Towards Obtaining High Quality Sentence-Aligned English-Croatian Parallel Corpus

Marija Brkić and Maja Matetić
Department of Informatics
University of Rijeka
Rijeka, Croatia
mbrkic@uniri.hr; maja.matetic@ri.t-com.hr

Sanja Seljan
Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Zagreb, Croatia
sseljan@ffzg.hr

Abstract— **This paper presents the acquisition of parallel bilingual corpus and all the steps involved in the process of unsupervised sentence alignment, such as tokenization, lowercasing, etc. The problem of sentence alignment is not trivial because translators do not necessarily translate one sentence in the source language into one sentence in the target language. Three different unsupervised and language independent approaches to sentence alignment are presented and implementations of these approaches through three different freely available tools are tested. A gold standard for English-Croatian automatic sentence alignment evaluation is created. Finally, a detailed analysis of the acquired corpus is given.**

Sentence alignment; alignment tools; sentence alignment evaluation; parallel corpus; sentence-length; word-correspondence

I. INTRODUCTION

There are two important factors that highly influence machine translation (MT) quality. These are domain and modality. As far as modality is concerned, MT can be applied to spoken and written language. If MT system is textual, then spoken language has to be converted into written form, either by manual transcription, or by automatic speech recognition system. However, spoken language is often ungrammatical and reliant on gestures and mutually understood knowledge. Therefore, a better language resource for building an MT system is a *parallel corpus*. A parallel corpus is a collection of text with its translation into another language [1].

In this paper we focus on the acquisition of sentence-aligned parallel corpus for the English-Croatian language pair and its preparation as the training data for a statistical machine translation (SMT) system. SMT systems are completely language independent, besides the fact that they require a huge parallel corpus that needs to be split into sentences and words [2]. Nowadays, the acquisition of a parallel corpus is straightforward, with the Internet as a valuable source. Information, as a product of daily activities, is published in multiple languages on a daily basis [3]. Since Croatia is in the process of negotiations for the European Union membership, the number of legal text translations from English into Croatian, and vice versa, is in constant increase. However, in order to be useful for applying machine learning methods to SMT, the parallel corpus needs to be sentence-aligned [4]. Many approaches to sentence alignment are either supervised or language dependent. In this paper we are interested into unsupervised language independent approaches, such as those presented in [4], [5], [6], [7], and [8].

The second section of this paper presents the acquisition of English-Croatian parallel corpus and all the preprocessing steps needed for unsupervised sentence alignment with three selected tools. In the third section three different approaches to sentence alignment, each implemented by one of the above mentioned tools, are presented. All three approaches are unsupervised and language independent. The fourth section deals with the evaluation of different sentence alignment tools. Last section outlines the experimental study conducted and analyzes of the acquired corpus. Directions for future work are given in the conclusion.

II. CORPUS

The corpus used in this experiment consists of a subset of 600 Acquis Communautaire documents. English documents can be obtained at http://eur-lex.europa.eu/en/index.htm, while Croatian documents can be obtained at http://ccvista.taiex.be/download.asp.

The documents, i.e. decisions, regulations and bylaws, have been obtained in doc format and aligned using the CELEX number.

Firstly, the conversion to txt format with UTF encoding has been done. All of the preprocessing programs take aligned documents as input (table I).

TABLE I. CORPUS

English	Croatian
Fuel tanks must be made so as to be corrosion resistant. They must satisfy the leakage tests carried out by the manufacturer at a pressure equal to double the working pressure…	Spremnici za gorivo moraju biti izrađeni tako da su otporni na koroziju. Moraju zadovoljiti ispitivanja na propuštanje koje provodi proizvođač pri tlaku dvostrukom od radnog tlaka…

A script written in Perl has been run on all the files in order to remove hard returns, tabulators, and extra spaces. This is an important preprocessing step which prevents oversegmentation in the process of sentence splitting. Moreover, for some of the tools used this is a pre-requirement. A sentence-split corpus has been obtained by running another Perl script.

The case-normalization is usually done by lowercasing or true-casing. This needs to be done since words may appear in lower-cased or upper-cased form in a text. True-casing preserves uppercase in names, and enables the distinction between Lončar (a surname) and *lončar* (a person who makes pots). We have used a script written in Perl to lowercase our corpus. Tokenization and lowercasing have been done by running two separate Perl scripts.

Lastly, while working with CorAl, non-English and non-Croatian words have been manually filtered out, since it has a user-friendly interface and only a limited degree of automation possible, i.e. the automatic sentence alignment process needs to be manually run for each document pair.

III. Sentence Alignment

In order to make the corpus useful for SMT systems, sentence alignment needs to be done [1]. That is not part of the translation process per se. It is mostly used for creating lexical resources like bilingual dictionaries or parallel grammars. It can also be exploited in word-sense disambiguation (words and phrases have different meanings in different domains – phenomenon known as polysemy) or information retrieval.

Furthermore, sentence alignment is a necessary step to fully exploit the benefits of computer-assisted (CAT) tools, e.g. translation memories (TMs). An illustration follows. Product manuals have multiple versions and need to be translated into multiple languages. Each new version can be aligned to a previous one in order to detect differences between the two, and then the previous version can be aligned to its translation. Only the newly added parts need to be translated and appended to the existing translation of the previous version [9].

The sentence alignment problem comes down to finding which group of sentences in one language corresponds to which group of sentences in another language, where either group can be empty to account for deletions and insertions [9]. Sentence alignment is only a first step toward the more ambitious task of word alignment [5]. The reason why the task of sentence alignment is not trivial is that translators do not always translate one sentence in the source language into exactly one sentence in the target language [9]. Long sentences may be broken up or short sentences may be merged [5]. The alignment type of a sentence pair is the number of sentences in the set [1]. Naturally, the most common situation is 1-to-1 sentence alignment, an alignment where one source sentence is aligned to one target sentence. Moreover, studies show that even around 90% of alignments fall into this category. The excerpt given in table I also exemplifies 1-to-1 alignment. However, there is a surprising number of crossing dependencies in real texts, i.e. the order of sentences in the translation changes [9]. Alignment problem algorithms are one class of string-matching problems which, unlike correspondence problem algorithms, do not account for crossing dependencies [5]. Therefore, rearrangements in the order of sentences must be described as many to many alignments. Each sentence may occur in only one alignment or *bead*. Sentences added in translations or deleted from translations lead to 1-to-0 and 0-to-1 alignments, alignments where there are no target sentences which could be aligned to some source sentences and alignments where there are no source sentences which could be aligned to some target sentences, respectively [9].

There are different approaches to sentence alignment, three of which will be described in the following subsections. Ideally, a sentence alignment method should be fast, accurate and language independent. Sentence-length-based methods are relatively fast and fairly accurate, while word-correspondence-based methods are more accurate but much slower. When texts to be aligned contain small deletions or free translations, the accuracy of sentence-length-based methods decreases drastically [6]. The following subsections outline three different approaches to sentence alignment.

A. Gale and Church

One of the early approaches to sentence alignment is the one by Gale and Church [5]. It is based on the simple statistical model of character lengths and on the notion that longer sentences in one language tend to be translated into longer sentences in another language, and vice versa. The algorithm Gale and Church propose is twofold. After paragraph alignment, the sentences within these paragraphs are aligned. Short headings and signatures usually have less than 50 characters, so the threshold of 100 characters is taken to differentiate between paragraphs and pseudo-paragraphs. Gale and Church use two components to define the match function. One component is a probability distribution for the alignment type, which is obtained from the hand-aligned data. Another component is a distance measure that considers the number of letters in each sentence. There are two preconditions and these are that all sentences need to be accounted for and that each sentence may occur in only one sentence pair.

B. Moore's method

Moore's method is a three-step hybrid method that uses sentence-length-based and word-correspondence-based models [4]. Sentence-split and word-split corpus is first aligned by using a modified version of Brown et al.'s sentence-length-based method, which has the same basis as that of Gale and Church described in the previous subsection. The sentence pairs assigned with the highest probability of alignment are used in the second step with the Expectation Maximization (EM) algorithm [1] to train a modified version of IBM Model 1. EM was first explained in 1977 by Dempster, Laird, and Rubin [10]. Only four iterations of EM are performed. In the third step of the Moore's method, the corpus is realigned, i.e. the initial model is augmented with IBM Model 1. Since the third step is confined to the minimal alignment segments that were assigned a non-negligible probability according to the initial model, the alignment is faster that the initial one, although the model is much more expensive to apply.

C. Braune and Fraser

Braune and Fraser present a two-pass method for sentence alignment, which augments a sentence-length based model with lexical statistics [6]. The alignment model they use is a slightly modified version of Moore's. In contrast to Moore's it allows extracting 1-to-many/many-to-1 alignments. In the first pass this method uses sentence-length-based statistics to extract the training data for the IBM Model 1 translation tables. In the second pass a model-optimal alignment composed of the smallest possible correspondences, i.e. 1-to-0/0-to-1 and 1-to-1, is found. These alignments are then merged into larger model-optimal clustered alignments, i.e. up to *R* sentences on each side of the cluster [6].

IV. Evaluation Metrics

In order to evaluate alignment methods, the sentence alignment story needs to be set on a more formal footing.

There are two sets of segments, S={s$_1$, s$_2$, …, s$_n$} and T={t$_1$, t$_2$, …, t$_m$}. S is a source language text, and T is its translation into a target language. The alignment A between S and T can be defined as a subset of the Cartesian product P(S) x P(T), where P(S) and P(T) stand for the set of all subsets of S and the set of all subsets of T, respectively. The triple (E, S, C) is then called a bitext, and each of the elements of the alignment is called a *bisegment*.

Recall and *precision* defined at the alignment level do not take into account partial correctness of bisegments.

The alignment A *recall* is defined with respect to the reference alignment A$_r$, as in (1), and represents the proportion of bisegments in A that are correct with respect to A$_r$.

$$recall = | A \cap A_r | / | A_r | . \qquad (1)$$

The alignment A *precision* is defined with respect to the reference alignment A$_r$, as in (2), and represents the proportion of bisegments in A that are correct with respect to the number of bisegments proposed.

$$precision = | A \cap A_r | / | A | . \qquad (2)$$

Equation (3) defines *F-measure*, which is a harmonic mean of precision and recall and enables combining recall and precision in a single efficiency measure.

$$F = 2 \times \frac{recall \times precision}{recall + precsion} . \qquad (3)$$

Out of the alignment A={a$_1$, a$_2$, …, a$_m$} and the reference A$_r$={ar$_1$, ar$_2$, …, ar$_n$}, with a$_i$=(as$_i$, at$_i$) and ar$_i$=(ars$_i$, art$_i$), the sentence-to-sentence recall and precision can be derived, as in (4) and (5), where $A^{'}$ and $A^{'}_{r}$ are defined in (6) and (7), respectively [11].

$$recall = | A' \cap A'_r | / | A'_r | . \qquad (4)$$

$$precision = | A' \cap A'_r | / | A' | . \qquad (5)$$

$$A' = \bigcup_i (as_i \times at_i) . \qquad (6)$$

$$A_r' = \bigcup_1 (ars_1 \times art_1) . \qquad (7)$$

V. Experimental Study

A. Tools

For the task of sentence alignment, three unsupervised and language-independent tools have been selected, i.e. CorAl [12], Bilingual Sentence Aligner [4] and Gargantua [6].

CorAl, developed at the University of Zagreb, Faculty of Electrical Engineering, is the Java implementation of the algorithm designed by Gale and Church. The second tool, Bilingual Sentence Aligner, is a set of Perl scripts that implement Moore's method. Gargantua, the third tool, is written in C++ and implements Braune and Fraser's two-pass method for sentence alignment.

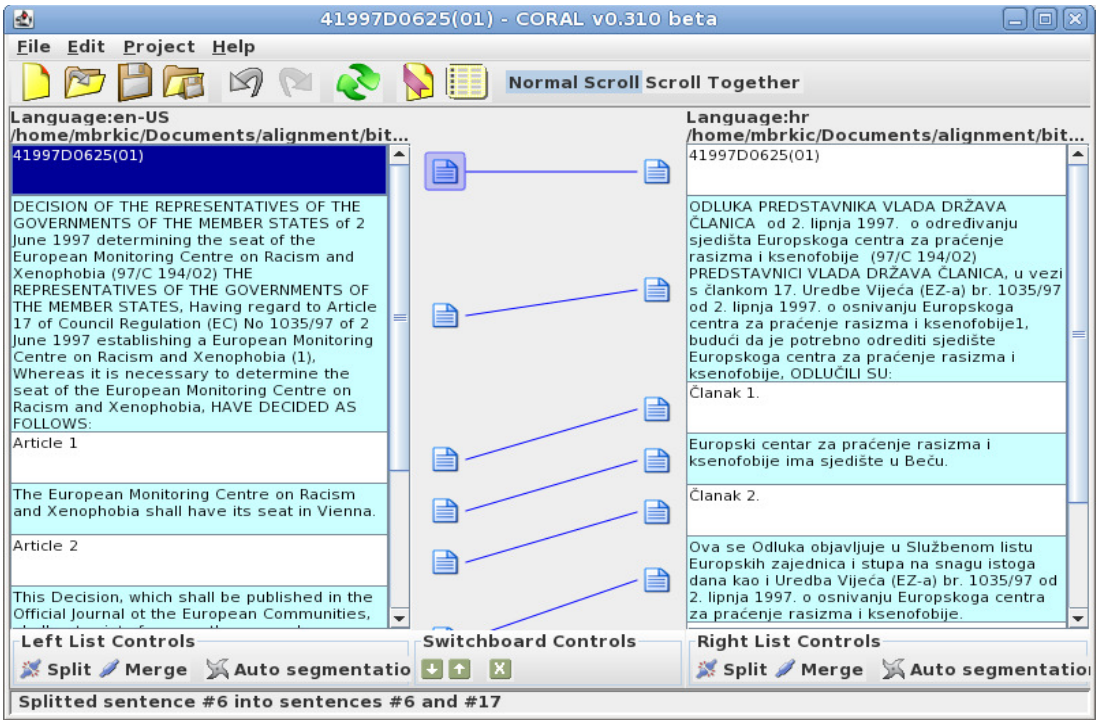
CorAl, unlike the other two tools, has GUI and is enriched by a sentence segmentation module. There are both, automatic and fully manual modes of usage [12]. However, the process of sentence aligning of multiple files is several orders of magnitude faster by using the other two tools because they operate fully automatic. As far as corpus preparation is concerned, CorAl poses almost no pre-requirements, Bilingual Sentence Aligner requires sentence-split corpus, while Gargantua requires cleaned, sentence-split, tokenized and lowercased corpus.

B. Procedure

1) CorAl

First, we describe the procedure of sentence alignment by using CorAl (Fig. 1). CorAl has been used for automatic and manual sentence alignment tasks.

After loading each pair of parallel documents into CorAl, semi-automatic sentence splitting based on the list of abbreviations has been done. Over-segmentation on the Croatian side of the parallel corpus has been observed in cases where ordinal numbers are followed by words starting with capital letters. On the English side of the corpus, automatic segmentation module has shown the tendency to under-segment because numbers are not followed by periods. Next, automatic sentence alignment has been done.



41997D0625(01)

DECISION OF THE REPRESENTATIVES OF THE GOVERNMENTS OF THE MEMBER STATES of 2 June 1997 determining the seat of the European Monitoring Centre on Racism and Xenophobia (97/C 194/02) THE REPRESENTATIVES OF THE GOVERNMENTS OF THE MEMBER STATES, Having regard to Article 17 of Council Regulation (EC) No 1035/97 of 2 June 1997 establishing a European Monitoring Centre on Racism and Xenophobia (1), Whereas it is necessary to determine the seat of the European Monitoring Centre on Racism and Xenophobia, HAVE DECIDED AS FOLLOWS:

Article 1

The European Monitoring Centre on Racism and Xenophobia shall have its seat in Vienna.

Article 2

This Decision, which shall be published in the Official Journal ot the European Communities,

41997D0625(01)

ODLUKA PREDSTAVNIKA VLADA DRŽAVA ČLANICA od 2. lipnja 1997. o određivanju sjedišta Europskoga centra za praćenje rasizma i ksenofobije (97/C 194/02) PREDSTAVNICI VLADA DRŽAVA ČLANICA, u vezi s člankom 17. Uredbe Vijeća (EZ-a) br. 1035/97 od 2. lipnja 1997. o osnivanju Europskoga centra za praćenje rasizma i ksenofobije1, budući da je potrebno odrediti sjedište Europskoga centra za praćenje rasizma i ksenofobije, ODLUČILI SU:

Članak 1.

Europski centar za praćenje rasizma i ksenofobije ima sjedište u Beču.

Članak 2.

Ova se Odluka objavljuje u Službenom listu Europskih zajednica i stupa na snagu istoga dana kao i Uredba Vijeća (EZ-a) br. 1035/97 od 2. lipnja 1997. o osnivanju Europskoga centra za praćenje rasizma i ksenofobije.

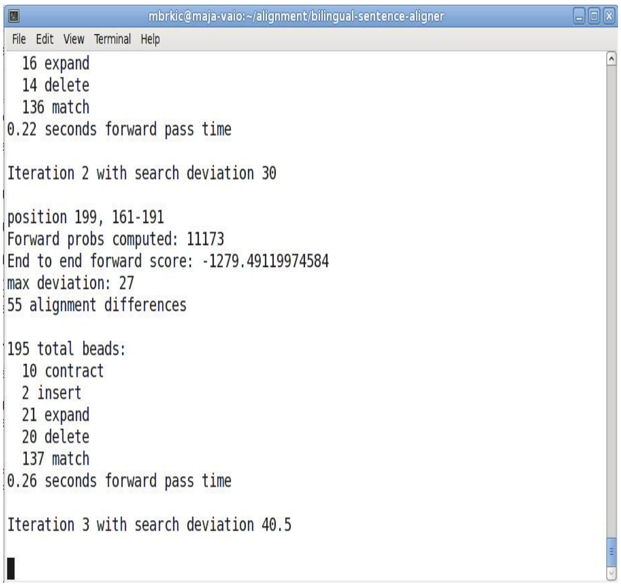
Splitted sentence #6 into sentences #6 and #17

Figure 1. CorAl tool.

In order to obtain a gold standard for automatic sentence alignment evaluation, manual sentence alignment has been done by one person and checked by a professional translator. In this task foreign phrases have been filtered out.

2) Bilingual Sentence Aligner

A screenshot taken during the sentence alignment task with the second tool is shown in Fig. 2.



```
  16 expand
  14 delete
  136 match
0.22 seconds forward pass time

Iteration 2 with search deviation 30

position 199, 161-191
Forward probs computed: 11173
End to end forward score: -1279.49119974584
max deviation: 27
55 alignment differences

195 total beads:
  10 contract
  2 insert
  21 expand
  20 delete
  137 match
0.26 seconds forward pass time

Iteration 3 with search deviation 40.5

▌
```

Figure 2. Bilingual Sentence Aligner tool.

The aligner makes an assumption that the alignable sentences in each file are in the same order. However, the assumption that there are only 1-to-1 alignments is not made. Sentences that are alignable but are out of order are not identified as alignable. Bilingual Sentence Aligner relies on enough data because it estimates a statistical word-translation model. Therefore, a minimum of 10.000 sentence pairs in the input is recommended. The output are all the sentences that align with probability greater than some threshold (0.5 is the default value) according to a statistical model computed by the aligner.

The sentences to be aligned need to be in paired files with one sentence per line and spaces between words. Lowercasing is done automatically and punctuation marks are cleaned out. Model 1 (Fig. 3) that has been obtained by the tool contains 231.153 entries.

3) Gargantua

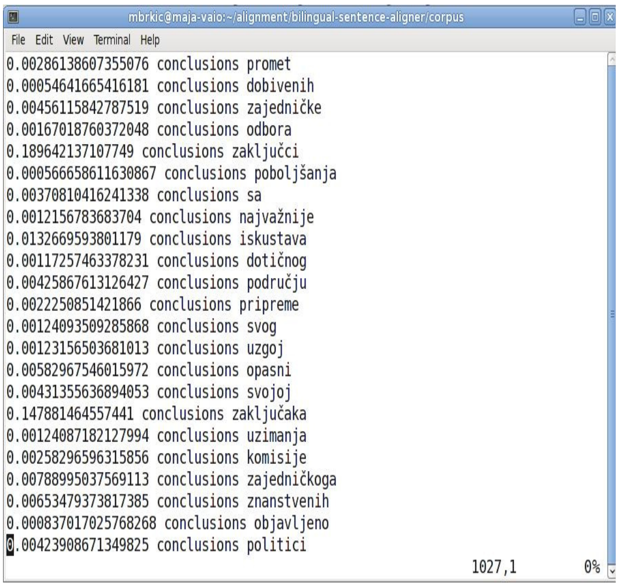
Braune and Fraser released a tool called Gargantua (Fig. 4). Prior to running the automatic sentence alignment task, the parallel corpus needs to be cleaned, i.e. empty lines need to be removed, and the corpus needs to be split into sentences. Additional requirements the tool imposes are tokenization and lowercasing. The aligner is only about 4 times slower than that of Moore's in aligning symmetrical documents [6].

C. Results

1) Corpus analysis

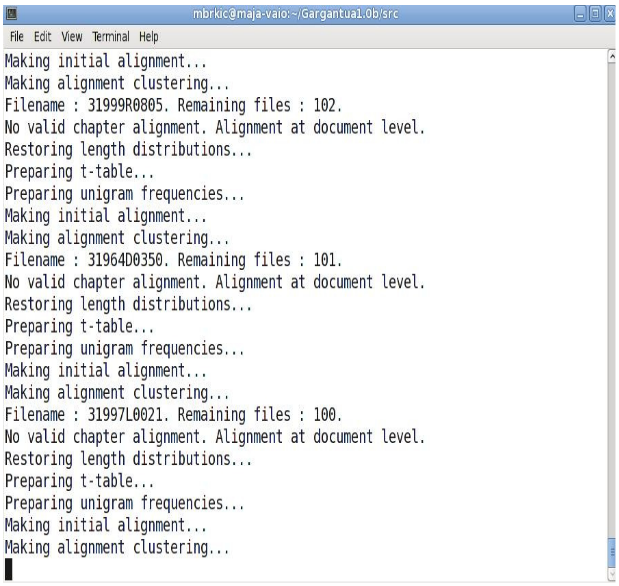
A detailed analysis of the obtained sentence-aligned parallel corpus follows. Word distribution tells how many words are used in a corpus. This is useful since answering the question of how many words there are in any language is an open-ended question due to the dynamic nature of languages, i.e. new words are constantly coined or borrowed and some existent words fall out of use.

For comparison, English Europarl corpus consists of about 29 million words and one million sentences, out of which 86.699 are different words. The top ten words in the English Europarl corpus are all function words. Function words come from a fixed class of words and they fulfill a specific role of how words relate together in a language. They pose a challenge for machine translation because a type of role that exists in one language may not exist in another language. The most frequent word is the article *the,* which makes up almost 7% of the corpus, and the top ten words, which include comma and end-of-sentence period token, make up 30% of the corpus. There are 33.447 words that occur only once [1].


mbrkic@maja-vaio:~/alignment/bilingual-sentence-aligner/corpus
File Edit View Terminal Help

```
0.00286138607355076 conclusions promet
0.00054641665416181 conclusions dobivenih
0.00456115842787519 conclusions zajedničke
0.00167018760372048 conclusions odbora
0.189642137107749 conclusions zaključci
0.000566658611630867 conclusions poboljšanja
0.00370810416241338 conclusions sa
0.0012156783683704 conclusions najvažnije
0.0132669593801179 conclusions iskustava
0.00117257463378231 conclusions dotičnog
0.00425867613126427 conclusions području
0.0022250851421866 conclusions pripreme
0.00124093509285868 conclusions svog
0.00123156503681013 conclusions uzgoj
0.00582967546015972 conclusions opasni
0.00431355636894053 conclusions svojoj
0.147881464557441 conclusions zaključaka
0.00124087182127994 conclusions uzimanja
0.00258296596315856 conclusions komisije
0.00788995037569113 conclusions zajedničkoga
0.00653479373817385 conclusions znanstvenih
0.000837017025768268 conclusions objavljeno
0.00423908671349825 conclusions politici
                                                1027,1          0%
```

Figure 3. Model 1.


mbrkic@maja-vaio:~/Gargantua1.0b/src
File Edit View Terminal Help

```
Making initial alignment...
Making alignment clustering...
Filename : 31999R0805. Remaining files : 102.
No valid chapter alignment. Alignment at document level.
Restoring length distributions...
Preparing t-table...
Preparing unigram frequencies...
Making initial alignment...
Making alignment clustering...
Filename : 31964D0350. Remaining files : 101.
No valid chapter alignment. Alignment at document level.
Restoring length distributions...
Preparing t-table...
Preparing unigram frequencies...
Making initial alignment...
Making alignment clustering...
Filename : 31997L0021. Remaining files : 100.
No valid chapter alignment. Alignment at document level.
Restoring length distributions...
Preparing t-table...
Preparing unigram frequencies...
Making initial alignment...
Making alignment clustering...
```

Figure 4. Gargantua tool.

In our sentence-aligned parallel corpus, there are 21.846 unique word-forms on the source side and 39.345 unique word-forms on the target side (table II). This is not surprising since English is morphologically poor language and Croatian morphologically rich language.

The list of the ten most frequent words (punctuation excluded) for both sides of the corpus is shown in table III. The most frequent word on the source side is the article *the,* which makes up about 8% of the corpus, and the top ten words make up about 27% of the corpus. There are 9.372 words that occur only once. As far as the target side is concerned, the most frequent word is the preposition *u,* which makes up 3% of the corpus, and the top ten words make up about 16% of the corpus. There are 17.589 words that occur only once. Almost all ten most frequent words on both sides of the corpus are function words (the ninth word on the source side is an exception). When punctuation is included, then the most frequent token on the target side is a dot, with the frequency that is almost 2 times bigger than that of the most frequent word. On the source side, the dot is on the fourth position as far as frequency is concerned, with the comma being on the third position. Furthermore, the English side of the corpus contains more commas than Croatian.

Zipf's law defines the distribution of words in a corpus. According to Zipf's law, the product of the rank r of each word (sorted by frequency) and its frequency f is roughly a constant, i.e. the frequency of any word is inversely proportional to its rank in the frequency table. Fig. 5 shows the distribution of words in the corpus. It is evident that the English side of the corpus is more in accordance with the Zipf's law than Croatian, which we attribute to the fact that Croatian is morphologically rich language and that the law was checked against unlemmatized corpus. Zipf's law should be treated as a roughly accurate characterization of certain empirical facts, rather than as a law. It is useful as a rough description of the frequency distribution of words in human languages [9].

If we want to have a corpus big enough to see a sufficient number of occurrences of rare words, let a sufficient number be defined as 10.000 occurrences of the 10.000th most frequent word, we need a corpus that consists of around 12.5 million words on the English side and around 33.5 million words on the Croatian side. This is calculated from (8) [1].

$$r \times f = \%mostFrequentWord \times s \qquad (8)$$

2) *Sentence alignment task*

A large reference bilingual corpus has been created, aligned at a sentence level. Table IV presents the total number of segments aligned by each the above described tools. The number of segments identified by Gargantua is about 33% higher than the number of segments identified by Bilingual Sentence Aligner. The number of segments identified manually is only about 2% higher than the number of segments identified by CorAl automatically.

VI. CONCLUSION

Parallel corpora, particularly sentence-aligned parallel corpora, are a valuable resource for numerous research experiments. Larger corpora (e.g. Europarl, Acquis Communautaire) have proven their value in SMT, information retrieval and the creation of language resources.

The next phase of our work will include evaluating the quality of automatic sentence alignment obtained by each of these tools by calculating recall and precision with respect to the reference, i.e. manual sentence alignment, as well as identifying the counts for each alignment type dependent on the tool used. We expect the error rate to be low because we are dealing with translations from the legal domain, which are more literal. We are also interested to check how the quality of a sentence alignment tool affects the output of an SMT system, which has motivated us to do this experiment.

To conclude, we would like to point out that the existence of these language-independent sentence alignment tools drives forward the development of language resources for resource-poor and minority languages. Languages like Croatian, which is spoken by only about 6 million people, highly benefit from the availability of such tools.

TABLE II. CORPUS STATSISTICS

	English				**Croatian**			
	Min	***Max***	***Avg***	***Total***	***Min***	***Max***	***Avg***	***Total***
Words	131	28.327	2.117	635.234	98	27.790	1.890	567.287
Tokens	149	31.965	2.501	750.564	116	35.517	2.326	697.944
Unique tokens	68	2.879	492	21.964	68	4.242	656	39.488
Characters	759	146.565	10.940	3.282.219	672	124.852	10.957	3.287.141
Characters with spaces	890	174.504	13.048	3.914.528	756	144.910	12.480	3.744.166
Punctuation	9	6.922	368	110.512	10	17.880	450	135.017
Sentences	3	1.751	80	24.204	5	1.672	85	25.752
Bytes	890	174.892	13.058	3.917.453	770	147.994	12.848	3.854.428

TABLE III. MOST FREQUENT WORDS

English		**Croatian**	
Words	***Counts***	***Words***	***Counts***
the	51.691	u	16.813
of	32.920	i	15.518
to	17.375	se	11.979
in	15.188	na	8.359
and	14.208	za	7.908
shall	9.052	ili	6.552
be	8.516	je	5.677
a	7.836	od	5.248
article	7.258	da	5.183
for	6.899	o	5.051

TABLE IV. SEGMENT COUNTS IN SENTENCE-ALIGNED PARALLEL CORPUS

Semi-manual sentence-splitting		**Automatic sentence-splitting**	
CorAl	***Manual***	***Bilingual Sentence Aligner***	***Gargantua***
35.650	36.227	12.993	19.541

REFERENCES

[1] P. Koehn, Statistical machine translation. Edinburgh, UK: Cambridge University Press, 2010.

[2] M. Brkić, T. Vičić, and S. Seljan, “Evaluation of the statistical machine translation service for Croatian-English,” Proc. International Conf. The Future of Information Sciences (INFuture 09), Sept. 2009, pp. 319-332.

[3] A. Lopez, “Statistical machine translation,” ACM Computing Surveys (CSUR), vol. 40(3), Aug. 2008, pp. 1-49, doi: 10.1145/1380584.1380586.

[4] R. Moore, “Fast and accurate sentence alignment of bilingual corpora,” Proc. 5th Conf. Association for Machine Translation in the Americas (AMTA 02), Oct. 2002, pp. 135-144.

[5] W. Gale and K. Church, “A program for aligning sentences in bilingual corpora,” Computational linguistics, vol. 19(1), March 1993, pp. 75-102, doi: 10.3115/981344.981367.

[6] F. Braune and A. Fraser, “Improved unsupervised sentence alignment for symmetrical and asymmetrical parallel corpora,” Proc. 23rd International Conf. Computational Linguistics (COLING 2010), Aug. 2010, pp. 81-89.

[7] Y. Deng and W. Byrne, W, “MTTK: An alignment toolkit for statistical machine translation,” Proc. Conf. of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume: demonstrations (HLT-NAACL 2006), June 2006, pp. 265-268, doi: 10.3115/1225785.1225789

[8] A. Ceausu, D. Stefãnescu and D. Tufis, “Acquis Communautaire sentence alignment using support vector machines,” Proc. 5th International Conf. Language Resources and Evaluation (LREC 2006), May 2006, pp.2134-2137

[9] C. Manning, H. Schütze, and MITCogNet, Foundations of statistical natural language processing. Cambridge, MA: MIT Press, 1999.

[10] G. Lugar and W. Stubblefield, Artificial intelligence: structures and strategies for complex problem solving. Redwood City, CA: Benjamin/Cummings, 2008.

[11] P. Langlais, M. Simard, and J. Véronis, "Methods and practical issues in evaluating alignment techniques," Proc. 17th International Conf. Computational Linguistics (COLING 1998), vol. 1, Aug. 1998, pp. 711-717, doi: 10.3115/980451.980964.

[12] S. Seljan, M. Tadić, Ž. Agić, J. Šnajder, B. Dalbelo Bašić, and V. Osmann, "Corpus Aligner (CorAl) Evaluation on English-Croatian parallel corpora," Proc. International Conf. Language Resources and Evaluation (LREC 2010), May 2010, pp. 3481-3484.

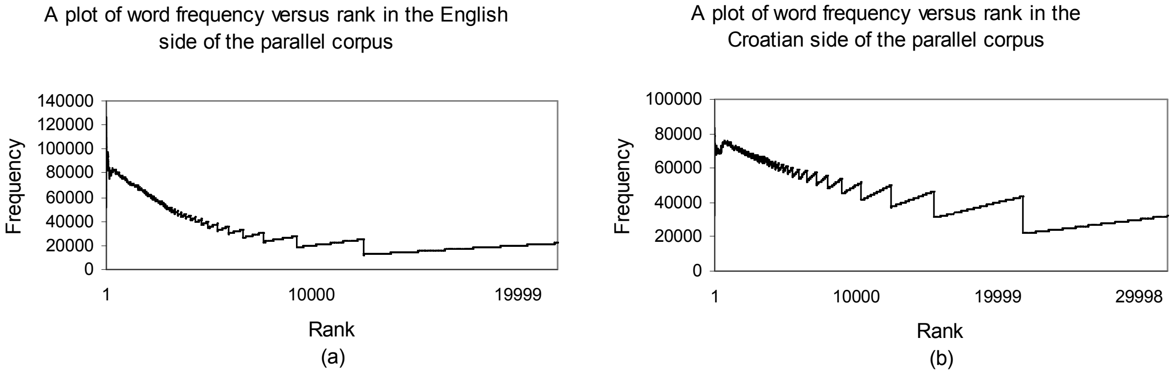


Figure 5. A plot of word frequency versus rank: (a) the English side of the parallel corpus (b) the Croatian side of the parallel corpus.