

Procedures in Building the Croatian-English Parallel Corpus

Marko Tadić
marko.tadic@ffzg.hr
<http://www.hnk.ffzg.hr/mt>

Department of Linguistics
Philosophical Faculty
University of Zagreb
Ivana Lučića 3
HR-10000 Zagreb
Croatia

This contribution gives a survey of procedures and formats used in building the Croatian-English parallel corpus which is being collected at the Institute of Linguistics at the Philosophical Faculty, University of Zagreb. The primary text source is the newspaper *Croatia Weekly* which has been published from the beginning of 1998 by HIKZ (Croatian Institute for Information and Culture). After a quick survey of existing English-Croatian parallel corpora, the article copes with procedures involved in text conversion and text encoding, particularly the alignment. There are several recent suggestions for alignment encoding, and they are listed and elaborated at the end of the article.

KEYWORDS: corpus linguistics, parallel corpora, Croatian language, English language, corpus encoding, alignment, CES, XML

1. Introduction

For any kind of research involving two or more languages such as multilingual lexicography, contrastive linguistics, machine translation, etc., parallel corpora are of essential importance. Knowing the role of English today as *lingua communis*, it is no surprise that the most common pairing of languages in parallel bilingual corpora is English : L_x. This is the reason why we chose English as a pair to the Croatian from the beginning.

Many scholars probably do not know that this very language pairing in parallel corpora started more than 30 years ago: Professor Rudolf Filipović launched the *Yugoslav Serbo-Croatian—English Contrastive Project*¹ in 1968. The preliminary idea was brought to Zagreb by Professor Bujas in 1967 when he returned from Austin, TX. (Bujas 1967). Until 1971, when the project ended, the Brown corpus was acquired, cut in half (505 822 tokens) preserving the original 15 genre balance, and morphosyntactically marked and translated (Bujas 1969:36). The concordance with morphosyntactic categories as keywords was produced as well as a bilingual sentence database (Bujas 1975:53).

As far as we know, this was the first implementation of computers in contrastive linguistics. Computer data tapes still exist at the Institute of Linguistics, but, unfortunately, it is impossible to find a computer system which would be able to read them — so they are of no practical use today. Nevertheless, the project resulted in a great number of publications, primarily in the field of contrastive linguistics, known as *Contrastive Studies*, *New Contrastive Studies* and *Chapters in Contrastive Linguistics*, all published by the Institute of Linguistics, Philosophical Faculty, University of Zagreb.

The second Croatian-English parallel corpus is the translation of Plato's *Republic*, published on TELRI CD-ROM (Erjavec, Lawson, Romary 1998); however, the Croatian-English language pair is not the only one, and it was certainly not of the primary interest. Since the whole work is well known to the TELRI community and wider, we will go on with our topic.

The third Croatian-English parallel corpus has been collected in the scope of example based machine translation project known to us only by paper from the LREC 1998 conference (Allen & Hogan 1998). The size of that corpus is about 0.85 Mw for Croatian part and about 0.78 Mw for English part (Allen & Hogan 1998:749).

2. Corpus

The *Croatian-English parallel corpus*, which is now being collected at the Institute of Linguistics at the Philosophical Faculty, University of Zagreb is the fourth Croatian-English corpus pair. Its primary aim was to investigate procedures of text-conversion, corpus collection/organization, sentence alignment, and corpus encoding which would be used in later parallel corpora projects, such as the *Croatian-Slovene parallel corpus*, which was approved by both Ministries of Science in July 1999 and was effectively launched in October 1999.

This fact must be pointed out. It is this very spot where TELRI could find affirmation of its international efforts: two members of the TELRI Association were able to launch a bilateral project approved at the formal national level of their Ministries of Science. Formally, this project exists outside the TELRI framework, but without the TELRI “stirring pot,” it would not have been possible.

In corpora collection, there are several factors which should be kept under control. The representativeness of the corpus is one of them — an ideal which is hard to achieve, yet everyone is trying to come to its vicinity. The situation is even worse in the case of parallel corpora since the demand for parallelism narrows the already limited choice of texts. Also for languages with a small number of speakers and/or translators such as Croatian, one can be happy to obtain any valuable translations. The outcome is usually a rather unbalanced set of bitexts because you have to take whatever you can get in digital form. It would be “methodologically cleaner” to have a corpus originating from one text source, which you could call *Corpus of This-and-That*. Fortunately, we found ourselves in such a situation.

The source of texts is the newspaper *Croatia Weekly*, published by HIKZ (Croatian Institute for Culture and Information) from the beginning of 1998 until May 2000. The publication is similar to *USA Today* in a Croatian way — it covers different domains: politics (internal and foreign), economy and finance, tourism, ecology, culture, art, sports, and events, and it is intended for the public abroad. It contains 16 pages (including 4 pages for advertising) giving us an average of 14 400 tokens per issue for Croatian and 17 400 for English. The last issue published has number 118, and we have access to the digital form of all texts in both languages except for the first 5 issues. Thus 113 issues provides approximately 1.6 Mw for Croatian and approximately 1.9 Mw for English.

The only problem which could cast a shadow on our “methodological happiness” is the fact that the most popular weekly in Croatia, *Nacional*, which is one of the most important Croatian language sources for our Croatian National Corpus, started with English translations on its Web page. These translations cover approximately 15% of the original Croatian texts. Now, choosing the text candidates for the corpus, we are in the position to decide between “methodological purity” and the size as well as topic variation. For the time being, we will stick to only one text source — *Croatia Weekly*. In future versions of the corpus, texts from other sources will be included.

3. Making the corpus

3.1. Platform

Surprisingly, our platform is not UNIX — all software (commercial, shareware, and custom made) runs on Windows 9*/NT. A few years ago this would be peculiar, but today, when language technologies have already descended to the market level, it seems to be a mere technical exercise.

3.2. Text formats

Croatian texts, delivered by the publisher to a professional translation bureau, are available in “bare ASCII” format, completely stripped of any markup. Thus, for the Croatian half of the pair, markup has to be done by macros and scripts used in

commercial text-processors (MS Word 97). The English texts are supplied in typesetting format (QuarkXPress 3.32); we extract them as RTF files and process them further.

3.3. Conversion

We have designed an application called 2XML and engaged an independent software company to do the programming work. The application performs conversion by applying user-defined scripts to input in the form of RTF or HTML file, resulting in output, delimited at the beginning and at the end with <BODY>...</BODY>, which is “full blown” XML. Figure 1 gives the overview of the script-editing page of the 2XML application.

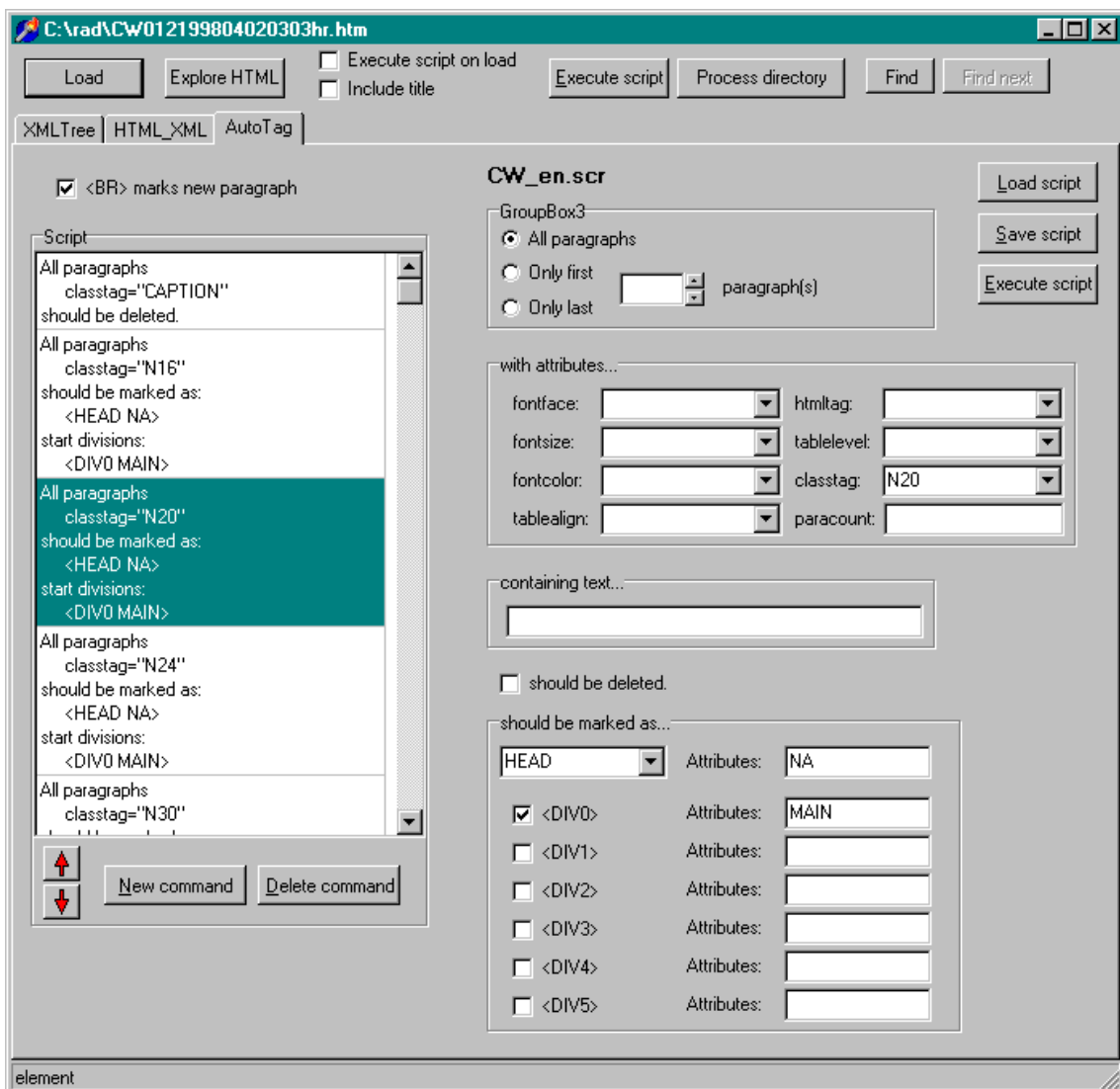


Figure 1.

The conversion is made in two steps:

- 1) the program produces the “dirty” XML with `</P>` marked only, where certain HTML and/or RTF attributes (typeface name, font size, margin alignment, style name, etc.) are preserved (Fig. 2 shows just few of them);

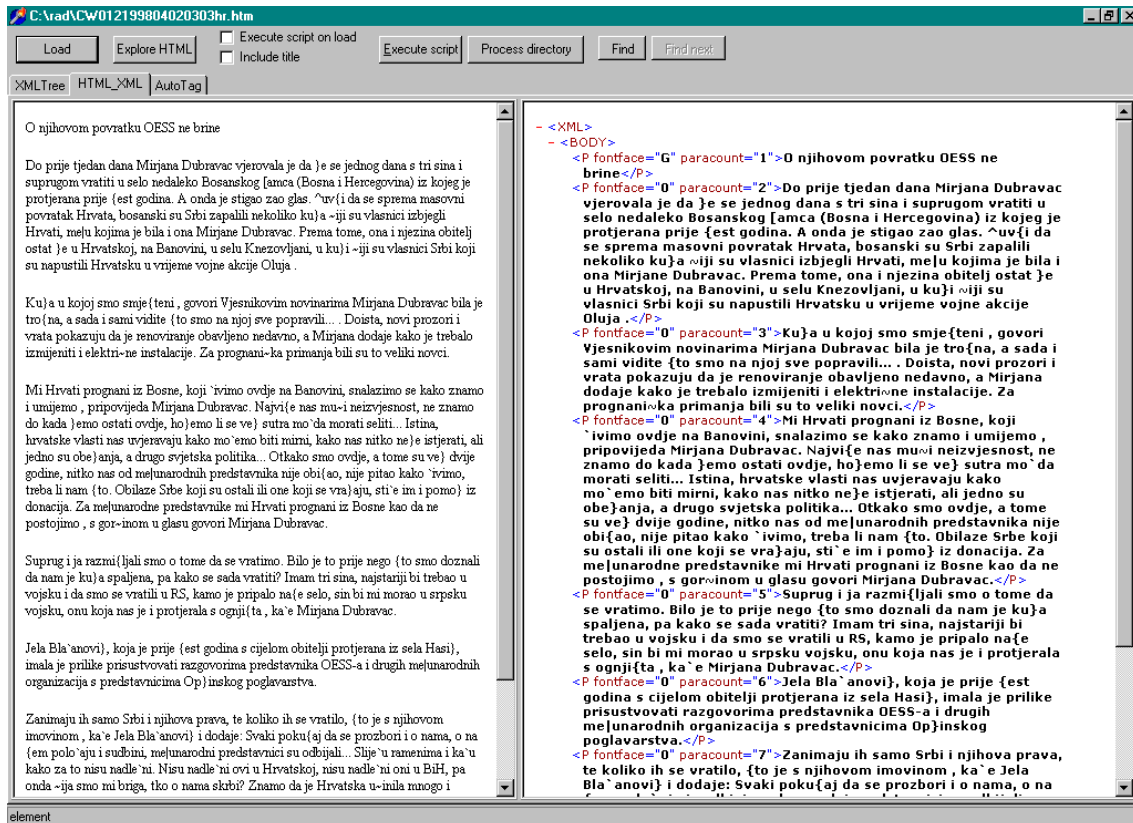


Figure 2.

- 2) The user-defined script is run on the “dirty” XML file, producing the final, “clean” XML file where HTML and/or RTF attributes, preserved from the first stage, are replaced by XML opening and closing tags. Usually these are different kinds of `<DIVs>` and `<HEADs>` with their specific attribute values defined by script (Fig. 3).

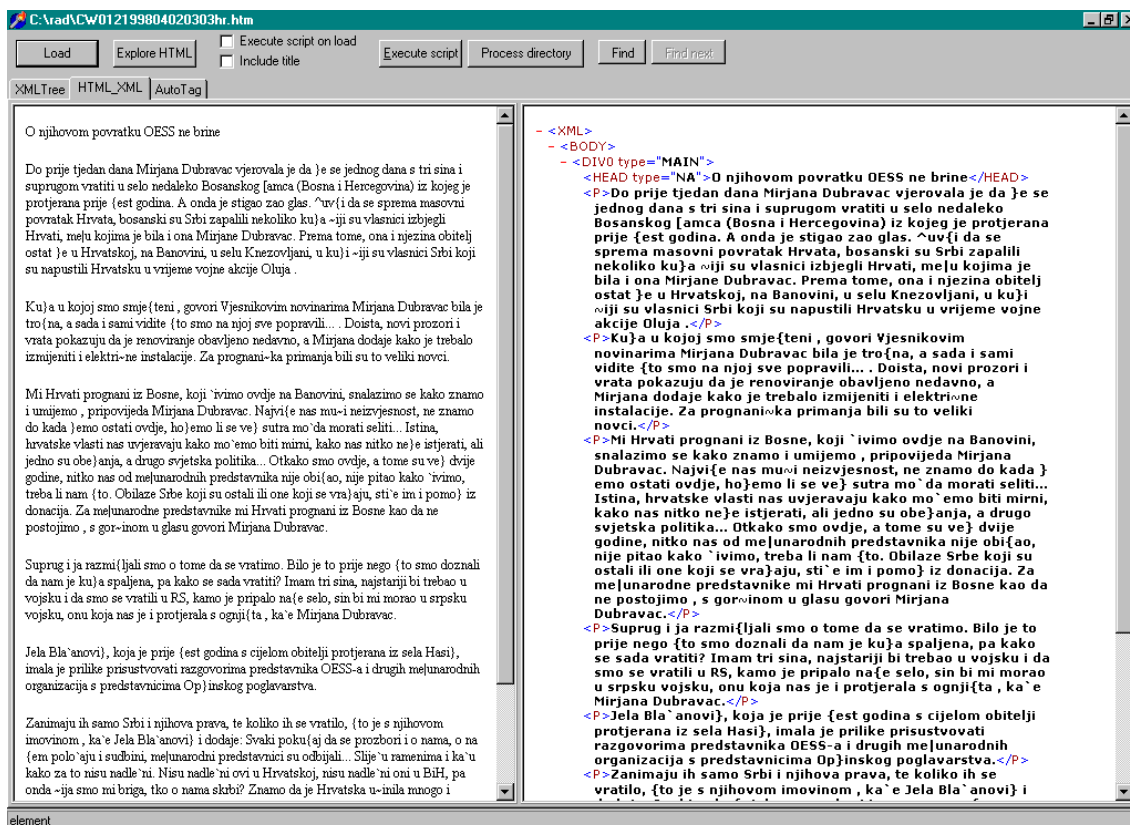


Figure 3.

All that has to be done after the conversion is to attach the header, and then the completely formatted XML document is ready for inclusion into the corpus. At the time of writing this paper, the 2XML application is in high beta stage.

3.4. Sentence delimiting

Sentence marking is accomplished by script applied by shareware Search&Replace V3.0 by Funduc Software Ltd. which allows regular expressions, scripts, etc. The </S><S> insertion is done in a familiar way, after punctuation followed by a capital letter. After that, output is filtered for exceptions like dr., prof., mr., ms., miss., ing., st., sv., initials, etc.

Tokenizer is another tool which comes in the bundle with 2XML. It analyzes XML input, delimits tokens, and flags them as R (= word), B (= number), I (= punctuation), X (= XML tag). Output can be a tabbed file for input in database (Fig. 4)

TOKEN	FILE NAME	BYTE OFFSET	FLAG
<BODY>	CW011199803260101hr	1	X
<DIV0 type="MAIN">	CW011199803260101hr	7	X
<HEAD type="NA">	CW011199803260101hr	25	X
<S>	CW011199803260101hr	41	X
Predsjednik	CW011199803260101hr	44	R
Tuđman	CW011199803260101hr	56	R
primio	CW011199803260101hr	63	R
Kinkela	CW011199803260101hr	70	R
,	CW011199803260101hr	77	I
Vedrinea	CW011199803260101hr	79	R
i	CW011199803260101hr	88	R
Primakova	CW011199803260101hr	90	R
</S>	CW011199803260101hr	99	X
</HEAD>	CW011199803260101hr	103	X
<HEAD type="PN">	CW011199803260101hr	110	X
<S>	CW011199803260101hr	126	X
Tuđman	CW011199803260101hr	129	R
:	CW011199803260101hr	135	I
Hrvatska	CW011199803260101hr	137	R
vojno	CW011199803260101hr	146	R
,	CW011199803260101hr	151	I
gospodarski	CW011199803260101hr	153	R
i	CW011199803260101hr	165	R
sigurnosno	CW011199803260101hr	167	R
orijentirana	CW011199803260101hr	178	R
na	CW011199803260101hr	191	R
europske	CW011199803260101hr	194	R
integracije	CW011199803260101hr	203	R
.	CW011199803260101hr	214	I
</S>	CW011199803260101hr	215	X
<S>	CW011199803260101hr	220	X
Ministri	CW011199803260101hr	223	R
Vedrine	CW011199803260101hr	232	R
i	CW011199803260101hr	240	R
Kinkel	CW011199803260101hr	242	R
uputili	CW011199803260101hr	249	R
zahtjev	CW011199803260101hr	257	R
Hrvatskoj	CW011199803260101hr	265	R
da	CW011199803260101hr	275	R
izradi	CW011199803260101hr	278	R
konkretan	CW011199803260101hr	285	R
plan	CW011199803260101hr	295	R
povratka	CW011199803260101hr	300	R
izbjeglica	CW011199803260101hr	309	R
.	CW011199803260101hr	319	I

Figure 4.

or tokenized in form like (Fig. 5):

```
<BODY><DIV0 type="MAIN"><HEAD type="NA"><S><W type="R">Predsjednik</W>
<W type="R">Tu&#273;man</W> <W type="R">primio</W>
<W type="R">Kinkela</W><W type="I">,</W> <W type="R">Vedrinea</W> <W type="R">i</W>
<W type="R">Primakova</W></S></HEAD>
<HEAD type="PN"><S><W type="R">Tu&#273;man</W><W type="I">:</W> <W type="R">Hrvatska</W>
<W type="R">vojno</W><W type="I">,</W> <W type="R">gospodarski</W> <W type="R">i</W>
<W type="R">sigurnosno</W> <W type="R">orijentirana</W> <W type="R">na</W>
<W type="R">europske</W> <W type="R">integracije</W><W type="I">.</W></S>
<S><W type="R">Ministri</W> <W type="R">Vedrine</W> <W type="R">i</W>
<W type="R">Kinkel</W> <W type="R">uputili</W> <W type="R">zahtjev</W>
<W type="R">Hrvatskoj</W> <W type="R">da</W> <W type="R">izradi</W>
<W type="R">konkretan</W> <W type="R">plan</W> <W type="R">povratka</W>
<W type="R">izbjeglica</W><W type="I">.</W></S>
```

Figure 5.

which is suitable for further processing. But prior to the word segmentation, sentence alignment has to be performed.

4. Aligning

Alignment at the sentence level is in the test stage. We are testing two alignment programs: a translation memory database system DéjàVu 2.3.82 by Atril and Vanilla aligner by Pernilla Danielson and Daniel Ridings (Danielsson and Ridings 1997).

4.1. Aligning with DéjàVu²

The demo version of the DéjàVu translation memory database system has a fully functional aligning module with a rather user friendly interface (Fig. 6.)

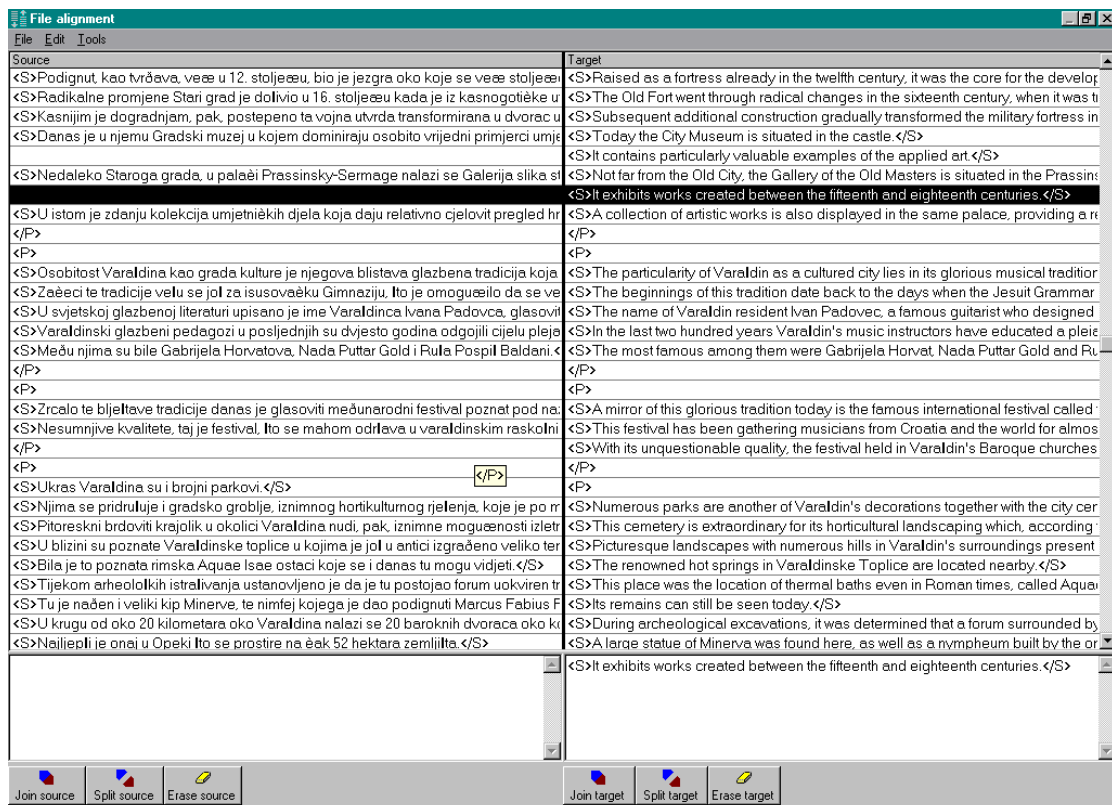


Figure 6.

Export from that translation memory database to TMX by means of a built-in export filter would look like in (Fig. 7).

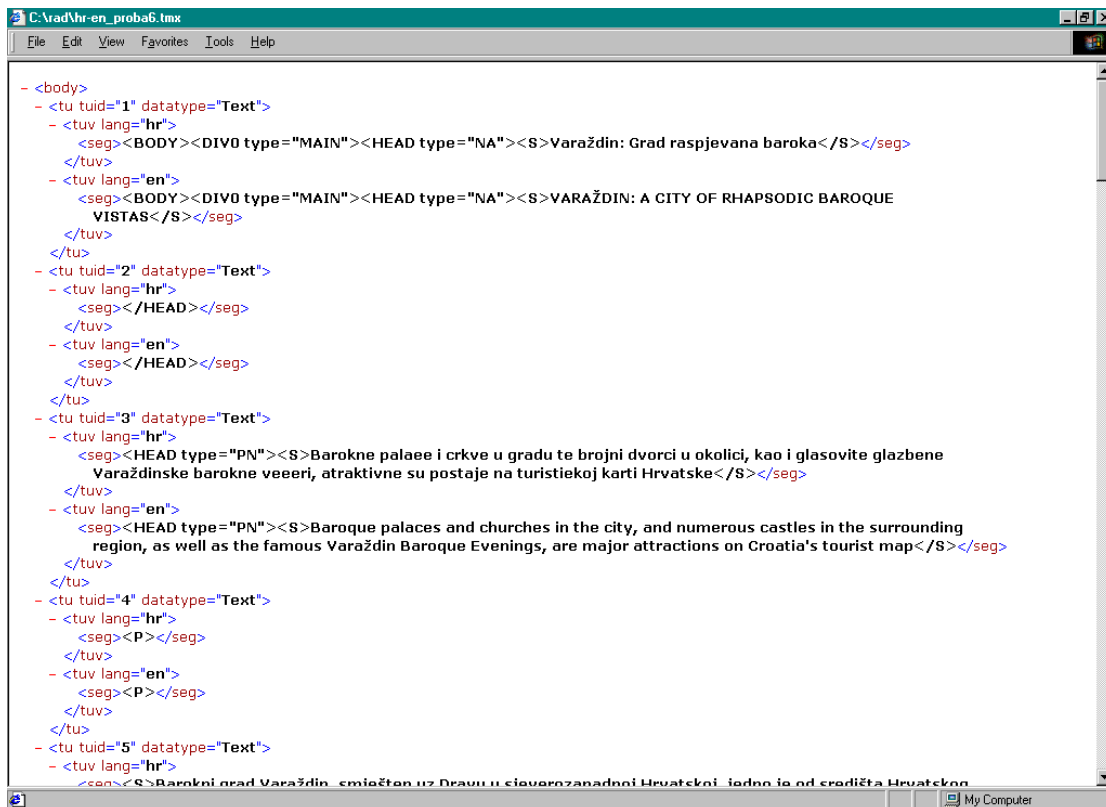


Figure 7.

Does it look OK? Definitely not, because all levels above <S> are incorporated in <TUV> and <SEG> elements and that is not what we would expect.

Besides, Figure 6 demonstrates that there is a lot of discrepancy in alignment between languages which requires a great deal of manual post-processing.

4.2. Aligning with Vanilla aligner³

Vanilla aligner (DOS version) gives better results with less alignment mistakes, even in one-to-many cases, but neither its interface is friendly nor is its output encoded the way we wanted (Fig. 8).

```

Cw_hr.txt - Notepad
File Edit Search Help
*** Länk: 1 - 1 ***
<BODY> <DIV0 type="MAIN"> <HEAD type="NA"> <S> Dvorci Hrvatskog zagorja : Romanti&#269;ni put u pro&#353;lost </S> </HEAD>
.EOS
<BODY> <DIV0 type="MAIN"> <HEAD type="NA"> <S> THE CASTLES OF ZAGORJE : A ROMANTIC TRIP TO THE PAST </S> </HEAD> .EOS
.EOP
*** Länk: 1 - 1 ***
<HEAD type="PN"> <S> Nigdje u Europi ne mo&#382;e se na tako malom prostoru na&#263;i tolika koncentracija , uglavnom ,
baroknih dvoraca </S> </HEAD> .EOS
<HEAD type="PN"> <S> Such a concentration of mostly Baroque castles on such a small area cannot be found anywhere else in
Europe </S> </HEAD> .EOS
.EOP
*** Länk: 1 - 1 ***
<P> <S> Hrvatsko zagorje prostor je u sjeverozapadnoj Hrvatskoj koje svojim osobitostima potvr&#273;uje iznimne
turisti&#269;ke mogu&#263;nosti . </S> .EOS
<P> <S> The region of Hrvatsko Zagorje is an area in northwest Croatia whose unique features confirm its extraordinary
tourism potential . </S> .EOS
*** Länk: 1 - 1 ***
<S> Blaga kontinentalna klima toga podru&#269;ja daje ugodno uto&#269;i&#353;te svima onima koji zaziru od prekomjerenih
ljetnih vru&#263;ina ili pak o&#353;trih zima . </S> .EOS
<S> The region's mild continental climate provides a pleasant refuge to all of those who are trying to escape from excessive
summer heat or harsh winters . </S> .EOS
*** Länk: 1 - 1 ***
<S> Osim prekrasnog bre&#382;uljkastog krajolika , kupice vina i zagorskih specijaliteta , Hrvatsko zagorje nudi turistima i
poprili&#269;an povijesno-umjetni&#269;ki inventar uglavnom baroknih dvoraca i ne&#353;to starijih pleni&#263;kih burgova .
</S> .EOS
<S> In addition to a beautiful landscape full of hills , a glass of wine and Zagorje culinary specialties , Zagorje offers
quite a number of historical and artistic sites - mostly Baroque castles and some older castles once held by the region's
nobility . </S> .EOS
*** Länk: 1 - 1 ***
<S> Po gusto&#263;i spomenika graditeljstva , nakon jadranskog prostora , to je najbogatije spomeni&#269;ko podru&#269;je
Hrvatske . </S> .EOS
<S> The sheer density of monuments makes this Croatia's richest heritage region after the Adriatic coast . </S> .EOS
*** Länk: 1 - 1 ***
<S> Me&#273;u stotinama za&#353;ti&#269;enih spomenika kulture , ovdje se nalazi devedesetak dvoraca i kurija . </S> .EOS
<S> Among its hundreds of protected cultural monuments , there are about ninety castles and mansions here . </S> .EOS
*** Länk: 1 - 1 ***
<S> Nigdje u Europi ne mo&#382;e se na tako malom prostoru na&#263;i tolika koncentracija dvoraca . </S> .EOS

```

Figure 8.

The same problem of higher element levels incorporated in aligned segments is still present. So we have...

4.3. Encoding problems

How to store alignments? Do we have a common way to encode them since we use XML? There is a number of ways how to do it now both in SGML and XML encoding:

1. Pointers stored in separate document:

1.1. Corpus encoding standard (Ide 1998 and CES⁴) are defined in SGML, with extensive use of ID attributes in <S> elements and pointers to them (example from CES 5.3.4.2):

```

DOC1: <s id=p1s1>According to our survey, 1988 sales of
mineral water and soft drinks were much higher than in 1987, reflecting
the growing popularity of these products.</s>
<s id=p1s2>Cola drink manufacturers in particular achieved above-
average growth rates.</s>

```

```
DOC2: <s id=pls1>Quant aux eaux minérales et aux limonades, elles
rencontrent toujours plus d'adeptes.</s>
<s id=pls2>En effet, notre sondage fait ressortir des ventes
nettement supérieures à celles de 1987, pour les boissons
à base de cola notamment.</s>
```

```
ALIGN DOC:
<linkGrp targType="s">
<link xtargets="pls1 ; pls1">
<link xtargets="pls2 ; pls2">
</linkGrp>
```

1.2. TEI Lite DTD was converted to XML in May 1999 by Patrice Bonhomme.⁵ Since we are using XML, this is the possible candidate for our encoding system. The usage of pointers to IDs and storage to different documents remains very much the same as in CES.

2. Translation memory (TMX⁶) inspired type of alignment encoding:

2.1. Since we have chosen XML, one would expect to use the PLUG project DTD⁷ which groups sentences in segments like the example from Tiedemann (1998:11):

```
<doc.body>
<align id='svenprf2' link='1-1'>
<seg lang='sv'>
<s>
Eders Majest&auml;ter, Eders Kungliga H&ouml;gheter, herr talman,
ledam&ouml;ter av Sveriges riksdag!
</s>
</seg>
<seg lang='en'>
<s>
Your Majesties, Your Royal Highnesses, Mr Speaker, Members of the
Swedish Parliament.
</s>
</seg>
</align>
```

The problem with this encoding system is that all upper levels of markup are lost since the <BODY> of the document is reorganized in a string of <ALIGN> elements. These elements further contain <SEG> elements which are actually aligned and accompanied with explicit language markers. Actual <S> elements are embedded in <SEGS>.

2.2. The ELAN Slovene-English parallel corpus⁸ was encoded in TEI SGML. The TEI <BODY> element was redefined to be a string of translation units (<TU> elements) which are formed by pairs of aligned <SEG> elements:⁹

```
<tu id="usta.14" lang="sl-en">
<seg lang="sl"><w>Slovenija</w> <w>je</w>
<w>ozemeljsko</w> <w>enotna</w> <w>in</w>
<w>nedeljiva</w> <w>dr&zcaron;ava</w><c>.</c>
</seg>
<seg lang="en"><w>Slovenia</w> <w>is</w>
```

```

<w>a</w> <w>territorially</w>
<w>indivisible</w> <w>state</w><c>.</c>
</seg>
</tu>

```

In this solution, it is important to notice that the <SEG> element is not composed of <S> but, unlike in the PLUG project, of <W> and <C> elements. The proper alignment between the sentences is not marked explicitly, but they are deductible from <SEG> opening and closing tags as well as from the <C> elements which could serve as the sentence-boundary markers in the case of alignment which is not one-to-one.¹⁰

Similar to the PLUG DTD, to which this solution also refers, all upper-level encoding (<DIVs>, <HEADs>, etc.) is lost.

Is there a way to keep aligned sentences together in the same element while retaining upper levels of text encoding? Could it be possible in the same document to have aligned only those parts of document structure which show actual translation and keep the rest of structure unique for both languages? Ideally that would look like a structure with preserved higher levels and aligned <SEG> elements just above the <S> level. That kind of encoding is certainly more readable for humans and needs less text storage (Fig. 9):

```

<DIV0 type="article">
  <HEAD type="NA">
    <ALIGN type="1-2">
      <SEG lang="hr">
        <S>Ovdje je re#269;enica 1 koja uklju#269;uje i 2.</S>
      </SEG>
      <SEG lang="en">
        <S>Here comes the sentence No 1.</S>
        <S>This is sentence No 2.</S>
      </SEG>
    </ALIGN>
    <ALIGN...> ...
  </ALIGN>
  ...
</HEAD>
<P>
  <ALIGN type="1-1">
    <SEG lang="hr">
      <S>Ovdje je re#269;enica 3.</S>
    </SEG>
    <SEG lang="en">
      <S>Here comes the sentence No 3.</S>
    </SEG>
  </ALIGN>
  <ALIGN...> ...
</ALIGN>
  ...
</P>
  ...
</DIV1>

```

Figure 9.

Although this kind of encoding looks attractive, there are several remarks, which can be said about it.

First of all, the DTD would have to be more complicated because the <ALIGN> element should be included in virtually any element which allows <P>. However, it stands in conflict with the general demand, formulated in CES, for keeping the original document unchanged as much as possible. That demand is even unavoidable with read-only source documents (see Thompson and McKelvie 1997).

Furthermore, the type of encoding shown in Figure 9 is actually redundant and can be generated from the documents encoded by the system mentioned in point 1.1. or 1.2. Since our Croatian-English Parallel Corpus project is at the beginning, the decision about the alignment encoding system remains to be made in the near future.

However, it seems that for <S> elements alignment we would have quite a lot of checking. The amount of “handwork” can be seen from statistics that show significant discrepancy in the number of <S> and <W> elements in Croatian and English:

		Hr	En	% increase
CW010:	<P>	195	195	
	<S>	729	796	9.2
	<W>	15483	18176	17.4
CW011:	<P>	178	178	
	<S>	675	754	11.7
	<W>	14853	17602	18.5
CW012:	<P>	174	174	
	<S>	652	733	12.4
	<W>	17317	20193	16.6
CW013:	<P>	174	174	
	<S>	679	767	13.0
	<W>	17163	19902	16.0
Avg.	<P>	180.25	180.25	
	<S>	683.75	762.50	11.5
	<W>	16204	18968.25	17.1

The first question coming to one's mind is: Is it a regular difference or the result of inadequate translation? The ELAN Slovene-English parallel corpus shows an even stronger tendency towards EN token prevalence: SI: 510 533 and EN: 632 218 meaning a 23.8% increase. The <S> correspondence between Slovene and English is also mentioned (SI: 25 572 and EN: 24 993 meaning a 2.3% decrease), but in Vintar (1999:64), it is not clear how those numbers were acquired. They could not have been investigated without a further sentence segmentation of the original corpus data because of the type of encoding used and described above in 2.2. Here the <S>-element Slovene-English correspondence is different from Croatian-English, and it is probably due to the fact that the Croatian-English corpus is collected from only one source while Slovene-English is compiled from 15 different text sources. It would be interesting to see data from other Slavic languages paired with English in a parallel corpus.

5. Conclusions

This paper presented the starting-point of the collecting and encoding of the Croatian-English Parallel corpus. As we proceed with the development of this language resource, where lack of Croatian language was more than evident, the referring data will be made available on http://www.hnk.ffzg.hr/hr-en_pcorp.

What is important at this point is the completion of the alignment along with the decision about its encoding. Further steps would be widening the corpus with texts from other sources and including the refined annotation, particularly at the <W> level. Lemmatization and MSD for English should not be a problem today, but for Croatian, we plan cooperation with our Croatian National Corpus project where the module for Croatian lemmatization and MSD annotation of corpora is being developed¹¹ in cooperation with MulTextEast V2 initiative.

6. Acknowledgements

The author would like to thank Ivana Simeon and Krešimir Šojat for work done in the process of converting the original files.

Thanks is due to the Croatian Institute for Culture and Information, the publisher of Croatia Weekly who provided us with source texts for this corpus.

7. Notes

1. The 'Serbo-Croatian', 'Croato-Serbian,' or 'Croatian or Serbian' was the official name for the Croatian language under communist authorities which tried to unify it with the Serbian language by force and to suppress any kind of Croatian language specifics which were considered dangerous for that unification process. The same name still persists in the Serbian part of former Yugoslavia and in many Slavistic handbooks. This name for the project was the only one allowed at that time.
2. Here I would like to express our thanks to Tomaž Erjavec and his colleagues from Ljubljana who gave us invaluable advice and tried to save us from wandering around. How much they succeeded in that matter is yet to be seen.
3. Here I would like to express our thanks to Milena Slavcheva, and the Mannheim TELRI team, who provided us with that software. See also <http://nl.ijs.si/telri/Vanilla>.
4. See <http://www.cs.vassar.edu/CES/>
5. See <http://www.loria.fr/~bonhomme/XML> and http://www.loria.fr/~bonhomme/xteelite-0_6.zip
6. See <http://www.lisa.unige.ch/tmx/>
7. In Tiedemann (1998:8). See also <http://numerus.ling.uu.se/~corpora/plug/>.
8. Erjavec (1999a:27). See also <http://nl.ijs.si/elan/>
9. Erjavec (1999b:4)
10. Part of the Slovene-English ELAN corpus, namely, the Orwell's *1984* component, has <S> elements marked inside <SEG> elements.
11. For the Croatian National Corpus visit <http://www.hnk.ffzg.hr>

8. References

Ahrenberg, Lars; Merkel, Magnus; Ridings, Daniel; Sågwall Hein, Anna; and Tiedemann, Jörg (1999) Automatic processing of parallel corpora: A Swedish perspective. (<http://numerus.ling.uu.se/~corpora/plug/>)

- Allen, Jeffrey and Hogan, Christopher. 1998. Expanding Lexical Coverage of Parallel Corpora. *First International Conference on Language Resources and Evaluation*, LREC'98. 747-754 Granada: ELRA.
- Bujas, Željko (1967) "Concordancing as a Method in Contrastive Analysis". *Studia Anglica et Romanica Zagrabiensia*, 23: 49-62.
- Bujas, Željko (1969) "Computers in the Yugoslav Serbo-Croatian/English Contrastive Analysis Project". *ITL Review for Applied Linguistics*, Spring 1969: 35-42.
- Bujas, Željko (1975) "Computers in the Yugoslav Serbo-Croatian — English Contrastive Project". *Bilten Instituta za lingvistiku Zagreb* 1: 44-58.
- Danielsson, Pernilla and Ridings, Daniel (1997) "Practical presentation of a "vanilla" aligner", ed. by Reyle, U. and Rohrer, C. Presented at the TELRI Workshop on Alignment and Exploitation of Texts. Institute Jožef Stefan, Ljubljana (<http://svenska.gu.se/PEDANT/workshop/workshop.html>).
- Erjavec, Tomaž ; Lawson, Ann; Romary, Laurent (1998) East meets West — A Compendium of Multilingual Resources. 2 CD-ROMs. Mannheim: TELRI-IDS
- Erjavec, Tomaž (1999a) "Making the ELAN Slovene/English Corpus". *Proceedings of the Workshop Language Technologies — Multilingual Aspects*, ed. by Vintar, Špela. 23-30. Ljubljana: Department of Translation and Interpreting, Faculty of Arts, Univ. of Ljubljana. (<http://nl.ijs.si/et>)
- Erjavec, Tomaž (1999b) "A TEI encoding of aligned corpora as translation memories". *Proceedings of the EACL-99 Workshop on Linguistically Interpreted Corpora (LINC-99)*, Bergen. ACL.
- Ide, Nancy (1998) "Corpus Encoding Standard: SGML guidelines for encoding linguistic corpora". First International Conference on Language Resources and Evaluation, LREC'98. 463-470 Granada: ELRA. (<http://www.cs.vassar.edu/CES/>)
- Thompson, Henry and McKelvie, David (1997) "Hyperlink semantics for standoff markup of read-only documents". SGML Europe'97. (<http://www.ltg.ed.ac.uk/~ht/sgmleu97.html>)
- Tiedemann, J (1998) Parallel corpora in Linköping, Uppsala and Göteborg (PLUG). Work package 1. Department of Linguistics, Uppsala University. (<http://numerus.ling.uu.se/~corpora/plug/>)

Vintar, Špela (1999) “A Lexical Analysis of the IJS-ELAN Slovene-English Parallel Corpus”. *Proceedings of the Workshop Language Technologies — Multilingual Aspects*, ed. by Vintar, Špela. 63-70. Ljubljana: Department of Translation and Interpreting, Faculty of Arts, Univ. of Ljubljana.