ACCURAT

Analysis and Evaluation of Comparable Corpora
for Under Resourced Areas of Machine Translation

www.accurat-project.eu

Project no. 248347

Deliverable D2.4
Aligned comparable corpora

Version No. 1.0
30/11/2011

Document Information

Deliverable number:	D2.4
Deliverable title:	Aligned comparable corpora
Due date of deliverable:	30/11/2011
Actual submission date of deliverable:	30/11/2011
Main Author(s):	Radu Ion, Nikos Glaros, Mārcis Pinnis, Mateja Verlic, Ahmet Aker, Gregor Thurmair, Fangzhong Su
Participants:	RACAI, TILDE, USFD, ZEMANTA, ILSP, LINGUATEC, CTS
Internal reviewer:	ILSP
Workpackage:	WP2
Workpackage title:	Multi-level alignment methods and information extraction from comparable corpora
Workpackage leader:	RACAI
Dissemination Level:	PU
Version:	V1.0
Keywords:	aligned comparable corpora, document pairing in comparable corpora

History of Versions

Version	**Date**	**Status**	**Name of the Author (Partner)**	**Contributions**	**Description/ Approval Level**
V0.1	08/11/2011	Draft	RACAI	Table of Contents	Created
V0.2	25/11/2011	Draft	RACAI	Added EN-RO chapter	EN-RO document alignments with statistics
V0.3	25/11/2011	Draft	RACAI	Added EN-DE chapter	EN-DE document alignments with statistics
V0.4	25/11/2011	Draft	RACAI	Added EN-SL chapter	EN-SL document alignments with statistics
V0.5	25/11/2011	Draft	RACAI	Added EN-LV chapter	EN-LV document alignments with statistics
V0.6	25/11/20	Draft	RACAI	Added EN-LT	EN-LT document

	11				alignments with statistics
V0.7	26/11/2011	Draft	RACAI	Added EN-ET	EN-ET document alignments with statistics
V0.8	26/11/2011	Draft	RACAI	Added conclusion and a tentative comparability assessment for corpora	Conclusions
V0.9	27/11/2011	Draft	ILSP	Added EN-EL aligned corpora info	EN-EL document alignments with statistics
V0.91	28/11/2011	Draft	RACAI	Added Introduction	Introduction and merged comments from ILSP
V0.95	30/11/2011	Draft	RACAI	ILSP comments	Reviewed ILSP comments.
V1.0	30/11/2011	Final	Tilde	Final review	Submitted to PO

EXECUTIVE SUMMARY

This report describes the document level alignments that have been obtained by aligning corpora collected in the ACCURAT project. This information is useful for subsequent tasks of MT-related data mining such as parallel data mining or translation lexicon extraction in that comparable document pair filtering will reduce the huge search space of such tasks.

The collected data are stored at the ACCURAT project FTP Server repository and are freely available after contacting the ACCURAT consortium: project@tilde.lv.

Table of Contents

Abbreviations

Abbreviation	Term/definition
ASCII	American Standard Code for Information Interchange
MT	Machine Translation
SMT	Statistical Machine Translation
RBMT	Rule Based Machine Translation
CC	Comparable Corpora
HTML	HyperText Markup Language
SMT	Statistical Machine Translation
MCC	Multilingual comparable corpus
URL	Uniform Resource Locator
UTF-8	Unicode Transformation Format

1. Introduction

The present deliverable (D2.4) reports on Comparable Corpora (CC) alignment issues based on results that have been produced by applying work package WP2[1] document alignment tools on a large subset of the various comparable corpora that have been collected in work packages WP3[2] and WP4[3].

It is easier and more cost-effective to collect CC rather than parallel corpora, for a large number of languages and for many specialised knowledge domains, given that comparable documents are by definition much less strictly inter-correlated than the parallel ones. The predominant paradigm for acquiring multilingual CC is the cross language information retrieval method based on seed lists of source documents URLs.

A direct consequence of the nature of CC and of their collection method is the generally large size of the retrieved comparable textual data as compared to (truly) parallel corpora collected from the WWW. The size difference is usually several orders of magnitude. Therefore, mining these high volumes of comparable data for parallel textual segments is far more computationally intensive and challenging than doing so for parallel corpora. Specifically, we are referring to the computation of the positional information of the translation units in parallel texts (paragraphs, sentences or phrases), which constitutes a natural pruning technique, when mining for parallel phrases in parallel texts. That is, given an *i*-th textual unit in the source document, one has only a limited window of ±k textual units around the *j*-th textual unit in the target document to search for alignment. However, this is not true in case of CC, where a pair of translation equivalents can be located anywhere in a given pair of source and target documents (k is much bigger here).

To lessen this large computational load, inherently evident in the process of extracting parallel data from CC, one has to reduce the search scope of the corresponding parallel data mining algorithms[4], among other things. One way to do so is to first detect the pairs of documents that have the greatest chances of actually containing parallel data, then mine those pairs only for parallel data. We call this step the "document alignment" step and this deliverable consolidates results of document alignments that have been **automatically** detected in comparable corpora having been collected in the frame of ACCURAT project.

Results for CC alignment at the document level are given here for a series of comparable corpora covering the following language pairs:

- English-Romanian (EN-RO)
- English-German (EN-DE)
- English-Slovenian (EN-SL)
- English-Latvian (EN-LV)
- English-Lithuanian (EN-LT)
- English-Estonian (EN-ET)

[1] WP2: "Multi-level alignment methods and information extraction from comparable corpora"

[2] WP3: "Methods and techniques for building a comparable corpus from the Web"

[3] WP4: "Comparable corpora in MT systems"

[4] These are algorithms that search for parallel textual units in all possible source and target document pairs in a bilingual corpus.

- English-Greek (EN-EL)

In the sequel, comparable corpora compilation details, document alignment statistics and information on where to find the actual document alignment data are provided in a separate section for every one of the above aligned CC.

2. Document Alignments for English-Romanian

2.1 Corpus Compilation

For English-Romanian we have document-aligned two corpora: by their short names "USFDNews" and "ILSPNewsDisasters". Since the USFDNews corpus was already document-aligned, we only had to align the ILSPNewsDisasters (EN-RO) corpus at the document level.

USFDNews is documented in the following ACCURAT deliverables:

- ACCURAT Deliverable D3.6 "Comparable corpora for under-resourced languages" from which we learn (Table 1, page 6) that for EN-RO, this corpus contains 5.516 English documents, 3.363 Romanian documents with 2.559.497 words in English and 1.206.919 words in Romanian;
- ACCURAT Deliverable D3.4 "Report on methods for collection of comparable corpora" where, on page 6, section 2.1.2, we find the document alignment method. In short, these documents were aligned based on the same date/time interval, lengths and similarity of their titles, etc. (see the deliverable for more details). This way, a number of 25.627 document pairs was generated.

ILSPNewsDisasters is a "narrow domain" (EN-RO) corpus and is presented in the following ACCURAT deliverables:

- ACCURAT Deliverable D3.4 "Report on methods for collection of comparable corpora" where in section 3.2, page 17, we learn that is a monolingual-driven corpus collected by focused crawling using a seed list of topic-specific terms and URLs;
- ACCURAT Deliverable D3.7 "Comparable corpora for narrow domains" where, on page 13, we learn that the "Topical News Disasters" corpus contains 24.555 English documents with 25.806.107 words and 4.494 Romanian documents with 4.367.014 words.

2.2 Document Alignment Statistics

Table 1 Quantitative description of the English-Romanian document alignments

	Total aligns.	**Avg. 1:n**	**Avg. n:1**	**Avg. score**	**Max. score**	**Min. score**	**% greater than avg.**	**% smaller than avg.**
USFDNews (EN-RO)	25627	2.34	3.22	1	1	1	0	0
ILSPNews Disasters (EN-RO)	10165	1.56	3.01	9.83e-5	0.005	6.23e-7	27.7%	72.3%

The columns in Table 1 above have the following meanings:

- "Total aligns." specifies the total number of document pairs that have been obtained;
- "Avg. 1:n" gives the average number of target documents that align to a single source document. This can be a measure of how weakly comparable the corpus is: the higher this number, the more likely is to deal with a weakly comparable corpus, because a high alignment productivity usually indicates an impossibility of a "sure" alignment;
- "Avg. n:1" gives the average number of source documents that align to a single target document. This is different from the "Avg. 1:n", because EMACC similarity measure is not symmetric;
- "Avg. score" is the average probability score for all document pairs in the alignment;
- "Max/Min. score" shows the maximum and the minimum alignment probabilities from the entire set;
- "% greater/smaller than avg." indicates the percentage of the document pairs that have an alignment probability that is higher/lower than the average "Avg. score". We hypothesise that if the two percentages are highly disproportionate, we are also dealing with a weakly comparable corpus.

ILSPNewsDisasters has been document aligned using EMACC (see ACCURAT Deliverable D2.6 "Toolkit for multi-level alignment and information extraction from comparable corpora", page 39) for which we do not have probability threshold limits as to what accounts for strongly and weakly comparable. Furthermore, EMACC normalizes the probabilities computed over the entire alignment set and thus, any high probability is eventually flattened. USFDNews corpus has been aligned with a technique that does not produce alignment probabilities and, because of this, we assigned the probability "1" to any document pair from the set.

Comparing USFDNews with ILSPNewsDisasters, we can see that they have a similar comparability degree if we are to judge by "Avg. n:1" and "Avg. score" alone. Furthermore, we are inclined to believe that, because of the fact that only 27% of document alignments are higher than the average, that ILSPNewsDisasters corpus is a weakly comparable corpus. The next figure plots the alignment probabilities for ILSPNewsDisasters corpus from the highest to the lowest scaled so that the maximum alignment probability 0.005 becomes 1 (200 times).

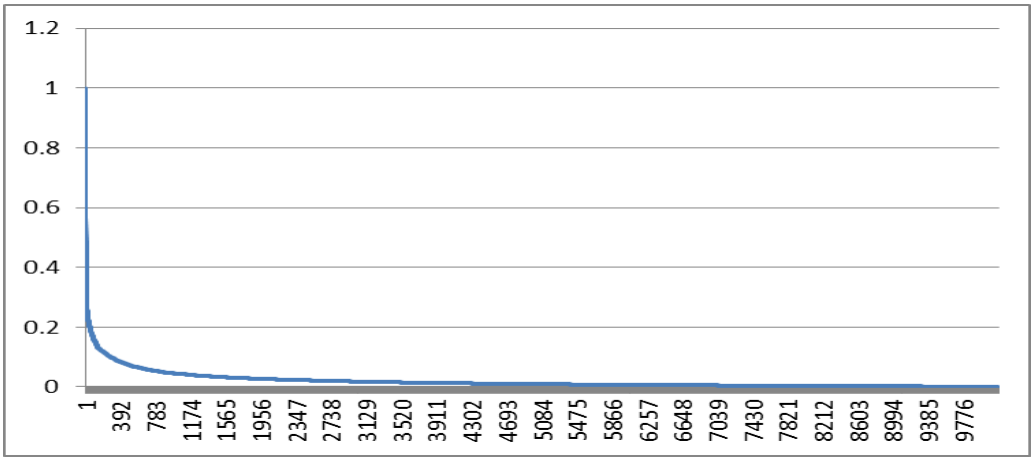

Figure 1: EMACC document alignment probabilities from the highest to the lowest for the English-Romanian ILSPNewsDisasters corpus

2.3 Document Alignment Location and Format

The English-Romanian document alignments are found on the ACCURAT FTP Server under the "/WP2/D2.4" directory. Each file name mentions the following:

- pair of source and target languages;
- the algorithm used to align the documents: currently, this can be EMACC ('emacc', see D2.6 "Toolkit for multi-level alignment and information extraction from comparable corpora" section 2.3), ComMetric ('commetric', see D2.6 "Toolkit for multi-level alignment and information extraction from comparable corpora" section 2.1) or its newer version DicMetric ('dicmetric') or the USFD algorithm described in the section 2.1.2 of the D3.4 deliverable ('usfd'). For instance, a document alignment file may read '`en-lv-USFDNews-commetric-docalign.lst`' meaning that the source document in a pair is in English, the target document is in Latvian, the processed corpus was called USFDNews and the algorithm used was ComMetric.

A document alignment file contains on each line (ended with the newline character ASCII 0x0A) a document pair along with its alignment score all separated by TAB characters (ASCII 0x09). For instance, a line in that file looks like:

```
ilsp/disasters/en-ro/en/file_94227_cleared.txt <TAB>
ilsp/disasters/en-ro/ro/file_14231_cleared.txt <TAB>
0.005 <NEWLINE>
```

3. Document Alignments for English-German

3.1 Corpus Compilation

For the English-German language pair and to be able to support the RBMT system's evaluation for Automotive domain, we have aligned the ILSP Automotive version 2 corpus, called hereafter ILSPAutomotiveV2. This is a narrow domain (EN-DE) corpus that is described by the following deliverables:

- ACCURAT Deliverable D3.4 "Report on methods for collection of comparable corpora", where in section 3.2, page 17, we learn that is a bilingual narrow domain comparable corpus collected by ACCURAT focused crawling tool using a seed list of topic-specific terms and URLs;
- ACCURAT Deliverable D3.7 "Comparable corpora for narrow domains" where, on page 13, we learn that the "Automotive engineering" corpus contains 5.629 English documents with 6.122.805 words and 11.651 German documents with 8.283.115 words.

3.2 Document Alignment Statistics

Table 2 Quantitative description of the English-German document alignments

	Total aligns.	Avg. 1:n	Avg. n:1	Avg. score	Max. score	Min. score	% greater than avg.	% smaller than avg.
ILSP AutomotiveV2 (EN-DE)	16293	16.26	11.24	0.06	0.167	0.05	36.3%	63.7%

The descriptions of the columns are the same as in the case of English-Romanian document alignments. Here we have to mention that due to the anticipated high running time of the parallel data miner, we were forced to select a fraction of the document alignments that EMACC produced for this pair of languages. That is, we selected the top 1% of the produced alignments in the decreasing order of the document alignment probabilities, a step which produced the 16.293 figure reported in Table 2. The next figure plots the distribution of the sorted alignment probabilities scaled so that the maximum value of 0.167 becomes 1 (6 times)

.

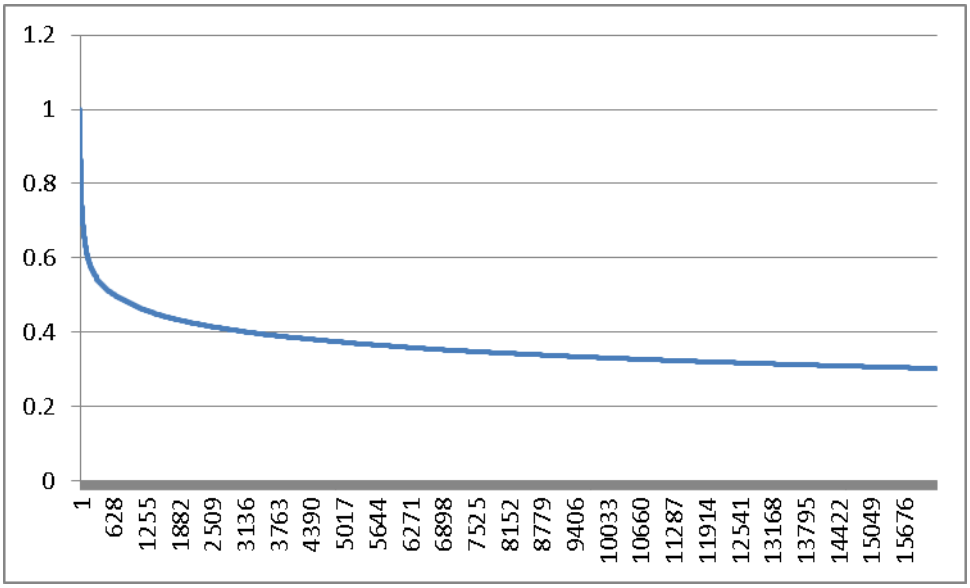

Figure 2: Top 1% of the EMACC document alignment probabilities from the highest to the lowest for the English-German ILSPAutomotiveV2 corpus

Judging the 36.6% of alignments that have an alignment probability above the average and the 16 target documents per a source document value, we may also conclude that we are dealing with a weakly comparable corpus. The fact that the lowest alignment probability is much larger by comparison with the English-Romanian case is a result of having used here only the top 1% of the alignments produced by EMACC.

3.3 Document Alignment Location and Format

The English-German document alignments are found on the ACCURAT FTP Server under the "/WP2/D2.4" directory. File name conventions and format are the same as in case of English-Romanian document alignments from section 2.3.

For instance, a line in the '`en-de-ILSPAutomotiveV2-emacc-docalign.lst`' file looks like

```
automotive-v2/en-de/en/file_11015_cleared.txt <TAB>
automotive-v2/en-de/de/file_117520_cleared.txt <TAB>
0.0835987674256443 <NEWLINE>
```

4. Document Alignments for English-Slovenian

4.1 Corpus Compilation

For English-Slovenian we have document-aligned the EN-SL part of the USFDNews corpus in order to test the document alignment tools. USFDNews is presented in the following ACCURAT deliverables:

- ACCURAT Deliverable D3.6 "Comparable corpora for under-resourced languages", from which we learn (Table 1, page 6) that for EN-SL, this corpus contains 2.237 English documents, 1.225 Slovenian documents with 1.043.117 words in English and 299.700 words in Slovenian;
- ACCURAT Deliverable D3.4 "Report on methods for collection of comparable corpora" where, at page 6, section 2.1.2 we find the document alignment method. Basically, these documents were aligned based on the same date/time interval, lengths and similarity of their titles, etc. (see the deliverable for more details). This way, a number of 3.642 document pairs were generated.

4.2 Document Alignment and Statistics

Table 3 Quantitative description of the English-Slovenian and Slovenian-English document alignments

	Total aligns.	**Avg. 1:n**	**Avg. n:1**	**Avg. score**	**Max. score**	**Min. score**	**% greater than avg.**	**% smaller than avg.**
USFDNews (EN-SL) EMACC	1174	1	1	1.07e-12	3.75e-11	2.62e-16	20.5%	79.5%
USFDNews (EN-SL) ComMetric	348	1.77	4	0.56	0.79	0.5	35.4%	64.6%
USFDNews (SL-EN) EMACC	1174	1	1	1.09e-12	6.02e-11	2.61e-16	20.4%	79.6%
USFDNews (SL-EN) ComMetric	145	3.45	1.06	0.56	0.77	0.5	34.5%	65.5%

EMACC was run with the default settings and these settings imposed the generation of 1:1 document alignments which do not correctly reflect the nature of this corpus. Only 20% of the alignments are above the average and approx. 34% in the case of the ComMetric algorithm. These values also suggest that we are dealing with a weakly comparable corpus.

Figures 3 and 4 plot the alignment probabilities sorted in descending order. ComMetric ran with 0.5 document alignment probability threshold (only document pairs with probability over 0,5 are output) and this is why its distribution does not resemble the EMACC one.

4.3 Document Alignment Location and Format

The English-Slovenian and Slovenian-English document alignments are found on the ACCURAT FTP Server under the "/WP2/D2.4" directory. File name conventions and format are the same as in case of English-Romanian document alignments from section 2.3.

For instance, a line in the '`en-sl-USFDNews-emacc-docalign.lst`' file looks like

```
newCrawl13-06-11\en-sl\en\file_en_3.txt <TAB>
newCrawl13-06-11\en-sl\sl\file_sl_2.txt <TAB>
3.75143816e-11 <NEWLINE>
```

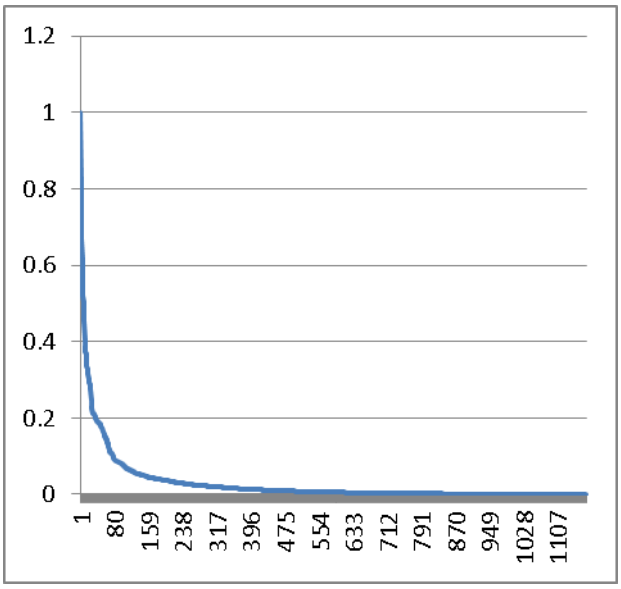

Figure 3: EMACC on EN-SL USFDNews

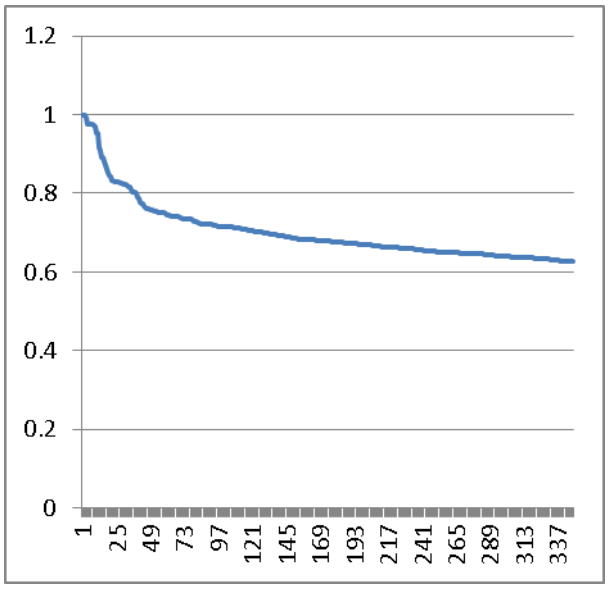

Figure 4: ComMetric on EN-SL USFDNews

5. Document Alignments for English-Latvian

5.1 Corpus Compilation

The following corpora have been document aligned in English-Latvian using ComMetric or its newer version, DicMetric:

- ILSPNewsDisaster (EN-LV) (aligned with ComMetric), which, according to deliverable D3.7 "Comparable corpora for narrow domains", has 24.555 documents in English with 25.806.107 words and 2.354 documents in Latvian with 1.812.255 words;
- ILSPNewsPolitical (EN-LV) (aligned with DicMetric) which, from deliverable D3.7, is known to include 9.494 documents in English with 24.939.703 words and 2.614 documents in Latvian with 6.617.927 words;
- ILSPNewsSports (EN-LV) (aligned with DicMetric) which, from the same deliverable, we find that it contains 6.936 documents in English with 8.818.092 words and 2.867 Latvian documents with 2.579.763 words;
- ILSPNewsTechnological (EN-LV) (aligned with ComMetric) which, from the afore-mentioned deliverable, we learn that it has 34.947 documents in English with 25.773.623 words and 2.534 Latvian documents with 3.126.694 words;

- ILSPRenewableEnergy (EN-LV) (aligned with ComMetric) corpus which contains 14.745 English documents with 18.925.084 words and 431 Latvian documents with 445.771 words according to the same D3.7 document;
- ILSPITLocalisation corpus (EN-LV) (aligned with ComMetric) with 7.723 English documents containing 4.207.731 words and 1.086 Latvian documents with 1.542.631 words (source: D3.7);
- USFDNews (EN-LV) (alignments provided with the corpus) corpus which, according to the ACCURAT Deliverable D3.6 "Comparable corpora for under-resourced languages", has 1.621 documents in English with 839.807 words and 770 documents in Latvian with 203.173 words.

The ILSP corpora are narrow domain bilingual corpora and have been collected by launching two focused monolingual crawlings ,while USFD corpora have been collected with the multilingual criteria in mind.

5.2 Document Alignment and Statistics

Table 4 English-Latvian document alignments: quantitative analysis of the corpora type.

	Total aligns.	**Avg. 1:n**	**Avg. n:1**	**Avg. score**	**Max. score**	**Min. score**	**% greater than avg.**	**% smaller than avg.**
ILSPNews Disasters (EN-LV)	2911	2.13	15.40	0.64	0.88	0.6	35.4%	64.6%
ILSPNews Political (EN-LV)	870	2.6	4.39	0.63	0.87	0.6	35.2%	64.8%
ILSPNews Sports (EN-LV)	35	1.94	2.05	0.66	0.82	0.6	42.8%	57.2%
ILSPNews Technology (EN-LV)	18595	4.41	55.34	0.63	0.91	0.6	39.2%	60.8%
ILSP Renewable Energy (EN-LV)	2400	2.32	104.34	0.65	0.85	0.6	39.8%	60.2%
ILSP ITLocalisation (EN-LV)	5335	3.15	13.47	0.56	0.97	0.5	38.4%	61.6%
USFD News (EN-LV)	7399	1.50	3.16	1	1	1	0	0

Studying the values from Table 4 we once again arrive at the conclusion that we dealt with weakly comparable corpora. It's clear that with an alignment productivity of 104 or 55 source documents per target document, in case of ILSPRenewableEnergy and ILSPNewsTechnological, we speak of weakly comparable corpora or about documents that are almost identical in the source language. And, because of the threshold (0.6 or 0.5) that was imposed at the document alignment stage, a clear picture of alignment probability distribution is not ready. We suspect that without the threshold, this distribution would look like the one in Figure 3. The USFDNews corpus was pre-aligned by USFD using techniques described in D3.4.

5.3 Document Alignment Location and Format

The English-Latvian document alignments are found on the ACCURAT FTP Server under the "/WP2/D2.4" directory. File name conventions and format are the same as in case of English-Romanian document alignments from section 2.3.

For instance, a line in the '`en-lv-ILSPNewsDisasters-commetric-docalign.lst`' file looks like

```
./EN/TextFiles/file_18321_cleared.txt <TAB>
./LV/TextFiles/file_69_cleared.txt <TAB>
0.7746 <NEWLINE>
```

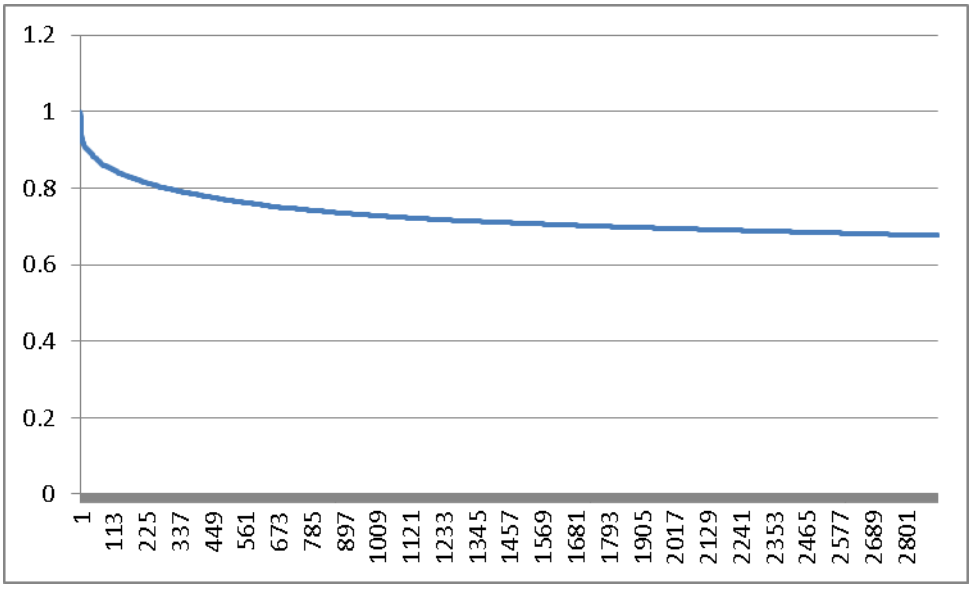

Figure 5: English-Latvian document alignments on ILSPNewsDisasters corpus. The shape of the alignment probability distribution for the rest of the ILSP corpora is the same except for ILSPITLocalisation.

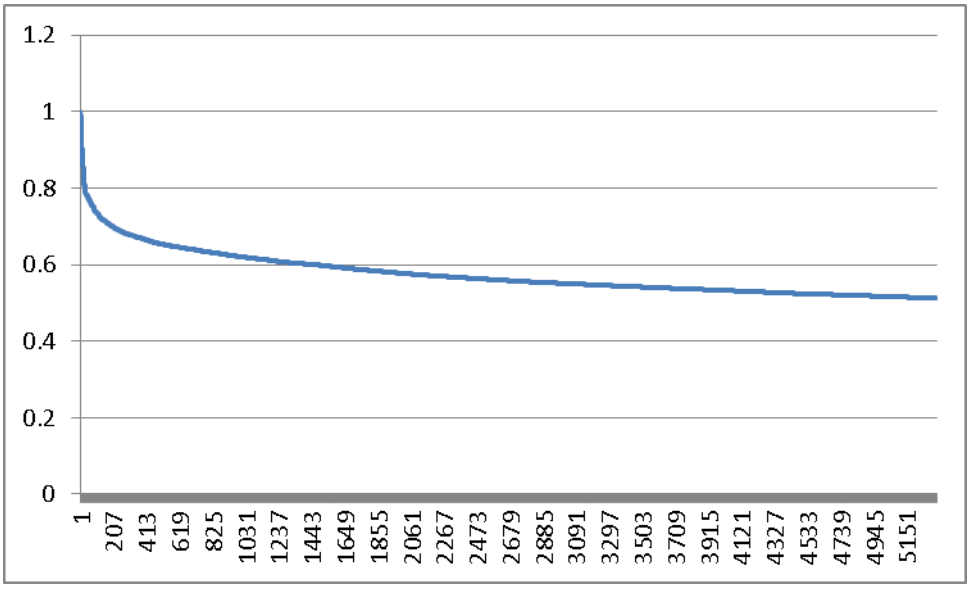

Figure 6: English-Latvian document alignments on ILSPITLocalisation corpus.

6. Document Alignments for English-Lithuanian

6.1 Corpus Compilation

The following corpora have been document aligned in English-Lithuanian using ComMetric or its newer version, DicMetric:

- ILSPNewsDisaster (EN-LT) (aligned with ComMetric) which, from deliverable D3.7 "Comparable corpora for narrow domains", we learn that it has 24.555 documents in English with 25.806.107 words and 3.000 documents in Lithuanian with 2.743.245 words;
- ILSPNewsPolitical (EN-LT) (aligned with DicMetric) which, from deliverable D3.7, we learn that it has 9.494 documents in English with 24.939.703 words and 888 documents in Lithuanian with 1.569.048 words;
- ILSPNewsSports (EN-LT) (aligned with DicMetric) which, from the same deliverable, we find that it has 6.936 documents in English with 8.818.092 words and 3.479 Lithuanian documents with 3.479.530 words;
- ILSPNewsTechnological (EN-LT) (aligned with ComMetric) which, from the afore-mentioned deliverable, we learn that it has 34.947 documents in English with 25.773.623 words and 3.285 Lithuanian documents with 2.564.797 words;
- ILSPRenewableEnergy (EN-LT) (aligned with ComMetric) corpus which has 14.745 English documents with 18.925.084 words and 467 Lithuanian documents with 614.997 words according to the same D3.7 document;
- USFDNews (aligned with ComMetric) corpus which, according to the ACCURAT Deliverable D3.6 "Comparable corpora for under-resourced languages", has 1.225 documents in English with 579.199 words and 568 documents in Lithuanian with 166.856 words.

The ILSP corpora are narrow domain corpora and have been collected monolingually while USFD corpora have been collected with the multilingual criteria in mind.

6.2 Document Alignment and Statistics

Table 5 English-Lithuanian document alignments: quantitative analysis of the corpora type.

	Total aligns.	**Avg. 1:n**	**Avg. n:1**	**Avg. score**	**Max. score**	**Min. score**	**% greater than avg.**	**% smaller than avg.**
ILSPNews Disasters (EN-LT)	15503	6.97	35.72	0.63	0.86	0.6	40%	60%
ILSPNews Political (EN-LT)	131	1.45	2.56	0.64	0.82	0.6	36.6%	63.4%
ILSPNews Sports (EN-LT)	90	3	1.25	0.63	0.77	0.6	37.8%	62.2%
ILSPNews Technology (EN-LT)	18893	5.29	47.23	0.63	0.86	0.6	39%	61%
ILSP Renewable Energy (EN-LT)	7886	4.52	88.6	0.65	0.86	0.6	40.5%	59.5%
USFD News (EN-LT)	1053	3	4.53	0.55	0.9	0.5	34.95%	65.05%

For English-Lithuanian, the shape of the alignment probability distribution is identical as in the case of English-Latvian so, we are also inclined to believe that we deal with weakly comparable corpora. A notable difference is that the English-Lithuanian USFDNews corpus has been aligned with the ComMetric algorithm and thus, our intuition that USFDNews will have the same distribution as the ILSP corpora is supported by the evidence in Figure 8.

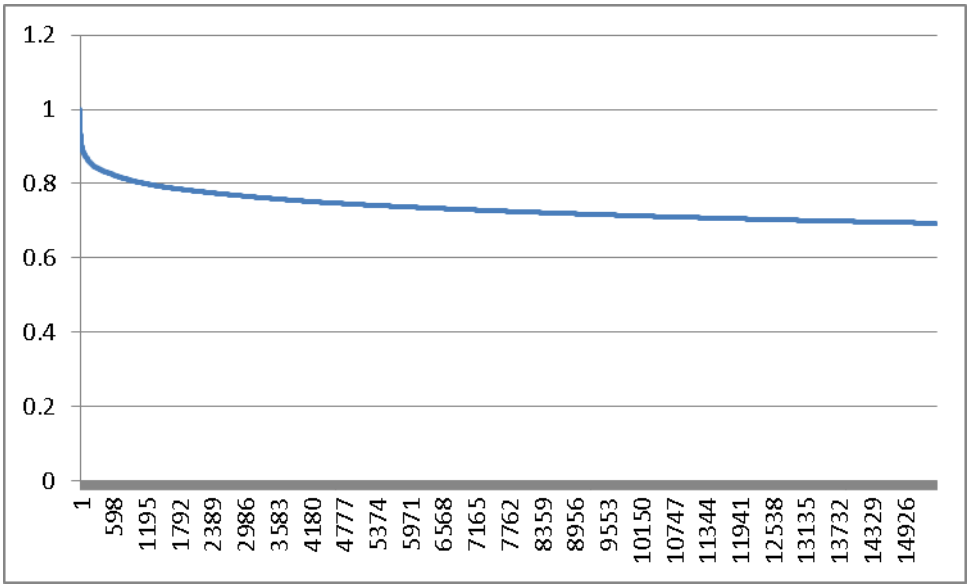

Figure 7: English-Lithuanian document alignments on ILSPNewsDisasters corpus. The shape of the alignment probability distribution for the rest of the ILSP corpora is the same.

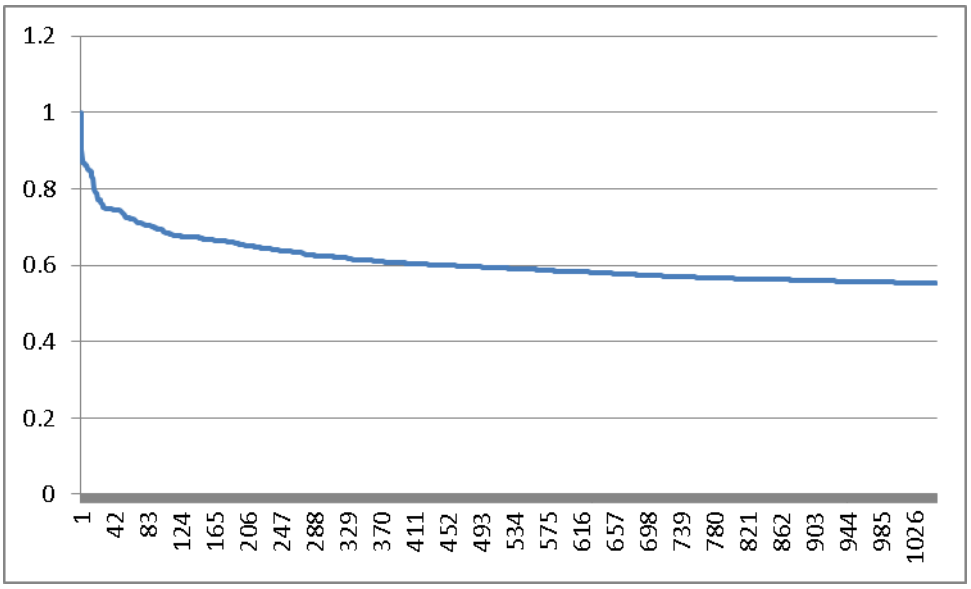

Figure 8: English-Lithuanian document alignments on USFDNews corpus.

6.3 Document Alignment Location and Format

The English-Lithuanian document alignments are found on the ACCURAT FTP Server under the "/WP2/D2.4" directory. File name conventions and format are the same as in case of English-Romanian document alignments from section 2.3.

For instance, a line in the '`en-lt-ILSPNewsDisasters-commetric-docalign.lst`' file looks like

```
./EN/Texts/file_55693_cleared.txt <TAB>
./LT/Texts/file_105255_cleared.txt <TAB>
0.7421 <NEWLINE>
```

7. Document Alignments for English-Estonian

7.1 Corpus Compilation

We have aligned the ILSPRenewableEnergy narrow domain (EN-ET) corpus and the USFDNews corpus. These corpora are characterized as follows:

- ILSPRenewableEnergy (EN-ET) (aligned with ComMetric) has 14.745 English documents with 18.925.084 words and 811 Estonian documents with 734.699 words according to ACCURAT Deliverable D3.7 "Comparable corpora for narrow domains";
- USFDNews (EN-ET) (aligned with ComMetric), according to the ACCURAT Deliverable D3.6 "Comparable corpora for under-resourced languages", has 661 documents in English with 292.130 words and 254 documents in Lithuanian with 37.274 words.

7.2 Document Alignment and Statistics

Table 6 English-Estonian document alignments: quantitative analysis of the corpora type.

	Total aligns.	Avg. 1:n	Avg. n:1	Avg. score	Max. score	Min. score	% greater than avg.	% smaller than avg.
ILSP Renewable Energy (EN-ET)	3777	3.11	151.08	0.64	0.85	0.6	40.3%	59.7%
USFD News (EN-ET)	214	1.13	2.4	0.57	0.99	0.5	41.6%	58.4%

The situation for English-Estonian is not different than what we have seen so far. Figures 9 and 10 suggest that the aligned corpora are weakly comparable.

7.3 Document Alignment Location and Format

The English-Estonian document alignments are found on the ACCURAT FTP Server under the "/WP2/D2.4" directory. File name conventions and format are the same as in case of English-Romanian document alignments from section 2.3.

For instance, a line in the '`en-et-USFDNews-commetric-docalign.lst`' file looks like

```
./newCrawl15-09-11/en-et/en/file_en_197.txt <TAB>
./newCrawl15-09-11/en-et/et/file_et_62.txt <TAB>
0.707 <NEWLINE>
```

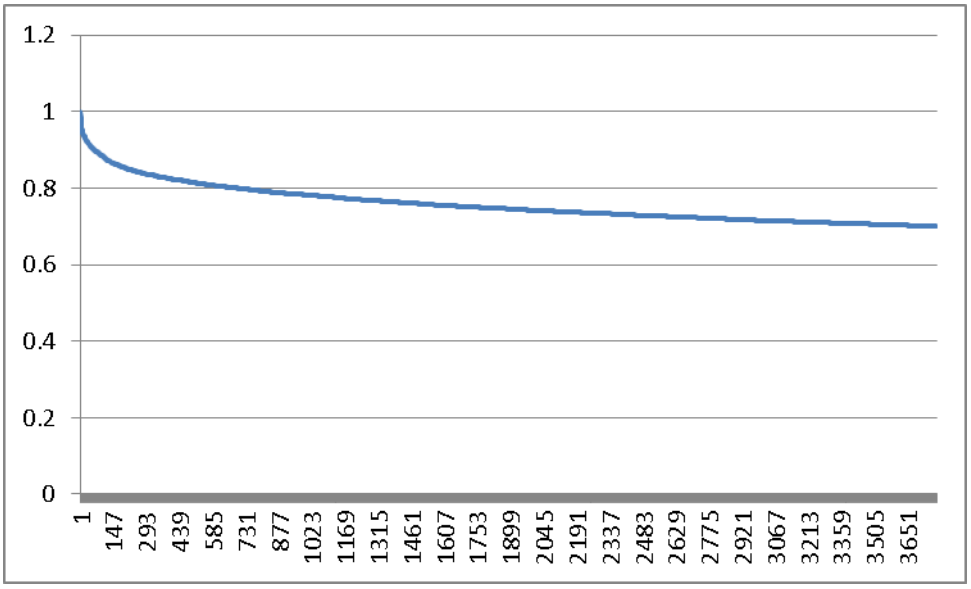

Figure 9: English-Estonian document alignments on ILSPRenewableEnergy corpus.

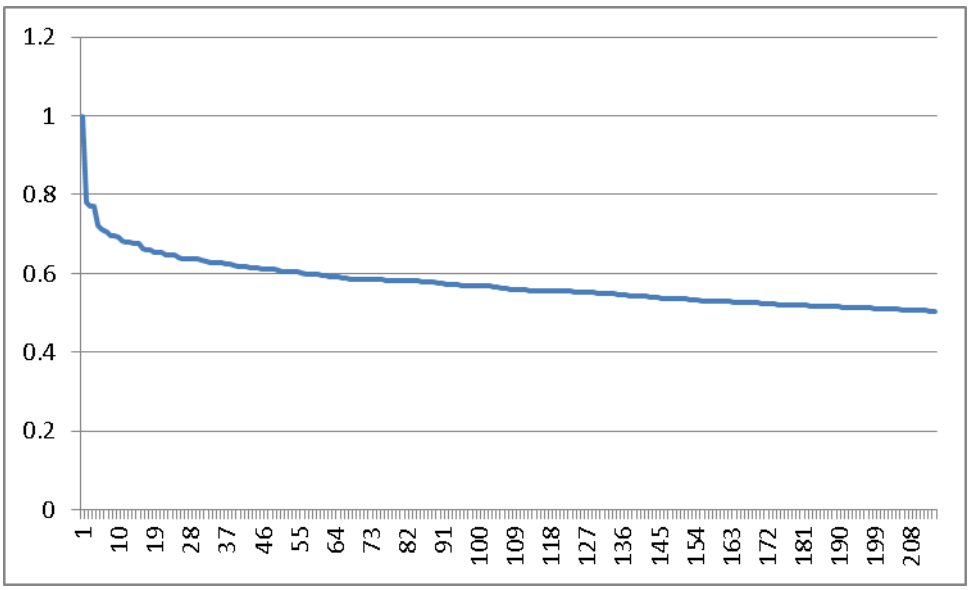

Figure 10: English-Estonian document alignments on USFDNews corpus.

8. Document Alignments for English-Greek

8.1 Corpus Compilation

For aligning English – Greek (EN-EL) documents, two bilingual comparable corpora have been used: one on the Topical News – Disasters narrow domain and another on News domain in general. These corpora are abbreviated as ILSPNewsDisasters and USFDNews and are briefly described as follows:

- ILSPNewsDisasters (EN-EL) includes 24.555 English documents comprising of 25.806.107 tokens and 5.512 Greek documents comprising of 8.753.473 tokens. This bilingual comparable corpus is further documented in the ACCURAT Deliverable D3.7 “Comparable corpora for narrow domains”.

- USFDNews corpus (EN-EL) contains crawled data from various internet news sites, as detailed in the ACCURAT Deliverable D3.6 "Comparable corpora for under-resourced languages". For the EN-EL documents alignment task the most recent crawled data have been used, that is 2.936 EN documents containing 1.287.598 words and 1.603 EL documents containing 350.456 words.

Both corpora have been document - aligned by means of the DicMetric tool. Only EN-EL document pairs that received a minimum of 0.3 comparability score (alignment probability) have been considered for the statistics given in Table 7 that follows.

8.2 Document Alignment and Statistics

Table 7 English-Greek document alignments: quantitative analysis of the corpora type.

	Total aligns.	**Avg. 1:n**	**Avg. n:1**	**Avg. score**	**Max. score**	**Min. score**	**% greater than avg.**	**% smaller than avg.**
ILSPNewsDisasters (EN-EL)	730	4.87	10.80	0.33	0.55	0.3	34.66%	65.34%
USFDNews (EN-EL)	239	2.59	6.64	0.36	0.64	0.3	39.33%	60.67%

The situation for English-Greek is similar to what we have seen so far. Figures 11 and 12 suggest that the aligned corpora are weakly comparable.

8.3 Document Alignment Location and Format

The English-Greek document alignments are found on the ACCURAT FTP Server under the "/WP2/D2.4" directory. File name conventions and format are the same as in case of English-Romanian document alignments from section 2.3.

For instance, a line in the '`en-el-USFDNews-dicmetric-docalign.lst`' file looks like

```
./Crawl01-08-11/en-el/en/file_en_20.txt <TAB>
./Crawl01-08-11/en-el/el/file_el_9.txt <TAB>
0.33 <NEWLINE>
```

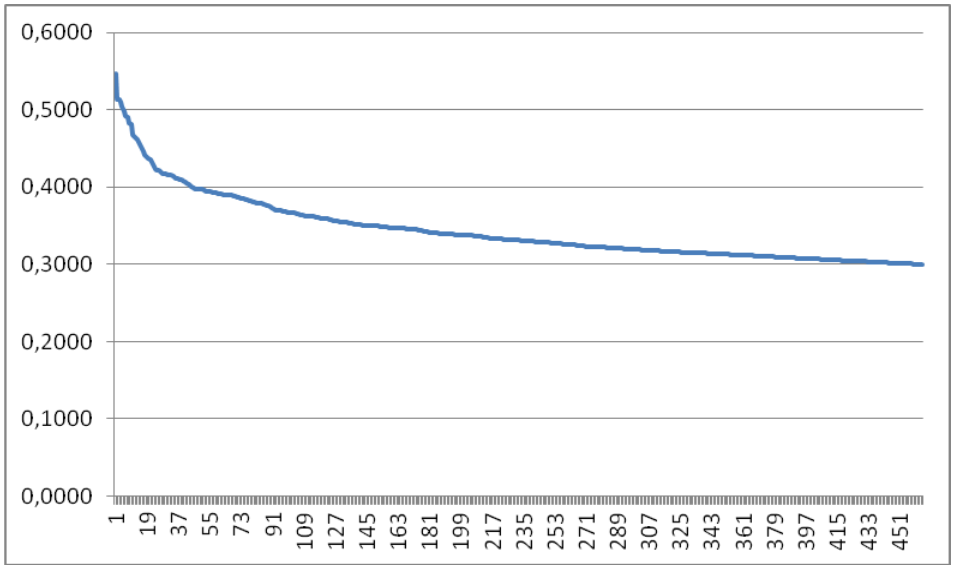

Figure 11: English-Greek document alignments on ILSPNewsDisasters corpus.

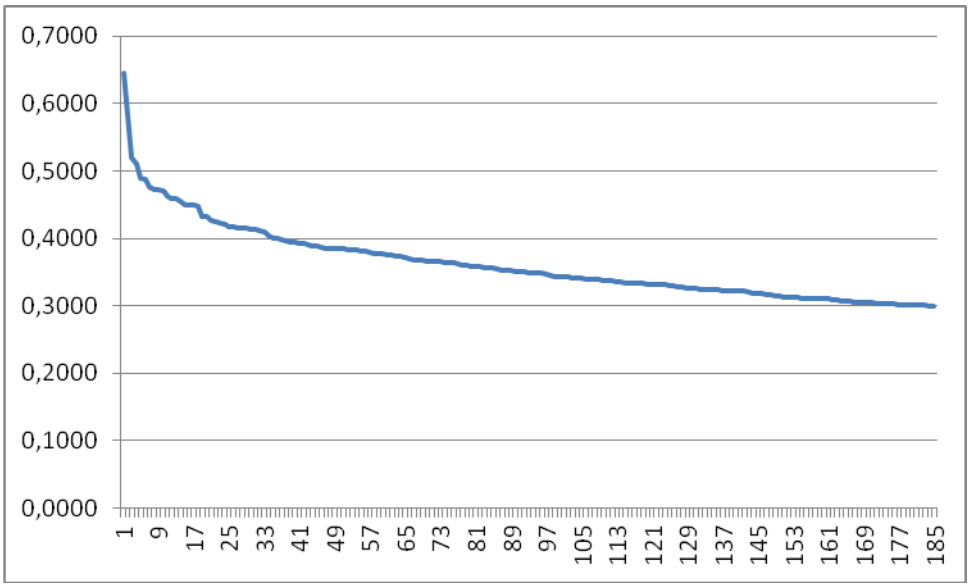

Figure 12: English-Greek document alignments on USFDNews corpus.

9. Conclusions

We have automatically aligned documents from a wide range of comparable corpora that have been collected in the ACCURAT project and this report was a good opportunity to study the collected corpora in order to assess their comparability level. We have hinted at a, to be further developed, methodology of graphing the comparability degree of a corpus based on the alignment probability distribution. Here, we attempt to refine this methodology.

The intuition about a comparable corpus containing M documents in the source language and N documents in the target language is that:

- if the corpus is parallel (and complete), then M = N and an accurate document alignment methodology would have to provide only 1:1 document alignments with high alignment probabilities. It also means that our measures of document alignment productivity “Avg. 1:n” and “Avg. n:1” described in section 2.2 of this document would also have to be close to 1;

- if the corpus is strongly comparable, M and N are not necessarily closer to one another, but the document alignment probabilities would still have to be high. The document alignment productivity is now greater than 1, but not much greater;
- finally, in the case of weakly comparable corpus, M and N are very different, the document alignment productivity would be large and relatively few pairs of documents will have their alignment probabilities greater than the average on the whole alignment set.

It is clear that the judgements above are heavily dependent on the ability of the document alignment technique to actually detect and assign high alignment probabilities to parallel or strongly comparable document pairs and to assign low alignment probabilities to unrelated document pairs.

In the next figure we shall attempt to graph the alignment probability distribution for the English-Estonian ILSPRenewableEnergy corpus in parallel with its document alignment productivity.

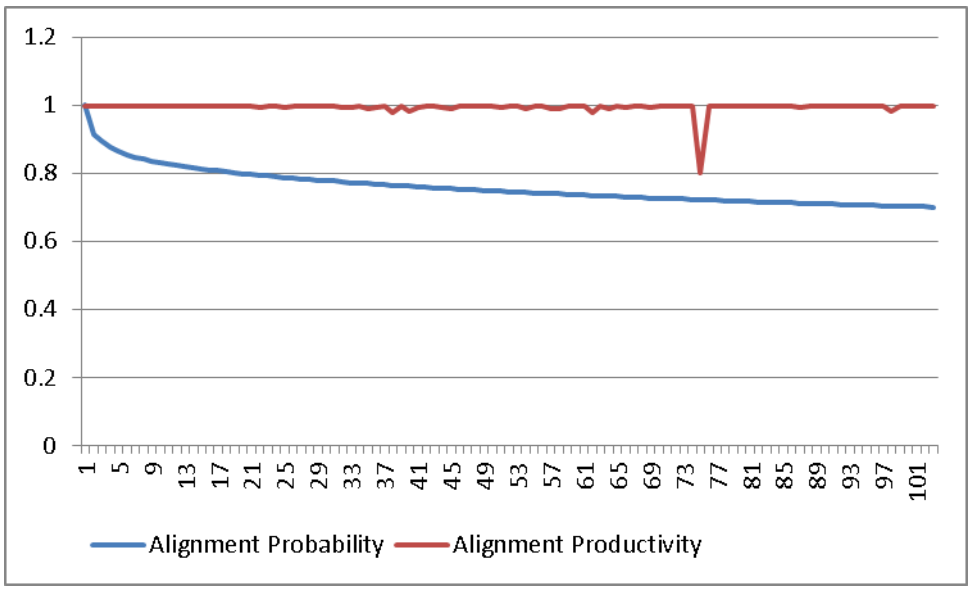

Figure 13: English-Estonian ILSPRenewableEnergy corpus comparability judgement.

In Figure 13 we have plotted the alignment probabilities from the highest to the lowest (the blue line), and, for every document pair (let's call it DP), we have also plotted in red (the red line) the alignment productivity *x* as $f(x) = 1 - \frac{1}{x}$ where x is the maximum between:

- the number of different target documents that align to the source document in DP and
- the number of different source documents that align to the target document in DP.

Thus, for a parallel corpus, the alignment probability (blue line) would have to be an almost straight line close to $y = 1$ and the alignment productivity (red line) would have to be an almost straight line close to y = 0 because of the fact that $f(1) = 0$. For a strongly comparable corpus, the alignment productivity (red line) will stay close to y = 0.5 or y = 0.66 for productivity levels of 2 and 3.

10. References

ACCURAT Deliverable D2.6 Toolkit for multi-level alignment and information extraction from comparable corpora, version 1.0, August 31, 2011.

ACCURAT Deliverable D3.4 Report on methods for collection of comparable corpora, version 1.0, October 31, 2011.

ACCURAT Deliverable D3.6 Comparable corpora for under-resourced languages, version 1.0, October 31, 2011.

ACCURAT Deliverable D3.7 Comparable corpora for narrow domains, version 1.0, October 31, 2011.