ACCURAT

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

www.accurat-project.eu

Project no. 248347

Deliverable D2.5
Extracted data for translation models of SMT and RBMT lexicon from aligned comparable corpora

Version No. 2.0
31/03/2012

Document Information

Deliverable number:	D2.5
Deliverable title:	Extracted data for translation models of SMT and RBMT lexicon from aligned comparable corpora
Due date of deliverable:	30/11/2011
Actual submission date of deliverable:	31/03/2012
Main Author(s):	Radu Ion, Mārcis Pinnis, Gregor Thurmair, Ahmet Aker, Rob Gaizauskas, Mateja Verlic, Nikos Glaros
Participants:	RACAI, TILDE, USFD, LINGUATEC, ZEMANTA, ILSP
Internal reviewer:	TILDE
Workpackage:	WP2
Workpackage title:	Multi-level alignment methods and information extraction from comparable corpora
Workpackage leader:	RACAI
Dissemination Level:	PU
Version:	V2.0
Keywords:	parallel data from comparable corpora, translation lexicons from comparable corpora, translated terminology from comparable corpora, translated named entities from comparable corpora

History of Versions

Version	Date	Status	Name of the Author (Partner)	Contributions	Description/ Approval Level
V0.1	08/11/2011	Draft	RACAI	Tentative Table of Contents	Created.
V0.2	29/11/2011	Draft	RACAI	RACAI	Description and statistics about the executed parallel data extraction workflow
V0.3	29/11/2011	Draft	RACAI	TILDE	Added Tilde's contribution about terminology mapping
V0.4	29/11/2011	Draft	RACAI	USFD	Added USFD's contribution about named entities mapping
V0.5	29/11/2011	Draft	RACAI	RACAI	Added Conclusions and Introduction

Version	Date	Status	Name of the Author (Partner)	Contributions	Description/ Approval Level
V0.6	30/11/2011	Draft	RACAI	RACAI, USFD, ZEMANTA, ILSP, LINGUATEC	Added section 1.2 and addressed comments from everybody.
V0.61	30/11/2011	Draft	RACAI	ILSP	Tables 5 and 6 fixes and some English corrections.
V1.0	30/11/2011	Final	Tilde	Final review	Submitted to PO
V1.1	16/03/2012	Draft	RACAI	RACAI	Addressed reviewer's comment no. 1
V2.0	31/03/2012	Final	Tilde	Final Reviw	Submitted to PO

EXECUTIVE SUMMARY

This report presents MT-related data that have been obtained from the aligned comparable corpora described in D2.4. By "MT-related data" we understand parallel texts, translated terminology and named entities. For each pair of languages of interest, we give quantitative and tentative qualitative analyses of the extracted data.

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
API	Application programming interface
CC	Comparable Corpora
CPU	Central processing unit
FMC	Focused Monolingual Crawler
URL	Uniform Resource Locator
MT	Machine Translation
NE	Named Entity
NER	Named-entity recognition
SMT	Statistical Machine Translation
RBMT	Rule Based Machine Translation
XML	Extensible Markup Language

Introduction

One of the main objectives of the ACCURAT project is to provide MT-related data to different types of MT systems, such as SMT or RBMT, by collecting and extracting these data from Comparable Corpora (CC). It has to be emphasized that extracting parallel data from CC is much more computationally intensive and algorithmically demanding than doing so for parallel corpora. This can be mainly attributed not only to the several orders of magnitude larger textual data volumes that have to be processed sequentially in the CC case, but also to the fact that there are many more positional possibilities that translation equivalents can have in a pair of aligned documents when these aligned documents are comparable rather than when they are parallel.

In what follows, we present the ACCURAT consortium's first attempt to do parallel data mining on CC in order to obtain parallel data that can benefit MT systems performance. This is a large-scale integration step of all of our tools responsible for collecting CC, aligning documents and finally, extracting parallel data and creating bilingual terminology and named entities lists from CC.

1 Parallel Data Extraction from Comparable Corpora

1.1 About the extraction process

After the document alignments described in the ACCURAT Deliverable D2.4 “Aligned comparable corpora” (CC) have been obtained, we proceeded to the parallel data mining step. This step was performed in order to assess how much/how good parallel data we are able to obtain from the CC that we collected. It is important to point out that this is not the last parallel data mining process we will attempt to obtain statistical machine translation (SMT)-ready training data, but, more importantly, this is the first large-scale integration step of all of our tools responsible for collecting CC, aligning documents and, finally, extracting parallel data from it.

In order to obtain document alignments, we have employed the services of the following three tools that were developed especially for this purpose in the ACCURAT project:

- EMACC, an Expectation-Maximization-based document aligner (described in the D2.6 “Toolkit for multi-level alignment and information extraction from comparable corpora” deliverable, section 2.3);
- ComMetric, a translation-based (Google and/or Bing) comparability metric tool (see D2.6 “Toolkit for multi-level alignment and information extraction from comparable corpora”, section 2.1);
- DicMetric, a refined version of ComMetric that does not rely on external translation services. However, it should be noted that an iterative procedure where existing internal SMT systems are improved with extracted data and then are used to detect new pairs of strongly comparable documents is a feasible scenario.

When the documents alignments were ready, we have used two tools to mine for parallel data in pairs of documents that we deemed as “comparable” by imposing certain thresholds on document alignment probabilities (0.3, 0.6 on ComMetric or 0.001 for EMACC – see deliverable D2.4 “Aligned comparable corpora” for a description of the document alignments):

- PEXACC, a translation-similarity algorithm for mining parallel data from CC (see Deliverable D2.2 “Report on multi-level alignment of comparable corpora”, section 2.2)
- MEExtract, a Maximum Entropy parallel sentence classifier (see Deliverable D2.2 “Report on multi-level alignment of comparable corpora”, section 2.3)

Each partner has run the selected document alignment tool and parallel data mining tool according to available computational resources. EMACC and PEXACC are CPU intensive tools that actually perform (almost) brute force searches for the best document alignments/parallel phrases. To quantify this assertion, we did a comparative assessment of the running times of PEXACC and MEExtract (tools from the category of parallel data mining from CC) on the latest extracted English-Romanian News comparable corpus by the USFD partner. The next table presents the corpus statistics.

Table 1: Statistcs on the English-Romanian USFD News corpus, version 14-02-12

	Documents	Sentences	Tokens	Size
en	17,845	464,961	9,309,338	53.7MB
ro	7,120	121,104	2,605,976	16.9MB

Both PEXACC and MEExtract ran on an Intel i7 980 @ 3.33GHz, 16 GB DDR3 @ 800MHz machine but PEXACC, with its parallel implementation, ran on all 12 cores of the mentioned CPU. The processing times were:

- PEXACC: 280 minutes; a single-core run was reported by the Windows Task Manager program as being nearly 30 hours.
- MEExtract: a single-core run of 33 minutes.

Unfortunately, comparing the outputs given by the two programs, we saw that, once again, MEExtract performed unsatisfactory on this corpus. The choise to use PEXACC (even with its high computation demands) was based on multiple tests with both tools done by all partners.

1.2 Statistics of the extracted parallel data

Table 1 contains the sizes of the extracted parallel data for the following language pairs: English-Romanian (en-ro), English-Latvian (en-lv), English-Lithuanian (en-lt), English-Estonian (en-et), English-Greek (en-el), English-Slovenian (en-sl) and English-German (en-de).

Table 2: PEXACC data (both sentences and chunks) on ACCURAT language pairs with parallelism threshold set to 0.3 and above and with filtered identical textual unit pairs

Lang. pair	Corpus	Unique counts		
		Pairs	Src. tokens	Trg. tokens
en-de	ILSPAutomotiveV2	37427	132862(3.3%)	133058(2.5%)
en-el	USFDNews	27	141(0%)	139(0%)
en-el	ILSPNewsDisasters	1512	5401(0%)	5441(0%)
en-et	ILSPRenewableEnergy	705	3193(0%)	3205(0%)
en-et	USFDNews	78	565(-)	555(-)
en-lt	ILSPNewsDisasters	9347	81012(0.2%)	88572(2.55%)
en-lt	ILSPNewsPolitical	359	2818(0%)	3044(0.15%)
en-lt	ILSPNewsSports	157	1099(0%)	1188(0%)
en-lt	ILSPNewsTechnological	24068	175054(0.5%)	190560(5.1%)
en-lt	ILSPRenewableEnergy	4827	37099(0.15%)	39827(4.2%)
en-lt	USFDNews	164	1509(-)	1458(-)
en-lv	ILSPITLocalisation	5818	46539(0.95%)	49833(2.9%)
en-lv	ILSPNewsDisasters	2177	16593(0%)	16920(0.75%)
en-lv	ILSPNewsPolitical	5627	68282(0.15%)	70440(0.95%)
en-lv	ILSPNewsSports	64	606(0%)	625(0%)
en-lv	ILSPNewsTechnological	20731	157941(0.45%)	173908(4.1%)
en-lv	ILSPRenewableEnergy	1772	11908(0%)	12854(1.85%)

Lang. pair	Corpus	Unique counts		
		Pairs	Src. tokens	Trg. tokens
en-lv	USFDNews	387	3927(-)	4050(-)
en-ro	USFDNews	8461	158099(13.4%)	165799(29.35%)
en-ro	ILSPNewsDisasters	2437933	7979279(5.1%)	8266110(23.18%)
en-sl	USFDNews	26	262(0%)	267(0.1%)
sl-en	USFDNews	11	92(0%)	91(0%)

The counts in Table 1 have been generated by uniquely counting textual unit pairs (source textual unit not identical to the target textual unit) that have a parallelism threshold above 0.3. The "identical" filter was imposed because of the fact that target documents contained English sentences and/or chunks that were found exactly in the English part of the corpus. We relaxed this filter to also take into account identical sentences and/or chunks with differing punctuation. One point to mention is that the significant count difference of the en-ro ILSPNewsDisasters corpus is due to the fact that we extracted chunk-level alignments.

In the source tokens ("Src. tokens") and target tokens ("Trg. tokens") columns, in parenthesis, there is an estimate of the percent of tokens that are found in the extracted parallel textual unit from the total number of tokens in the monolingual corpora per language (the sizes of the monolingual corpora are listed in the ACCURAT Deliverables D3.6 and D3.7). This is to show what percent of collected CC can be used as parallel data so as to make an estimate of the size a particular CC must have to obtain a parallel corpus of a given size.

Regarding the 0.3 threshold, this is a good threshold for sentence level extraction while for chunk level extraction, 0.6 is a more suitable choice. All resulting pairs with parallelism probabilities above these thresholds can be considered parallel or almost parallel. Please note that the pairs of phrases with a score below 0.6 may still be useful for tasks such as named entity or terminology mapping.

2 Translated Terminology

In this section of the deliverable a description of two approaches to acquiring both bilingually mapped terminology dictionaries and bilingual terminology dictionaries from comparable corpora is provided. The first approach uses mostly tools developed as part of the ACCURAT project published within Deliverables D2.6 and D3.5 in order to create a comparable corpus from the Web, monolingually tag terms in the corpus and then bilingually map the terms. The second approach, on the other hand, starts from aligned phrases and filters term candidates from them.

2.1 The First Approach

The first approach uses the methodology and tools published in previous ACCURAT project's deliverables, more precisely:

- multilingual narrow domain comparable corpora in the disasters news domain were collected using the *Focused Monolingual Crawler* (FMC) developed by the ILSP partner (included in the public Deliverable D3.5 of the ACCURAT project). FMC implements monolingual focused crawling for every specified language and ensures that the collected multilingual narrow domain comparable corpora contain at least weakly comparable documents.

- Comparability metrics were applied to the narrow domain multilingual corpora in order to evaluate the comparability of the collected corpora between different language pairs and to minimize the search space for the more resource intensive term mapping task. The comparability metrics tools were developed as part of the ACCURAT project's public Deliverable D2.6.
- Terms in every monolingual narrow domain corpus were tagged using either existing terminology extraction tools or tools developed as part of the ACCURAT project's Deliverable D2.6. For Greek an existing terminology extraction tool from the ILSP partner was used. For Latvian, Lithuanian and English *Tilde's Wrapper System for CollEx* (developed by the partners from FFZG and Tilde) was used. For Romanian the RACAI partner's terminology extraction tool was used. The latter two were developed as part of the ACCURAT project and are published within Deliverable D2.6.
- Terms in the bilingual comparable corpora were then mapped using two tools developed within the ACCURAT project – the RACAI partner's *TerminologyAligner* and the USFD partner's *MapperUSFD* (both published within the Deliverable D2.6). Note that the MapperUSFD is a cognate based approach and was designed only for mapping named entities. To investigate its potential we also applied it also for mapping terms.

The task of mapped terminology dictionary creation using the first approach was limited to the disaster news narrow domain corpora for the following two reasons:

- The process chain from corpus crawling, comparability estimation, monolingual corpora term tagging and bilingual term mapping is very resource intensive and requires a lot of computing power (the whole processing chain can take from 1 to 2 weeks for a single language pair depending on the monolingual corpora size).
- The corpus provides a wide variety of possible term combinations and, thus, is suitable for a terminology dictionary creation and a demonstration of the capabilities of the methods developed and applied in the ACCURAT project.

2.1.1 Multilingual Narrow Domain Corpora Statistics

The corpora statistics after monolingual corpora crawling with FMC and tagging of terms with the respective language dependent term tagging tools is shown in Table 3.

Table 3: Monolingual disaster news corpora statistics

Language	Documents	Terms	Unique Terms
English	24555	1756576	123721
Greek	5512	772168	273624
Latvian	2354	225506	41445
Lithuanian	3000	375712	69059
Romanian	4497	683282	34956

2.1.2 Aligned Narrow Domain Comparable Corpora Statistics

The comparability metrics were applied for language pairs where English was the source language. The comparable corpora statistics of the processed language pairs is shown in Table 4.

Table 4: Bilingual disaster news comparable corpora statistics

Language pairs	Comparability metric	Document pairs	Terms		Unique Terms	
			Source	Target	Source	Target
English-Greek	DicMetric	730	59985	6051	8361	3436
English-Latvian	ComMetric	2911	226451	30929	24060	9352
English-Lithuanian	ComMetric	15503	315583	59045	27149	13435
English-Romanian	EMACC	10165	707278	594690	68334	33676

Three different comparability metric solutions to align the bilingual narrow domain comparable corpora have been developed. For languages, where third party translation API's (*Google*, *Bing*, others) were available, *ComMetric* (the second version of the comparability metric from the CTS partner) was used. Due to the fact that translation API access was recently heavily limited (by both *Google* and *Microsoft*) *DicMetric* (the third version of the comparability metric from the CTS partner) was used for other languages except for English-Romanian where existing document alignment results from the RACAI partner's *EMACC* tool were used. Although different comparability metrics were used, we showed in ACCURAT D2.4 deliverable "Aligned comparable corpora" that ComMetric, DicMetric and EMACC produce similar results and thus, we believe that the impact of using different metrics on the mapping results is limited.

The total number of document pairs aligned by the metrics is presented in the column "*Document pairs*". The column "*Terms*" contains the total numbers of terms in the bilingual comparable corpora separately for source (in this task it is always English) and target (Greek, Latvian, etc.) languages. The column "*Unique terms*" contains the total number of unique terms in the bilingual narrow domain comparable corpora, which represents the full search space of each bilingual narrow domain comparable corpus. The term mapper's search space, however, is smaller than the unique term total search space as the mappers operate on the aligned document pair level, because mapping of all document pairs in the comparable corpora could cause a drop in precision as uncomparable document pair terms may be wrongly mapped.

2.1.3 Bilingually Mapped Terminology

The bilingual disaster news comparable corpus was processed with at least one terminology mapping tool (for some pairs both tools were not applied due to language specific constraints). The quantitative results of the mapping task are shown in

Table 5.

Table 5: Bilingual terminology mapping statistics

Language pairs	Document pairs	Extracted term pairs	
		MapperUSFD	TerminologyAligner
English-Greek	730	30	--
English-Latvian	2911	1158	489
English-Lithuanian	15503	2750	770
English-Romanian	10165	--	1828

2.1.4 Analysis of results

Both terminology mapping tools produce mapped terms together with a probabilistic score reflecting how confident the tool is that two terms are a translation of each other. The deliverable contains data acquired using the default set up. The default threshold for the confidence score for *MapperUSFD* is 0.5 and for *TerminologyAligner* - 0.6. If the user wants to use the dictionaries, additional filtering may be required depending on the task the user wishes to use the dictionaries for.

Results of the analysis for the English-Latvian pair show that with default values the precision for *MapperUSFD* is 30.83% and for *TerminologyAligner* – 85.89%. As the default thresholds are set relatively low, it is possible to raise precision by applying a higher threshold. For instance, when applying a threshold of 0.7 for both tools, the precision of *MapperUSFD* and *TerminologyAligner* is 95.33% and 96.30% respectively. This, however, is only an indicative value as for different language pairs the thresholds may differ.

The results also show that both tools perform relatively well on unigram term mapping (with default settings the unigram mapping precision is 67.63% and 97.87% for *MapperUSFD* and *TerminologyAligner* respectively), but lack in precision (as well as recall, after manually checking the corpora) for multi-word term mapping. For instance, the error rate for multi-word term mapping with the default settings on the English-Latvian language pair is over 99% for *MapperUSFD* (with 633 mismapped term pairs) and 38.75% for *TerminologyAligner* (with 62 mismapped term pairs). However, the error rates may be lowered if higher thresholds are applied. Furthermore, as noted above, the MapperUSFD tool uses a solely cognate based approach designed for named entity mapping. Clearly this approach is not sufficient for mapping multiword terms. The approach is now being adapted for term mapping by extending it to exploit dictionaries, as well as a higher matching threshold for cognate matching.

2.2 The Second Approach

This approach does filtering first, and then it performs term extraction, i.e., it extracts terms from aligned phrases, not from monolingual texts (the second approach is used by the Linguatec English-German RBMT system which needs mapped terminology with certain specificities). Thus, a comparable corpus of automotive data for the English-German language pair was prepared as it is described below.

2.2.1 Data Collection

Data were prepared in three steps: crawling and document alignment, sentence alignment, and phrase alignment.

2.2.1.1 Crawling and document alignment

Automotive websites in the area of transmissions/gearboxes were identified, and comparable texts were crawled. The link structure and hints on languages in the URLs were used to guide the crawling process. Also, they gave an indication for *document level* alignment as with bilingual crawlers.

After several crawls with improved precision the effort resulted in 116.056 comparable documents.

The next task was to clean the data, boilerplate parts were removed and the documents were converted into XML structured texts, keeping 40.958 English and 40.958 German documents, considered similar enough to be used for further processing.

2.2.1.2 Sentence alignment

It quickly turned out that the data, even when aligned on document level, were far from parallel; most of them were strongly comparable at best.

For alignment, a 'computer-assisted manual' approach was used as the ACCURAT tools were not yet available.

1. The documents were segmented into paragraphs. Those paragraphs, that were found to have the same size (in kB) in the English and German corpora, were deleted from the corpus, because they were considered to be identical, i.e., written in the same language. This was confirmed by sample manual checking.

2. Furthermore the paragraphs were segmented into sentences. The sentences were cleaned, for example all sentences not containing alphanumeric characters deleted, etc. This process resulted in approx. 350.000 sentences for each language.

3. The next step was the sentence alignment with the Hunaligner. As a result, we got ca. 270.000 aligned bilingual sentences. They were filtered:

- sentences with a score less then 0.5 were deleted
- sentences identical in both languages were deleted (the same texts on both sides = the same language)
- many translations (more than 5) for the same sentence were an indicator for wrong alignments; such pairs were also deleted

4. The result: 44.483 parallel sentences

After all these activities, there were 44.483 sentence pairs left; they were used for phrase alignment.

2.2.1.3 Phrase Alignment

Phrase Alignment was done with the resulting sentence pairs. Alignment was done with GIZA++ and MOSES phrase based translation.

These tools built a translation table containing 7.973 mio phrases. This table was used as an input to term mapping.

2.2.2 Term Extraction

Unlike other approaches to term mapping, which first do term extraction and then do term mapping, the present data were computed with Linguatec's *P2G* tool

(*PhraseTable2Glossary*); this tool supports a workflow which does first alignment, and then extracts terms from the aligned data, by applying filters to aligned phrases. The tool works in the following steps:

2.2.2.1 Probability Filter

Only phrases with a frequency > 1 and a probability > 0.6 (P(f|e) in phrase tables are considered. These phrases have been evaluated to have a very good precision (> 0.9).

2.2.2.2 Linguistic Filter

As terms have an internal linguistic structure, such structures can be used as filters; phrases which do not follow one of the linguistic patterns on both source and target side are eliminated.

There are 86 such patterns in German, and about 450 in English.

The two filters reduced the amount of phrases which are really terms to about 16.000 (15.974), meaning that every 500th phrase in the translation table contains a valid term.

2.2.2.3 Term Creation

The identified terms are brought into a proper form, with correct lemmatisation, truecasing, creation of agreements in multiword where necessary, POS annotation, etc.

This step can create duplicates (e.g. if the phrase table contains a term in both singular and plural form, and is normalised to the same lemma); such duplicates were removed in the output list.

2.2.2.4 Term Filter

The resulting terms can be filtered, e.g. to eliminate general purpose terms from a domain-specific (automotive) glossary, or to eliminate unwanted terms with a stop list.

For this delivery, such a filter was not applied, so the full term list as output by P2G is output.

The output format is:

German term <tab> German POS <tab> English term <tab> English POS

2.2.3Tool Evaluation

The P2G tool was evaluated by manually inspecting a random sample of about 2500 terms (about 15% of the output), and the error rates are:

- translation errors (wrong mappings in the phrase tables): 5.68%
- term creation errors (wrong lemmata etc.): 5.0% for German, and 5.6% for English lemmata

As the output terms are bilingual the error rates accumulate; this means that in the term candidate list, about 3 out of 20 entries need human intervention. This seems to be a reasonable ratio.

2.3 Usage

All bilingual term dictionaries acquired in the term mapping task described in this chapter are compressed in a ZIP archive named

"*D2_5_Section_2_TranslatedTerminology.zip*" located under the "WP2/D2.5/Terminology" folder. The archive contains the following dictionaries:

- Focused disaster news term dictionaries:
    - English-Latvian dictionary acquired with *TerminologyAligner* – "*EN-LV_Disaster-News_RACAI-TA_Terms.txt*"
    - English-Latvian dictionary acquired with *MapperUSFD* – "*EN-LV_Disaster-News_MapperUSFD_Terms.txt*"
    - English-Lithuanian dictionary acquired with *TerminologyAligner* – "*EN-LT_Disaster-News_RACAI-TA_Terms.txt*"
    - English-Lithuanian dictionary acquired with *MapperUSFD* – "*EN-LT_Disaster-News_MapperUSFD_Terms.txt*"
- Focused automotive domain dictionary:
    - English-German dictionary acquired with P2G tool by Linguatec – "EN-DE_Automotive_P2G_Terms.txt".

3 Named Entity Mapping

In this section we provide a description of acquiring bilingual mapped named entities from comparable corpora. In our approach we use the USFD NE (Named Entity) mapping tool (MapperUSFD) described in ACCURAT Deliverable D2.6 to map NEs between texts written in different languages. The texts are the comparable corpora described in the ACCURAT Deliverable D3.6 (USFD News and USFD Wikipedia corpora).

3.1 Pre-processing

MapperUSFD takes as input two comparable documents in text format and outputs pair of NEs with scores indicating their level of mapping. The texts on both sides require pre-processing such as sentence splitting and NE tagging. For both English and foreign (non English) documents we use OpenNLP[1] to identify sentence boundaries. Next, on the English text the mapper applies the OpenNLP NER to extract English NEs. On the foreign text side it assumes that the NEs are identified. We used the NER tools described in D2.6 to tagged the foreign documents with NEs. However, please note that the mapper does not always assume that one of the languages must be English. It can also work on pairs of documents where neither of the languages is English. In this case the mapper just assumes that both texts are already NE tagged.

Having both documents NE tagged (PERSON, LOCATION, ORGANIZATION) the mapper uses cognate based methods (see ACCURAT Deliverable D2.3) to map the NEs with each other. Each mapped NE pair is assigned a score between 0 and 1. In this task we restricted the mapper in returning pairs with scores >= 0.5.

The performance of the NE mapping using the MapperUSFD was evaluated in the ACCURAT Deliverable D2.3. In the evaluation we used 6 language pairs (en-de, en-hr, en-lv, en-ro, en-el and en-lt). Accuracy scores on a small sample of 100 document pairs for each of the 6 language pairs ranged from 51 to 93% with a micro-averaged accuracy of 81% -- where we get 81 from adding all the correct + partially correct scores and

[1] 1http://incubator.apache.org/opennlp/

dividing them by the total scores. The score of 51% was obtained for English-Croatian. However, the next lowest language pair achieved a score of 76% shown in general the high performance of the mapper.

3.2 NE-Mapping Results

The results of the mapping task is shown in Table 6 and Table 7. Table 6 contains the results from the USFD news and Table 7 from the USFD Wikipedia comparable corpora. In both tables we list for each language pair the number of documents pairs used in the mapping process, the full output of the MapperUSFD (third column) and the pairs which have a mapping score of 1.0 meaning exact match (fourth column).

Table 6: Bilingual NEs using USFD News comparable corpora

Language pairs	Document pairs	Mapped NEs	Mapped NEs (exact match)
English-Greek	6396	434	397
English-Latvian	2438	245	23
English-Lithuanian	1735	469	92
English-Romanian	11285	2160	1274
English-German	29341	2098	1083

Table 7: Bilingual NEs using USFD Wikipedia comparable corpora

Language pairs	Document pairs	Mapped NEs	Mapped NEs (exact match)
English-Greek	3668	4629	4330
English-Latvian	4273	2561	427
Latvian-Lithuanian	1027	1275	208
English-Lithuanian	10308	2897	1274
English-Romanian	48880	39206	20829
English-German	149891	207715	38435

3.3 Usage

3.3.1 Translated named entities

All bilingual named entities dictionaries acquired in the NE mapping task described in this chapter are compressed in a ZIP archive named "*D2_5_Section_3_NEMapping.zip*" located under the "/WP2/D2.5/NamedEntities" directory on ACCURAT FTP Server. The archive contains the following dictionaries:

- "ne-mappings-news" which contains all the NE dictionaries for language pairs described in Table 5;

- “ne-mappings-wiki” which contains all the NE dictionaries for language pairs from Table 6.

4 Conclusions

We have reported on the results we obtained by running the two workflows that are described in the ACCURAT Deliverable D2.6 “Toolkit for multi-level alignment and information extraction from comparable corpora”:

- The parallel data mining workflow which produces a set of parallel textual units (both sentences and phrases);
- The NE/Term mapping workflow which produces a set of translation dictionaries composed of translated terms and named entities.

Besides the immediate use of these automatically produced resources in SMT and RBMT, this was the first large-scale experiment of acquiring MT-related data for the majority of ACCURAT language pairs using the ACCURAT-developed machinery, and, together with ACCURAT Deliverable D2.4 “Aligned comparable corpora”, this report was the ideal occasion to study the behavior of our tools. We have learned that:

- The sizes of collected CC demand a running time optimization of both document alignment/comparability metrics tools and parallel data mining tools;
- Recall improvement on all data extraction tools is desired since, from Deliverable D2.4 we learned that the vast majority of collected CC are on the “weakly comparable” side;

We will use the data described in this report to selectively improve baseline MT systems (both statistical and rule-based) but taking into account the relative sizes of the baseline training data and CC extracted data. In the case of SMT with Moses, we plan to use multiple translation tables and back-off models in order to add useful translations only if these are not to be found in the baseline model.

5 References

ACCURAT Deliverable D2.3 Report on information extraction from comparable corpora, version 1.0, September 01, 2011.

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.