

The enhanced corpus management and developed and flexible search tools are of utmost importance for users of any corpus. We expect that Manatee corpus manager would give that flexibility to HNK users and enable both, single/multi-word and POS/MSD queries.

The plans for future would be to start with syntactic annotation and try to build for the beginning a modes treebank for Croatian. Which grammatical formalism and annotation tools will be used remains to be decided.

The system for named-entity recognition in Croatian will be developed on the basis of HNK in next two years. That will allow not only better processing of names, collocations but will eventually help in parsing.

The semantic annotation is also one of points on agenda but that will unfortunately have to be reserved for better future times.

5. References

Biber, D., Conrad, S., Reppen, R., 1998. *Corpus Linguistics. Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Bujas, Ž., 1975. *Ivan Gundulić »Osman«. Kompjutorska konkordancija.* Zagreb: Sveučilišna naklada Liber.

Čermák, F. Czech National Corpus: A Case in Many Contexts. *International Journal of Corpus Linguistics* 2:181-197.

EAGLES: Expert Advisory Group on Language Engineering Standards, 1996. (http://www.ilc.pi.cnr.it/EAGLES96/home.html).

Erjavec, T., 2001a. MULTEXT-East Resources Revisited. *Elsnews*, 10.1:3-2.

Erjavec, T., 2001b. MULTEXT-East Resources, Concede Edition. (http://nl.ijs.si/MTE/V2 and http://nl.ijs.si/ME/V2/msd/html/).

Erjavec, T., Ide, N., 1998. The MULTEXT-East Corpus. In *Proceedings of the First International Language Resources and Evaluation Conference*, 2:971-974.

Erjavec, T., Lawson, A., Romary, L., 1998. *East meets West. A Compendium of Multilingual Resources*. Mannheim: TELRI (CD-ROM, ISBN 3-922641-46-6).

Erjavec, T., Gorjanc, V., Stabej M., 1998. Korpus FIDA. In *Proceedings of the Conference 'Language Technologies for the Slovene Language'*. Ljubljana: Institute Jožef Stefan.

Erjavec, T., Džeroski, S., Zavrel, J. 2000. Morphosyntactic Tagging of Slovene: Evaluating Taggers and Tagsets. In *Proceedings of the Second International Language Resources and Evaluation Conference*, 2:1099-1104.

Gral, 1998. *Mediapanel tisak. Čitanost tiskovina u Hrvatskoj*. Zagreb: Gral marketing (internal report 2/98).

Ide, N., Bonhomme, P., Romary, L., 2000. An XML-based Encoding Standard for Linguistic Corpora. In *Proceedings of the Second International Language Resources and Evaluation Conference*, 2:825-830. (http://www.cs.vassar.edu/XCES).

Kačić, M., 1997. *Croatian and Serbian. Delusions and distortions*. Zagreb: Novi most.

Manning, C., Schütze, H., 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Oraić Tolić, D., 2001. Suvremena hrvatska proza: Izazov zbilje. *Republika* 5-6:39-51.

Pala, K., Rychlý, P., Smrž, P., 1997. DESAM - Annotated Corpus for Czech (DESAM - Annotated Corpus for Czech). In *Proceedings of SOFSEM 97*. Heidelberg: Springer Verlag. (http://nlp.fi.muni.cz/publications/sofsem1997_pala_pary_smrz/).

Pala, K., Rychlý, P., Smrž, P., 1998. Corpus Annotation in Inflectional Languages: Czech. In Min Tjoa, A., Roland R. Wagner (eds.), *Ninth International Workshop on Database and Expert Systems Applications. Los Alamitos*, California: IEEE Computer Society. (http://nlp.fi.muni.cz/publications/dexa1998_pala_pary_smrz/)

Podgornik, B., 1998. Dnevne novine čita 75 posto građana, *Novi list*, 25-11-98.

Rychlý, P., 2000. *Korpusové manažery a jejich efektivní implementace (Corpus Managers and their effective implementation)*. Ph.D. thesis, University of Brno. (http://www.fi.muni.cz/~pary/disert.ps)

Sinclair, J., 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Tadić, M., 1994. *Računalna obradba morfologije hrvatskoga književnoga jezika*. Ph.D. thesis, University of Zagreb. (http://www.hnk.ffzg.hr/txts/mt_dr.pdf).

Tadić, M., 1996. Računalna obradba hrvatskoga i nacionalni korpus. *Suvremena lingvistika*, 41-42:603-611. (English version at http://www.hnk.ffzg.hr/txts/mt4hnk_e.pdf).

Tadić, M., 1997. Računalna obradba hrvatskih korpusa: povijest, stanje i perspektive. *Suvremena lingvistika*, 43-44:387-394 (English version at http://www.hnk.ffzg.hr/txts/mt4hnk3_e.pdf).

Tadić, M., 1998. Raspon, opseg i sastav korpusa hrvatskoga suvremenog jezika. *Filologija*, 30-31:337-347. (English version at http://www.hnk.ffzg.hr/txts/mt4hnk2_e.pdf).

Tadić, M., 2000. Building the Croatian-English Parallel Corpus. In *Proceedings of the Second International Language Resources and Evaluation Conference*, 1:523-530.

Váradi, T., 1999. On Developing the Hungarian National Corpus. In Vintar, Š. (ed.), *Proceedings of the Workshop Language Technologies — Multilingual Aspects*. Ljubljana: Department of Translation and Interpreting, Faculty of Arts, University of Ljubljana.