EVALUATING SENTENCE ALIGNMENT ON CROATIAN-ENGLISH PARALLEL CORPORA

Sanja Seljan*, Željko Agić*, Marko Tadić**

*Department of Information Sciences
**Department of Linguistics
Faculty of Humanities and Social Sciences
University of Zagreb
Ivana Lučića 3, HR-10000 Zagreb
{sanja.seljan,zeljko.agic,marko.tadic}@ffzg.hr

ABSTRACT

This paper describes an experiment in applying sentence alignment methods to Croatian-English parallel corpora and systematically evaluate their performance within the recall, precision and F-measure framework. It is our primary goal to provide an insight and a reference point on sentence alignment accuracy for Croatian-English language pair and also to extend the scope of (Tadić, 2000) – to our knowledge, the first experiment dealing with automatic sentence alignment of Croatian-English parallel corpora – by utilizing newly implemented tools, creating corpora subsets defined by genre and finally by expanding and formalizing its preliminary observations on alignment accuracy. Therefore, in this paper we start off by briefly describing and argumenting sentence alignment paradigms of choice and presenting available language resources, subset of Croatian-English parallel corpus described in (Tadić, 2000) being our primary asset. These descriptions are followed by a formal definition of our testing framework. Results are then discussed in detail and conclusions are stated along with a brief insight on possible future work.

1. Introduction

Parallel corpora and especially sentence-aligned bilingual corpora can be very effectively used as a resource for numerous research projects or in creation of new resources, such as sentence or word alignment projects for computer-assisted translation, machine translation, multilingual information retrieval, language learning, multilingual terminology bases and semantic networks. Translation memories used in computer-assisted translation, machine translation or for terminology extraction, created by the sentence alignment process from the parallel corpus, directly reflect two main problems: corpus size and divergences in the layout of parallel texts, which can differ regarding the expert intervention in the set up of the alignment program. For this reason, corpus aligning has caused a great interest and numerous aligners relying on different methods.

For lesser spread languages, but still having rich cultural and historic pool of texts, creation of electronic tools and resources is of considerable interest. In the situation when Croatia approaches the European Union, use of common resources has become important not only for translators, but also for researchers and everyday users. Use of common resources and translation tools has become an absolute demand in any kind of cooperation activities and international communication.

One of the most well known examples of shared resources is Europarl: the European Parliament corpus, published on the web (URL http://www.statmt.org/europarl), which has numerous applications in NLP, significant because of its size, number of languages and various linguistic phenomena included, but also because of its provenance since the sources included are mainly from the United Nations, European Union or member governments (Koehn, 2005).

The JRC-Acquis Communautaire corpus and its documentation, consisting of 20+ languages are freely available from the web (at URL http://wt.jrc.it/lt/Acquis). This most voluminous multilingual parallel corpus, consisting of 22 languages, is especially suitable for cross-information language retrieval, machine translation and machine learning. Part of it is sentence-aligned using the Vanilla aligner, applying Gale-Church algorithm (Gale & Church, 1993) and HunAlign (Varga et al., 2005), which allows benchmarking of alignment tools and algorithms.

Therefore, the first experiment dealing with evaluation metrics of sentence-aligned Croatian-English corpora has been made and is described in this paper. As sentence-aligned corpora are more efficient and generally more valuable as a resource than non-aligned parallel corpora, while planning this investigation, we considered presenting results of automatic sentence-level alignment of Croatian-English parallel corpora using several different sentence alignment tools, namely SDL Trados WinAlign, Vanilla aligner (Danielsson & Ridings, 1997), Hunalign (Varga et al., 2005) and CORAL aligner in order to derive a decision on which of these tools should be utilized in aligning corpora of language pairs Croatian-Language$_x$ in the future. We also considered Moore's aligner (Moore, 2002) as one of possible solutions but since it was built with the main purpose of compiling the language models for SMT, which feature only exact (one-to-one) alignments and avoiding other types of alignments (non-one-to-one), we shifted our interest towards other algorithms. Most of them were discussed and compared in (Och & Ney, 2003) but what we actually wanted to put in the focus of our interest were the implementations. However, being that various sentence alignment tools are in fact based on underlying paradigms provided in forms of published or proprietary algorithms, we chose the scientific method and approach and decided to evaluate sentence alignment of Croatian-English language pair on paradigms exclusively, focusing on Gale-Church algorithm implemented in CORAL aligner.

In the following section of the paper, we present in more details these tools and resources utilized in the experiment and provide argumentation on choices made. Section 3 provides insight on test environment setup, while results are discussed and future research paths indicated in section 4.

2. Resources and tools

In order to provide a measure sentence alignment accuracy for Croatian-English language pair, there are two basic assets required – the language sample, i.e. manually (or otherwise) previously aligned parallel corpora and a sentence aligner, i.e. a program implementing an algorithm that automatically aligns non-aligned corpora. Manual and automatically aligned corpora are then compared and various methods of inspecting differences provide measures of alignment accuracy. In this section, we provide insight on these two basic assets while the following section focuses on evaluation framework.

Aligned subsets Croatian-English parallel corpora used in this research consist of:

1. Croatian-English parallel corpus of legislative documents (JOC – Journal of the European Community) and
2. Subset of Croatian-English parallel corpus from the newspaper Croatia Weekly, presented in (Tadić, 2000).

The legislative Croatian-English subset consists out of 6 Croatian legislative documents of about 20 pages consisting of bylaws, regulations, and decisions of the Croatian government and their translation into English documents. The documents consist of 6791 words in Croatian and corresponding 8510 words in English, i.e. 48982 characters in Croatian and corresponding 55567 characters in English due to highly inflective nature of the Croatian language. Documents were provided in plain text format, manually aligned using CORAL aligner and used as a reference point for automatic alignment evaluation. Document stats are given in Table 1.

	Pages		Words		Characters	
document	CRO	ENG	CRO	ENG	CRO	ENG
Bylaws	6	5.5	1311	1982	9238	12301
AMI	4	4.5	1577	1805	11143	11460
e-Sig	10.5	9	3903	4723	28601	31806
Total	20.5	19	6791	8510	48982	55567

Table 1. Legislative documents subcorpus statistics

Legislative documents are included in the experiment because various statistical machine translation platforms – relying on sentence- and word-alignment preprocessing – are largely utilized exactly in tasks of translating legal documents i.e. in multilingual environment of the European Union. It was therefore important to provide results of aligning Croatian and English in this domain in order to indicate the quality of sentence alignment platform on which future research is to build machine(-aided) translation systems.

However, regardless of overall importance of achieving high alignment accuracy on legislative documents, hand-annotated subset of Croatian-English parallel corpus was the main resource for this specific investigation, being linguistically well-formed and properly annotated by XML structure. The parallel corpus itself was previously described in detail (Tadić, 2000) as being sourced from the Croatia Weekly newspaper corpus and consisting of approximately 1.6 million tokens for Croatian and 1.9 million tokens for English.

Stats given in Table 2 indicate that subset size is approximately 32% of the entire parallel corpus when comparing token counts. As expected, both Croatian and English parts of the parallel subcorpus consist of same numbers of articles, sections, main titles, subtitles and paragraphs. Minor differences in section and main title counts are caused by human annotation errors as numbers match exactly when checking for specific errors; i.e. section count for Croatian lacks one section that is easily found if <DIV0> is replaced by <DIV in search query, clearly indicating a mistyped tag. Findings are similar for main title counts and a claim can be made that counts of document structure entities are the same.

Croatian	English	
1435	1435	Articles </BODY>
1599	1600	Sections </DIV0>
1597	1600	Main titles <HEAD type='NA'>
493	493	Subtitles <HEAD type='PN'>
6327	6327	Paragraphs </P>
22985	25412	Sentences </S>
514428	618462	Words
3402532	3300409	Characters
3918959	3920825	Characters with spaces
4913825	5017745	Bytes

Table 2. Cro-Eng subcorpus statistics

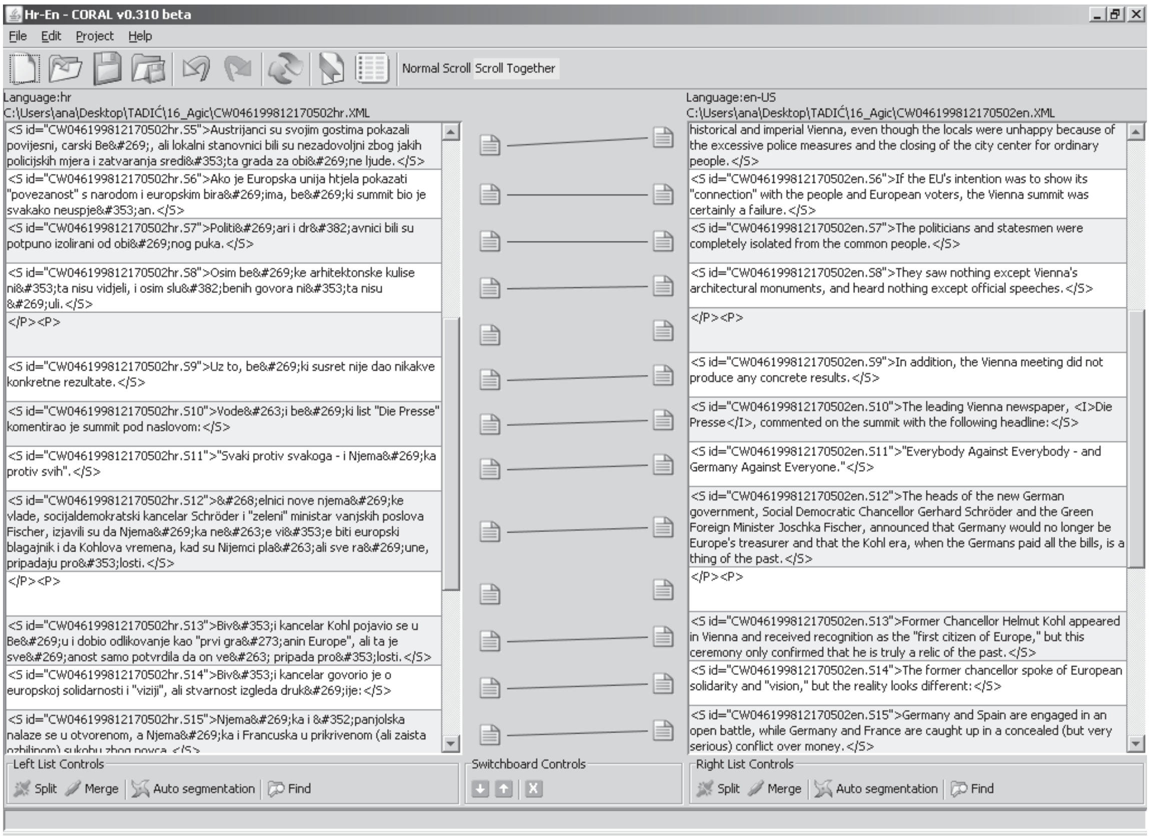
When comparing occurrence counts for sentences, words and characters, Croatian and English subcorpora start to differ, English getting higher numbers in all figures. This is an expected distribution and desired behaviour as structure of newspaper articles – captured by rows 1 to 5 of Figure 1 – remains the same in Croatian and English, while translation functions for sentences are never bijective. This conclusion also applies on word selection and subsequently character counts. As with the entire corpus, manually annotated gold standard subcorpus implements higher token count on English side with a factor of 1.2 compared to 1.19 overall. Subcorpus sample given in Figure 1 contains XML-wrapped article in English and Croatian in which structural and textual similarities and differences are illustrated.

```
<BODY><DIV0 type="MAIN"><HEAD type="NA">

<S id="CW047199812241407hr.S1">Bestseler u Aucklandu</S>

</HEAD><P>

<S id="CW047199812241407hr.S2">Roman "Croatia Mine" novozelandske Hrvatice Floride Vele
za kratko je vrijeme postao bestselerom u Aucklandu, a autorica se ove godine našla
među 20 odabranih pisaca tamošnje utjecajne književne priredbe "World Book Day".</S>

<S id="CW047199812241407hr.S3">Roman "Croatia Mine" (u izdanju Quin Pressa iz
Christchurcha) prvi je roman Floride Vele i nadahnut je stanovitim autobiografskim
elementima.</S>

<S id="CW047199812241407hr.S4">Prati sudbinu jedne hrvatske obitelji iz Podogore - od
iseljeništva iz Jugoslavije pa do snova i ljubavi prema dalekoj domovini Hrvatskoj.</S>

</P>

<BYLINE>(Večernji list)</BYLINE></DIV0></BODY>
```

```
<BODY><DIV0 type="MAIN"><HEAD type="NA">

<S id="CW047199812241407en.S1">BOOK BY CROATIAN AUTHORS BECOMES BESTSELLER IN
AUCKLAND</S></HEAD>

<P>

<S id="CW047199812241407en.S2">The novel <I>Croatia Mine</I>, by Florida Vela, a Croat
from New Zealand, quickly became the bestseller in Auckland.</S>

<S id="CW047199812241407en.S3">The author was placed on the list of twenty select
writers of World Book Day, an influential local literary event. <I>Croatia Mine</I>
(published by Quin Press from Christchurch) is the first novel by Florida Vela, and it
was inspired by certain autobiographical elements.</S>

<S id="CW047199812241407en.S4">It follows the fate of a Croatian family from Podogora -
from their emigration from Yugoslavia to their later dreams and yearning for their
distant homeland of Croatia.</S>

</P>

<BYLINE>(<I>Večernji list</I>)</BYLINE></DIV0></BODY>
```

Figure 1. Croatian-English parallel corpus sample

Second choice was that of sentence aligner. As stated in first section, among many alignment tools implementing many standard and specialized algorithms, we chose to provide figures on Croatian-English pair using a less-known aligner named CORAL (CORpus ALigner), being developed in Java at the Faculty of Electrical Engineering and Computing, University of Zagreb. There are two main reasons behind this choice: one is that it implements a standard Gale-Church algorithm that we wanted to evaluate and the other is encompassed by joint programme *Computational Linguistic Models and Language Technologies for Croatian* and its goals described in (Dalbelo Bašić et al., 2007): CORAL aligner is envisioned to be a default platform for sentence alignment (automatic and human assisted) of language pairs Croatian-Language$_X$ and thorough evaluation is required in order to develop newer and better versions of the tool. CORAL was previously evaluated on English-Slovene parallel corpus extracted from MULTEXT-East v3 specification (Erjavec, 2004) but was not yet presented to the community by the time this paper was published.


Hr-En - CORAL v0.310 beta
File Edit Project Help
Normal Scroll Scroll Together
Left List Controls
Split Merge Auto segmentation Find
Switchboard Controls
Right List Controls
Split Merge Auto segmentation Find

```
Language:hr
C:\Users\ana\Desktop\TADIĆ\16_Agic\CW046199812170502hr.XML
<S id="CW046199812170502hr.S5">Austrijanci su svojim gostima pokazali povijesni, carski Be&#269;, ali lokalni stanovnici bili su nezadovoljni zbog jakih policijskih mjera i zatvaranja sredi&#353;ta grada za obi&#269;ne ljude.</S>
<S id="CW046199812170502hr.S6">Ako je Europska unija htjela pokazati "povezanost" s narodom i europskim bira&#269;ima, be&#269;ki summit bio je svakako neuspje&#353;an.</S>
<S id="CW046199812170502hr.S7">Politi&#269;ari i dr&#382;avnici bili su potpuno izolirani od obi&#269;nog puka.</S>
<S id="CW046199812170502hr.S8">Osim be&#269;ke arhitektonske kulise ni&#353;ta nisu vidjeli, i osim slu&#382;benih govora ni&#353;ta nisu &#269;uli.</S>
</P><P>
<S id="CW046199812170502hr.S9">Uz to, be&#269;ki susret nije dao nikakve konkretne rezultate.</S>
<S id="CW046199812170502hr.S10">Vode&#263;i be&#269;ki list "Die Presse" komentirao je summit pod naslovom:</S>
<S id="CW046199812170502hr.S11">"Svaki protiv svakoga - i Njema&#269;ka protiv svih".</S>
<S id="CW046199812170502hr.S12">&#268;elnici nove njema&#269;ke vlade, socijaldemokratski kancelar Schröder i "zeleni" ministar vanjskih poslova Fischer, izjavili su da Njema&#269;ka ne&#263;e vi&#353;e biti europski blagajnik i da Kohlova vremena, kad su Nijemci pla&#263;ali sve ra&#269;une, pripadaju pro&#353;losti.</S>
</P><P>
<S id="CW046199812170502hr.S13">Biv&#353;i kancelar Kohl pojavio se u Be&#269;u i dobio odlikovanje kao "prvi gra&#273;anin Europe", ali ta je sve&#269;anost samo potvrdila da on ve&#263; pripada pro&#353;losti.</S>
<S id="CW046199812170502hr.S14">Biv&#353;i kancelar govorio je o europskoj solidarnosti i "viziji", ali stvarnost izgleda druk&#269;ije:</S>
<S id="CW046199812170502hr.S15">Njema&#269;ka i &#352;panjolska nalaze se u otvorenom, a Njema&#269;ka i Francuska u prikrivenom (ali zaista ozbiljnom) sukobu zbog novca.</S>
```

```
Language:en-US
C:\Users\ana\Desktop\TADIĆ\16_Agic\CW046199812170502en.XML
historical and imperial Vienna, even though the locals were unhappy because of the excessive police measures and the closing of the city center for ordinary people.</S>
<S id="CW046199812170502en.S6">If the EU's intention was to show its "connection" with the people and European voters, the Vienna summit was certainly a failure.</S>
<S id="CW046199812170502en.S7">The politicians and statesmen were completely isolated from the common people.</S>
<S id="CW046199812170502en.S8">They saw nothing except Vienna's architectural monuments, and heard nothing except official speeches.</S>
</P><P>
<S id="CW046199812170502en.S9">In addition, the Vienna meeting did not produce any concrete results.</S>
<S id="CW046199812170502en.S10">The leading Vienna newspaper, <I>Die Presse</I>, commented on the summit with the following headline:</S>
<S id="CW046199812170502en.S11">"Everybody Against Everybody - and Germany Against Everyone."</S>
<S id="CW046199812170502en.S12">The heads of the new German government, Social Democratic Chancellor Gerhard Schröder and the Green Foreign Minister Joschka Fischer, announced that Germany would no longer be Europe's treasurer and that the Kohl era, when the Germans paid all the bills, is a thing of the past.</S>
</P><P>
<S id="CW046199812170502en.S13">Former Chancellor Helmut Kohl appeared in Vienna and received recognition as the "first citizen of Europe," but this ceremony only confirmed that he is truly a relic of the past.</S>
<S id="CW046199812170502en.S14">The former chancellor spoke of European solidarity and "vision," but the reality looks different:</S>
<S id="CW046199812170502en.S15">Germany and Spain are engaged in an open battle, while Germany and France are caught up in a concealed (but very serious) conflict over money.</S>
```

Figure 2. CORAL screenshot

3. Evaluation method

Evaluation method used in this experiment was highly influenced by the one presented in (Langlais et al., 1998), i.e. set of methods used during the ARCADE text alignment evaluation project we chose as a starting point for our own experiment. Figure 3 provides an example on which evaluation techniques are demonstrated.

A_R	s_1	*Ovo je prva rečenica.*	t_1	*This is the first sentence.*
	s_2	*Ovo je druga rečenica i nalik je prvoj*	t_2	*This is the second sentence.*
			t_3	*It looks like the first.*

A_S	s_1	*Ovo je prva rečenica.*	t_1	*This is the first sentence.*
			t_2	*This is the second sentence.*
	s_2	*Ovo je druga rečenica i nalik je prvoj.*	t_3	*It looks like the first.*

Figure 3. Example reference and system alignment

Formal description is as follows. We consider source text S and output text T as a sequence of alignments {s1, …, sn} and {t1, …, tm}, respectively. This basic setting is also shown by Figure 3. An alignment A is then defined rather straightforward as a subset of Cartesian product of power-sets 2S × 2T. We then call the 3-tuple (S, T, A) a bitext and each of its elements is called a bisegment. Given these definitions, we set up two basic evaluation methods and consider two additional tweak or helper methods as proposed by (Langlais et al., 1998).

3.1. Basic F1-measure

Recall and precision are easily defined on a bitext:

$$recall = \frac{|A_S \cap A_R|}{|A_R|}, \quad precision = \frac{|A_S \cap A_R|}{|A_S|}$$

Being that recall basically measures coverage alone and precision deals with counting correct hits, F1-measure (Rijsbergen, 1979) is chosen for merging these two outputs:

$$F_1 measure = 2\ \frac{recall \times precision}{recall + precision}$$

On example in Figure 2, the measures calculate as follows:

$A_R = \{(\{s_1\}, \{t_1\}), (\{s_2\}, \{t_2, t_3\})\}$

$A_S = \{(\{s_1\}, \{t_1\}), (\{\}, \{t_2\}), (\{s_2\}, \{t_3\})\}$

$A_R \cap A_S = \{(\{s_1\}, \{t_1\})\}$, $|A_R \cap A_S| = 1$; $|A_R| = 2$; $|A_S| = 3$

recall = 0.50; precision = 0.33; F1-measure = 0.40

Being that this framework is rather harsh – an average observer would intuitively state that the alignment presented in Figure 2 is better than F1-measure indicates – and also rather high-level-oriented, we introduced, once again according to (Langlais et al., 1998) metrics, other and more finely-grained bi-segment subdivisions and cast the F1-measure framework onto them. In the presented example some segments are only partially correct, e.g. $(\{s_2\}, \{t_3\})$, which is the reason to measure recall and precision at the sentence level, and not at the alignment level.

3.2. Sentence track F1-measure

Given alignments $A_R=\{ar_1,…,ar_n\}$ and $A_S=\{a_1,…,a_m\}$, with $a_i=(as_i,at_i)$ and $ar_j=(ars_j,art_j)$, sentence-to-sentence level metrics can also be defined:

Once again, on example set in Figure 2, the sets are defined and measures calculated as follows:

$A'_R = \{(\{s_1\}, \{t_1\}), (\{s_2\}, \{t_2\}), (\{s_2\}, \{t_3\})\}$

$A'_S = \{(\{s_1\}, \{t_1\}), (\{s_2\}, \{t_3\})\}$

$A'_R \cap A'_S = \{(\{s_1\}, \{t_1\}), (\{s_2\}, \{t_3\})\}$, $|A'_R \cap A'_S| = 2$; $|A'_R| = 3$; $|A'_S| = 2$

recall = 2/3=0.66; precision = 2/21; F-measure = 0.80

It is now obvious that sentence granularity and measure is much more forgiving than that of an alignment granularity and that it is also somewhat closer to human evaluation. We thus chose sentence track F1-measure as a solid base for our experiment.

Method of (Langlais et al., 1998.) suggests tuning sentence track F1-measure by added granularity as A'_R and A'_S set cardinality could be expressed in terms of token count and character count. These tweaks are called word granularity and character granularity by (Langlais et al., 1998.) and we chose to waive them for purposes of this experiment. We find them somewhat useful, as they introduce reward to partial correctness of sentence alignment, but also judge them as inherent to the Gale-Church algorithm by default and therefore not to be of major effect to overall results. We proceed to results presentation for alignment track and sentence track evaluation in the following section.

4. Results and discussion

Evaluation results on alignment level and sentence level F1-measure track are provided in Table 3 for both legislative documents corpora and Croatian-English parallel subcorpus.

When considering results provided by (Gale and Church, 1993) for the core algorithm and results of (Langlais et al, 1998.) for various specific algorithms, pre- and post-processing steps encapsulating Gale-Church algorithm, these results delivered by CORAL are rather predictable and expected. Being that Gale-Church algorithm is proven to work excellent in detecting one-on-one alignments and legislative texts provided in our test case are both really small – 6791 words for Croatian, 8510 words for English overall – and straightforward in terms of manual alignment complexity, figure of 97-98% correct alignments is not surprising.

	Alignment track			Sentence track		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Legislative	96.80	96.15	96.47	97.82	97.47	97.64
Cro-Eng	91.40	93.65	92.51	94.20	93.97	94.08

Table 3. Alignment and sentence track F1-measure on Croatian-English parallel corpora

Alignment on Croatian-English parallel corpus could be improved, on the other hand. Lower results are seen as a direct consequence of increased complexity on newspaper texts, where the basic Gale-Church algorithm encounters larger number of non-one-to-one (one-to-n, n-to-one, one-to-zero, zero-to-one, n-to-n, n-to-m) manual alignments and resolves these with decreased accuracy, as reported in (Gale and Church, 1993) and many papers that followed.

5. Future work

The work presented in this paper certainly leaves room for improvements. Results of this research would without doubt be better with larger and/or annotated corpora, which then could be used for tasks such as word alignment, terminology extraction, creation of thesauri, online dictionaries, semantic networks, etc. If corpora were also POS/MSD tagged and/or lemmatized, it would considerably reduce information search, ambiguities and would enable significantly better exploitation of the text. This we would like to leave for further directions of investigation. Integration of the Croatian language into this kind of multilingual surrounding that exists for other European languages, would enable adding one more language and additional research on new cross-language relations and identities.

Beyond the scope of building additional corpora and enriching existing ones with linguistic annotation, technical improvements might include implementing pre- and post-processing steps around the core Gale-Church algorithm in order to handle possible non-one-to-one alignments with higher recall and precision. The algorithm itself – as a dynamic programming method – might enable additional tweaks or integration with other language preprocessing modules. Future research activities could include alignment experiments at the lower linguistic level i.e. word level or they could include building basic language models and finally experimental systems for statistical machine translation on results presented in this paper.

6. Acknowledgements

This work has been supported by the Ministry of Science, Education and Sports, Republic of Croatia, under the grants No. 130-1300646-0645, 130-1300646-1776, 130-1300646-0909, 036-1300646-1986.

7. References

Arcade. 2007. Evaluation of parallel text alignment system. URL http://www.up.univ-mrs.fr/veronis/arcade/arcade1/index-en.html

Ceausu, A., Stefanescu, D.; Tufiş, D. 2006. Acquis Communautaire Sentence Alignment using Support Vector Machines. In Proceedings of the LREC2006, Genoa-Paris: ELRA-ELDA.

Dalbelo Bašić, B., Dovedan, Z., Raffaelli, I., Seljan, S., Tadić, M. 2007. Computational Linguistic Models and Language Technologies for Croatian. Proceedings of the 29th ITI Conference. Zagreb : SRCE, 2007. pp. 521-528

Danielsson, P., Ridings, D. 1997. Practical presentation of a "vanilla" aligner. TELRI Workshop on Alignment and Exploitation of Texts. Institute Jožef Stefan, Ljubljana.

EC-DG-JRC. The JRC-Acquis multilingual parallel corpus and Eurovoc (v. 3.0). Italy, 2007. (http://wt.jrc.it/lt/Acquis/JRC-Acquis.3.0/doc/README_Acquis-Communautaire-corpus_JRC.html)

Erjavec, T. 2004. MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. Proceedings of the LREC 2004, ELRA, Paris.

Gale, W. A., Churck, K. W. 1993. A Program for Aligning Sentences in Bilingual Corpora. Computational Linguistics, vol. 19, pp. 75 – 102. MIT Press, Cambridge, Massachusetts, SAD, 1993.

Koehn, P. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. Machine Translation Summit 2005. URL http://people.csail.mit.edu/koehn/publications/europarl

Langlais, P., Simard, M., Veronis, J. 1998. Methods and practical issues in evaluating alignment techniques. In COLING-ACL98, 1998.

Moore, R. C. 2002. Fast and Accurate Sentence Alignment of Bilingual Corpora. In Machine Translation: From Research to Real Users. Proceedings, 5th Conference of the Association for Machine Translation in the Americas. Springer-Verlag, Heidelberg, Germany.

Och, F. J., Ney, H. 2003. A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics, volume 29, number 1, pp. 19-51 March 2003.

Seljan, S., Gašpar, A., Pavuna, D. 2007. Sentence Alignment as the Basis For Translation Memory Database. INFuture2007 – The Future of Information Sciences: Digital Information and Heritage. Zagreb: Odsjek za informacijske znanosti, Filozofski fakultet, 2007.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., Varga, D. 2006. The JRC-Acquis : A Multilingual Aligned Parallel Corpus with 20+ Languages. Proceedings of the 5th International Conference on Language Resources and Evaluation LREC2006, Genoa, Italy, 24-26 May 2006.

Tadić, M. 2000. Building the Croatian-English Parallel Corpus. In Proceedings of the Second International Conference on Language Resources and Evaluation. ELRA, Paris – Athens 2000, pp. 523-530.

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., Nagy, V. 2005. Parallel corpora for medium density languages. In Proceedings of the Recent Advances in Natural Language Processing 2005 Conference, pp. 590-596.