

Building the Croatian-English Parallel Corpus

Marko Tadić

Department of general linguistics and oriental studies, Faculty of philosophy, University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
marko.tadic@ffzg.hr

Abstract

The contribution gives a survey of procedures and formats used in building the Croatian-English parallel corpus which is being collected in the Institute of Linguistics at the Philosophical Faculty, University of Zagreb. The primary text source is newspaper *Croatia Weekly* which has been published from the beginning of 1998 by HIKZ (Croatian Institute for Information and Culture). After 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 elaborated. Preliminary statistics on numbers of <S> and <W> elements in each language is given at the end of the article.

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 choose English as a pair to the Croatian from the beginning.

Many scholars probably don't know that this very language pairing in parallel corpora started more than 30 years ago, as prof. Rudolf Filipović launched the *Yugoslav Serbo-Croatian—English Contrastive Project*¹ in 1968. The preliminary idea was brought to Zagreb by Željko 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 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 in 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 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 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 et al., 1998), although the Croatian-English lan-

guage pair is not the only one and it was certainly not of the primary interest. Since the whole work is well known, we will go on with our topic.

2. Corpus

Croatian-English parallel corpus, which is now being collected at the Institute of Linguistics at the Faculty of philosophy, University of Zagreb is the third 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 *Croatian-Slovene parallel corpus*, which was approved by both Ministries of science in July 1999 and was effectively launched in October 1999.

2.1. Representativeness issue

In corpora collecting 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. Situation is even worse in the case of parallel corpora since the demand for parallelism narrows the already limited choice of texts. Also as for languages with small number of speakers and/or translators such as Croatian, one can be happy to get any valuable translations. The outcome is usually rather unbalanced set of bitexts because you have to take whatever you can get in digital form. Wouldn't it be "methodologically cleaner" then 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.

2.2. Text source

The source of texts is the newspaper *Croatia Weekly*, being published by HIKZ (Croatian Institute for Culture and Information)² as from the beginning of 1998. The publication is sort of *USA Today* in Croatian way — it covers different domains: politics (internal and foreign), economy and finance, tourism, ecology, culture, art, sports and events and is intended for the public abroad. It is being published on 16 pages (including 4 pages for advertising) giving us an average of 16.200 tokens per

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 Serbian language by force and 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 textbooks. That name of the project was the only one allowed at that time.

² See <http://www.croatia.hr>.

issue for Croatian and 18.950 for English. The issue number 110 is just being published and we have access to the digital form of all texts in both languages except for first 5 issues. Thus having 100 issues leads us to approximately 1.6 Mw for Croatian and approximately 1.9 Mw for English.

The only problem which could cast a shade 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 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 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. Few years ago that 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 professional translations bureau, come to us in “bare ASCII” format, completely stripped off of any markup. Thus, as to 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.

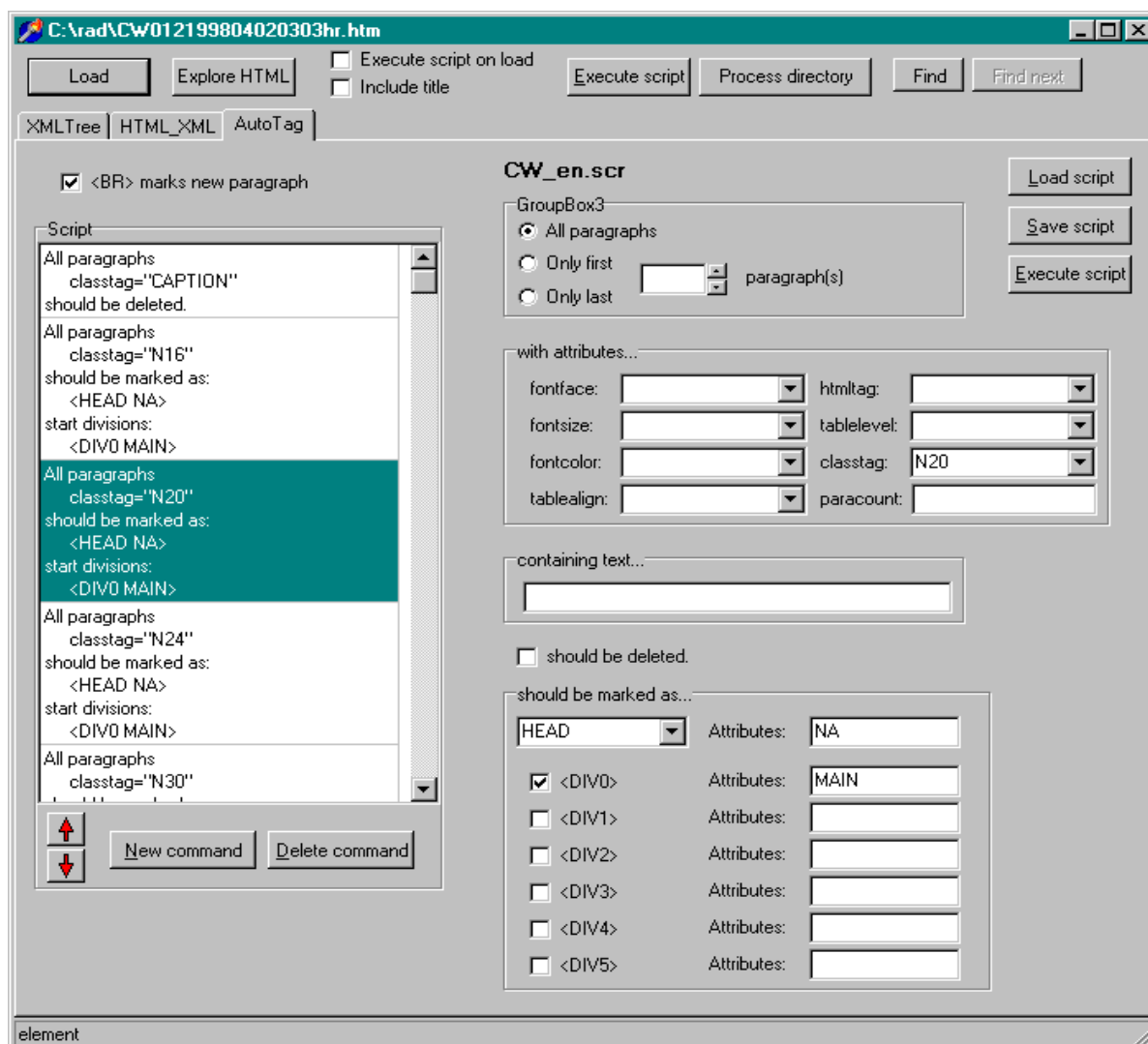


Figure 1: 2XML, Script editor page

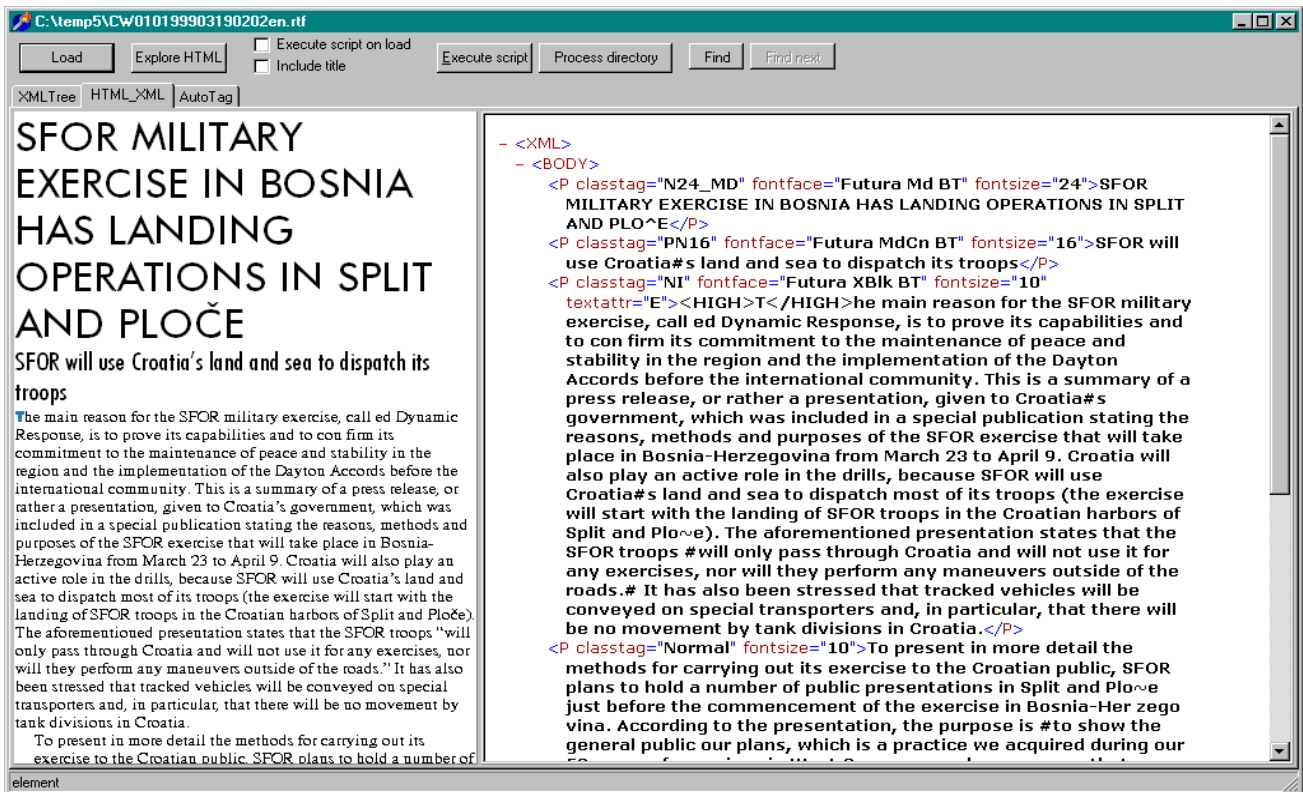


Figure 2: 2XML, the first step of conversion

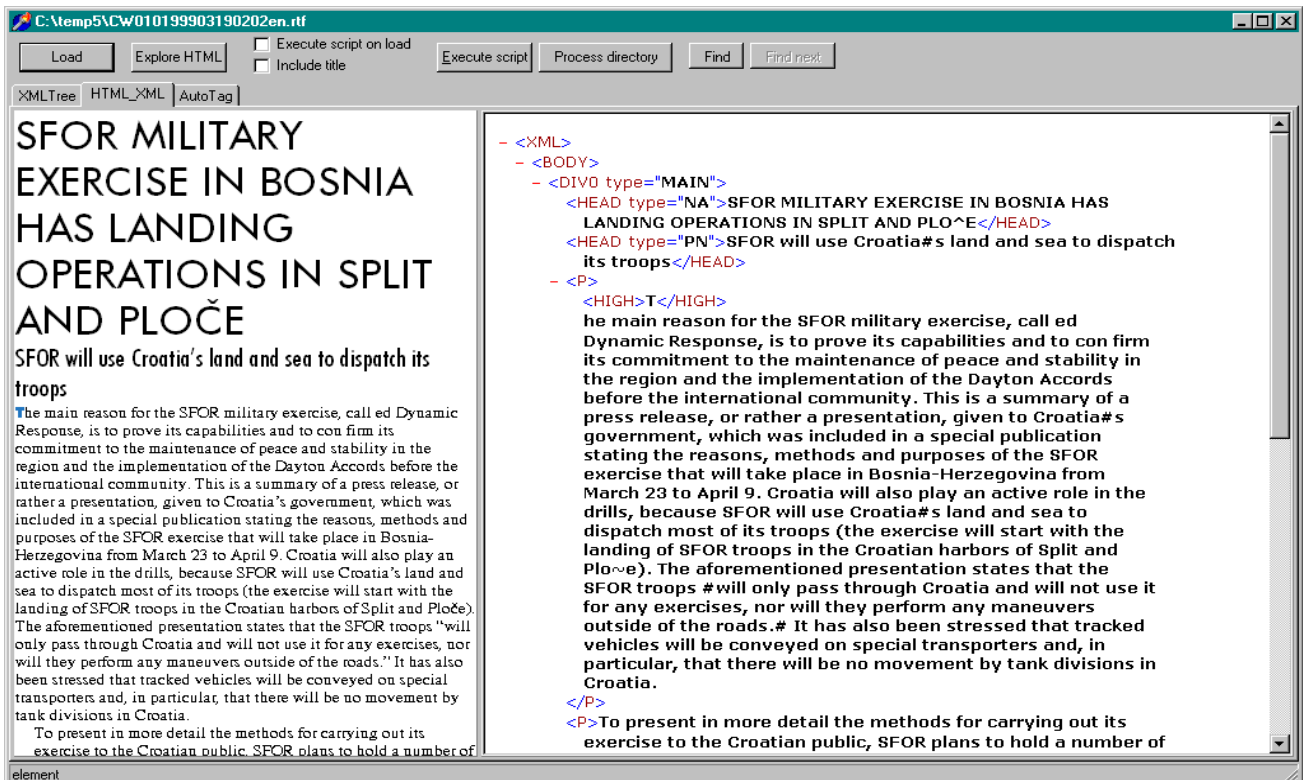


Figure 3: 2XML, the second step of conversion

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 (rtf style name, typeface name, font size, paragraph justification, style name etc.) are preserved (Figure 2 showing just few of them).

- 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 different <DIVs> and <HEADs> with their specific attribute values defined by script (Figure 3).

All that has to be done after the conversion and minor “cleaning”, is to attach the header and the completely formatted XML document is ready for inclusion into the corpus. The 2XML application is in prerelease stage, and it will be available soon.

3.4. Sentence delimiting

Sentence boundary markup is accomplished by means of a 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 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. The ordinal numbers represent particularly complex cases because by Croatian orthographic rules, ordinal numbers must be followed by period in order to be distinguished from cardinal numbers.

Thus, about 28% of arabic numerals written with period in Croatian texts are the sentence endings at the same time and that is something what had to be checked manually.

4. Aligning

Two aligning programs were used in the test stage of aligning on the sentence level. The first is a translation memory database system DéjàVu 2.3.82 by Atril, and the second is Vanilla aligner by Pernilla Danielson and Daniel Ridings (Danielsson & Ridings 1997).

4.1. Aligning with DéjàVu

The demo version of DéjàVu translation memory database system has a fully functional aligning module with a rather friendly user interface. Export from that translation memory database to TMX format by means of a built-in export filter would yield a result which looks like this:

```

- <body>
- <tu tuid="1" datatype="Text">
- <tuv lang="hr">
- <seg><BODY><DIV0 type="MAIN"><HEAD type="NA"><S>Varaždin: Grad raspjevana baroka</S></seg>
</tuv>
- <tuv lang="en">
- <seg><BODY><DIV0 type="MAIN"><HEAD type="NA"><S>VARAŽDIN: A CITY OF RHAPSODIC BAROQUE
VISTAS</S></seg>
</tuv>
</tu>
- <tu tuid="2" datatype="Text">
- <tuv lang="hr">
- <seg></HEAD></seg>
</tuv>
- <tuv lang="en">
- <seg></HEAD></seg>
</tuv>
</tu>
- <tu tuid="3" datatype="Text">
- <tuv lang="hr">
- <seg><HEAD type="PN"><S>Barokne palaee i crkve u gradu te brojni dvorci u okolici, kao i glasovite glazbene
Varaždinske barokne veeeri, atraktivne su postaje na turistiekoj karti Hrvatske</S></seg>
</tuv>
- <tuv lang="en">
- <seg><HEAD type="PN"><S>Baroque palaces and churches in the city, and numerous castles in the surrounding
region, as well as the famous Varaždin Baroque Evenings, are major attractions on Croatia's tourist map</S></seg>
</tuv>
</tu>
- <tu tuid="4" datatype="Text">
- <tuv lang="hr">
- <seg><P></seg>
</tuv>
- <tuv lang="en">
- <seg><P></seg>
</tuv>
</tu>
- <tu tuid="5" datatype="Text">
- <tuv lang="hr">
- <seg><S>Barokni grad Varaždin, smješten uz Dravu u sjeverozapadnoj Hrvatskoj, jedno je od središta Hrvatske

```

Figure 4: TMX export from DéjàVu alignment module

Figure 4 clearly shows that this output is not immediately usable because all levels above <S> are incorporated in <TUV> and <SEG> elements and that is not what we would expect.

Besides, there is a lot of discrepancy in alignment between languages, which requires a lot of manual post-processing.

4.2. Aligning with Vanilla aligner

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 (see Figure 5).

```

*** 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 potur&#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 utu&#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&#263;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 5: Vanilla aligner, alignment with upper levels included

The same problem of higher element levels incorporated in aligned segments is still present. So we may say that we encountered the...

4.3. Encoding problem

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

1. Alignment by storing pointers in separate document
 - 1.1. Corpus encoding standard (Ide 1998 and CES³) 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=pls1>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=pls2>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.⁴ Since we are using XML, this is the possible candidate for our encoding system.
- 1.3. In February 2000 the beta version of XCES (XML version of CES) has been announced.⁵ with XSL stylesheets for cesAna and cesAlign still under development. It seems that the usage of pointers to IDs and storing of alignment information to separate document 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 the usage of the PLUG project DTD,⁷ which groups sentences in segments like in the example from Tiedemann (1998:11):

```

<doc.body>
<align id='svenprf2' link='1-1'>
<seg lang='sv'>

```

⁴ See <http://www.loria.fr/~bonhomme/XML> and http://www.loria.fr/~bonhomme/xteelite-0_6.zip

⁵ See <http://www.cs.vassar.edu/XCES>

⁶ See <http://www.lisa.unige.ch/tmx/>

⁷ In Tiedemann (1998:8). See also <http://numerus.ling.uu.se/~corpora/plug/>

³ See <http://www.cs.vassar.edu/CES/>

```

<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 that 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 <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 from the <C> elements which could serve as the sentence-boundary markers in the case when alignment is not one-to-one.¹⁰

But like in 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. It could look like this:

```

<DIV0 type="article">
<HEAD type="NA">
<ALIGN type="1-2">
<SEG lang="hr">
<S>Ovdje je re&#269;enica 1 kao i broj 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>
...
</DIV0>

```

Although this kind of encoding looks attractive there are several remarks which could 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>. Besides, it stands in the 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 & McKelvie 1997).

Furthermore, the type of encoding shown in example above is actually redundant and can be generated from the documents encoded by the system mentioned in the point 1.1. to 1.3. above. That is why we decided to use that system of alignment encoding:

```

DOC 1:
<DIV0 type="MAIN">
<HEAD type="NA">
<S id="CW010199803190201hr.S1">Do 1. kolovoza
zabranjeni skupovi u ...</S></HEAD>
<HEAD type="PN">
<S id="CW010199803190201hr.S2">Vlada je
ocijenila kako je provo&#273;enje mirne ...</S>
<S id="CW010199803190201hr.S3">Stoga, treba
izbje&#263;i svaki &#269;in koji ...</S></HEAD>
<P>
<S id="CW010199803190201hr.S4">Vlada Republike
Hrvatske obvezala je ...</S> ...</P> ... </DIV0>

```

```

DOC 2:
<DIV0 type="MAIN">
<HEAD type="NA">
<S id="CW010199803190201en.S1">POLITICAL
RALLIES ...</S> </HEAD>
<HEAD type="PN">

```

⁸ Erjavec (1999a:27). See also <http://nl.ijs.si/elan/>

⁹ Erjavec (1999b:4)

¹⁰ Part of the Slovene-English ELAN corpus, namely *Orwell 1984* component, has <S> elements marked inside <SEG> elements.

```

<S id="CW010199803190201en.S2">The Government
has assessed that the ...</S> </HEAD>
<P>
<S id="CW010199803190201en.S3">The Croatian
Government has charged ...</S> ... </P> ... </DIV0>

```

DOC 3 (Alignment):

```

<link xtargets="CW010199803190201hr.S1 ;
                CW010199903190201en.S1">
<link xtargets="CW010199803190201hr.S2
                CW010199803190201hr.S3 ;
                CW010199903190201en.S2">
<link xtargets="CW010199803190201hr.S4 ;
                CW010199903190201en.S3">

```

The ID attributes are quite exhaustive: CW010 gives the issue number, 19980319 the date, 02 page number, 01 number of text on that page, hr/en encodes language, S gives the number of sentence in that <DIV> element.

5. Preliminary statistics

By simple count of elements in several newspaper issues, it seems that for <S> elements aligning we would have quite a lot of checking. The amount of “handwork” can be estimated from preliminary statistics that show significant discrepancy in number of <S> as well as <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>	652	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%

Table 1: Number of <P>, <S> and <W> elements in four issues of *Croatia Weekly*

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 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: 25572 and EN: 24993 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 point 2.2. Here the <S>-element Slovene-English correspondence is different from Croatian-English and that is probably due to the fact that Croatian-English corpus is collected from only one source while Slovene-English is compiled from 15 different text

sources. Anyway, it would be interesting to see data from other Slavic languages paired with English.

6. Conclusions

The starting-point of the collecting and encoding of the Croatian-English Parallel corpus has been presented. As we proceed with development of this language resource, which lack for Croatian language was more than evident, the referring data will be made available on <http://www.hnk.ffzg.hr/pcorp>.

What is important at this point is the completion of the alignment. 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 nowadays but as to Croatian, we plan the cooperation with our Croatian National Corpus¹¹ project where the module for Croatian lemmatization and MSD annotation of corpora is being developed in cooperation with MulTextEast V2 initiative.

7. 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.

8. References

- Ahrenberg, Lars; Merkel, Magnus; Ridings, Daniel; Sågvall Hein, Anna & Tiedemann, Jörg. (1999). Automatic processing of parallel corpora: A Swedish perspective. (<http://numerus.ling.uu.se/~corpora/plug/>)
- Bujas, Željko. (1967). Concordancing as a Method in Contrastive Analysis, *Studia Anglica et Romanica Zagabiensia*, 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 & Ridings, Daniel. (1997). Practical presentation of a “vanilla” aligner. In U. Reyle & C. Rohrer (Eds.), 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. In Špela Vintar (Ed.), Proceedings of the workshop Language technologies — Multilingual Aspects, (pp. 23--30). Ljubljana: Department of Translation and Interpreting, Faculty of Arts, Univ. of Ljubljana.
- Erjavec, Tomaž. (1999b). A TEI encoding of aligned corpora as translation memories. In Proceedings of the EACL-99 Workshop on Linguistically Interpreted Corpora (LINC-99), Bergen: ACL.

¹¹ For the Croatian National Corpus see <http://www.hnk.ffzg.hr>

- Ide, Nancy. (1998). Corpus Encoding Standard: SGML guidelines for encoding linguistic corpora. In Proceedings of the First International Conference on Language Resources and Evaluation, LREC'98. (pp. 463--470) Granada: ELRA.
- Thompson, Henry & McKelvie, David. (1997). Hyperlink semantics for standoff markup of read-only documents. In 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. In Špela Vintar (Ed.), Proceedings of the workshop Language technologies — Multilingual Aspects, (pp. 63--70). Ljubljana: Department of Translation and Interpreting, Faculty of Arts, Univ. of Ljubljana.