Publishable summary

1. A summary description of project context and objectives

ACCURAT is a 2.5 year long EU-funded research project that aims to research methods and techniques to overcome one of the central problems of machine translation (MT) – the lack of linguistic resources for under-resourced languages and domains. The main goals are to find, analyze and evaluate novel methods that exploit comparable corpora in order to compensate for the shortage of linguistic resources, and ultimately to significantly improve MT quality for under-resourced languages and narrow domains.

Traditional ways of building SMT engines that produce acceptable translation quality are often not possible for many domain / language combinations. The ACCURAT project is addressing this issue by developing the technology for using comparable corpora as resources for statistical machine translation (SMT) systems. The **key innovation** of the ACCURAT project will be the creation of a **methodology** and **tools** to measure, find, and to use comparable corpora to improve the quality of MT for under-resourced languages and domains.

The **scientific objectives of the ACCURAT** project are to:

- **Create comparability metrics** – to develop the methodology and determine criteria to measure the comparability of source and target language documents in comparable corpora;
- **Research methods for the alignment and extraction** of lexical, terminological and other linguistic data from comparable corpora;
- **Research methods for automatic acquisition** of a comparable corpus from the Web;
- **Measure improvements** from applying acquired data against baseline results from statistic machine translation and rule-based machine translation (RBMT) systems.

2. Work performed since the beginning of the project and the main results achieved so far

2.1 Criteria of comparability and comparability metrics

A key concept of the project is the notion of comparability. In the ACCURAT project comparability can be defined by how useful a pair of documents or segments of text are for machine translation (MT). Therefore initially four levels of comparability were introduced:

- **Parallel corpora** – collections of traditional parallel texts that are either true and accurate translations of each other, or approximate translations with minor variations, which can be aligned on the sentence level.
- **Strongly comparable corpora** – collections of closely related texts reporting the same event or describing the same subject, which typically can be aligned on text level.
- **Weakly comparable corpora** – collections of texts in the same subject domain and genre, but describing different events. These corpora typically cannot be aligned on the text level, but still can contain collections of translation equivalents.
- **Non-comparable:** pairs of tests drawn at random from a pair of very large collections of texts (e.g. the web) in the two languages.

During the first year of the project we have identified an initial set of criteria of comparability that can guide our procedure to construct comparable corpora. We primarily focused on comparability on higher levels (corpus and document comparability), with the task for selecting comparable corpora, texts and paragraphs for further alignment and use within MT. Features that can be used to identify the comparability level of a pair of documents are summarized in Table 1.

Table 1. Features of comparability

Language-dependent Features	Language-independent Features
LM divergence (words, phrases, N-grams)	String match overlap or letter N-gram overlap (without translation)
Cosine similarities (words, phrases, N-grams)	Out-link overlap
Named entities LM divergence / cosine similarities	Number of links to each other
Named entity tags LM divergence / cosine similarities	Genre overlap (binary)
String match overlap or letter N-gram overlap	Domain overlap (binary)
Parts of speech (including POS N-grams, frequent discontinuous POS signatures)	Date proximity
	Document length difference / ratio
	URL character overlap
	URL slash overlap

All language dependent features identified in Table 1 appear to be good indicators of the comparability level of a document pair. In particular an identifiable gap in the values between parallel, strongly comparable, weakly comparable and non-comparable document pairs is present for most of the features. Only, the cosine similarity among term frequencies seems not to be able to separate strongly from weakly comparable.

Language independent features were useful for identifying parallel documents, but do not identify any of the other degrees of comparability well, displaying little difference in the values of these features between comparable and non-comparable documents. One exception is the Image Link Overlap, which appears to identify both strongly comparable and parallel documents.

The features are used for developing and implementing comparability metric, i.e., the software package, which allows to compute multidimensional features and feature combinations on the level of individual texts and the whole corpora that show their level of comparability. The comparability metrics processes corpora and texts in a modular way with the aim to extract and compare features in a unified standard format. The framework generalises the way how features are annotated and extracted from corpora and texts, and allows us to test novel feature combinations of comparability which work best for specific translation directions.

2.2 Alignment methods

The term alignment is used in the context of machine translation to describe the pairing of text in one document with its translation in another. Alignment is commonly performed for texts that are translations of each other, but it is also possible to produce a type of alignment between texts that are not parallel but may be comparable to each other.

We have studied and evaluated existing alignment strategies designed for parallel corpora, comparable corpora, and non-comparable corpora. We focused particularly on the appropriateness of these techniques to corpora of different levels of comparability. A case study was performed making use of four different alignment methods (Giza++, Moses, cognate based alignment, co-occurrence based alignment) and applying them to corpora of different levels of comparability. We tested the accuracy of these methods for alignment of words or phrases by comparing the alignments produced to human word alignments. We showed that the most widely used existing alignment methods (Giza++ and Moses) are not well suited for use directly on strongly or weakly comparable texts, but for parallel corpora it is possible to obtain 85% of the correct alignments using this method. Additionally, we showed that for

weakly comparable corpora it is possible to correctly identify only around 35% of the alignments in text using word co-occurrence information.

A more powerful method we recently developed relies on Expectation Maximization (EM) algorithm for automatic generation of parallel and quasi-parallel data from any degree of comparable corpora ranging from parallel to weakly comparable. The method does not rely on strongly comparable documents already paired, but can work on collection of documents with various comparability degrees. This new algorithm extracts related textual units (documents, paragraphs or sentences) from comparable corpora relying on the hypothesis that, in a given corpus, certain pairs of translation equivalents are better indicators of a correct textual unit correspondence than other pairs of translation equivalents. We evaluated our method on the Initial Comparable Corpora that contain pairs of documents in English and each language of the interest: Romanian, Latvian, Lithuanian, Estonian, Greek, Slovene and German. We are able now to report document alignment precisions between 75% and 91% and paragraph alignment precisions (computed on automatically extracted pairs of paragraphs from the parallel corpus of DGT-TM using the exact same algorithm as in the case of document alignment) between 77% and 95%.

2.3 Methods for building a comparable corpus from the Web

At first the initial comparable corpora (ICC) have been collected for nine ACCURAT language pairs: Estonian-English, Latvian-English, Lithuanian-English, Greek-English, Romanian-Greek, Croatian-English, Romanian-English, Romanian-German and Slovenian-English. Every language corpus in ICC consists of approximately one million words per language, there are some small deviations in word count for lesser used language pair, i.e. Romanian-Greek. Parallel and strongly comparable documents are aligned at the document level. For the German-English language pair texts have been collected for automotive, medicine and software domains.

Several novel approaches how to build a comparable corpus from the Web that are applicable to under-resourced languages have been researched. Initial methods of searching for comparable corpora are implemented and used to automatically gather very large collections of comparable documents for all project language pairs. Specifically, methods have been developed to target the identification of comparable documents on Wikipedia, in narrow domains (like "Wind Energy"), in news documents, on Twitter, and on the Web as a whole. We have used our method for identifying comparable news documents to automatically gather a large collection of comparable news documents in all project language pairs (totalling over 100 million words). This collection was evaluated using initial metrics for measuring comparability. Initial results indicate that a large percentage of the corpora seem to be strongly comparable.

2.4 Comparable corpora in MT systems

To evaluate efficiency and usability of above mentioned methods and techniques for under-resourced languages and narrow domains research results will be integrated into ACCURAT baseline MT systems. The ACCURAT baseline SMT systems have been set up for 17 translation routes: English⇨Latvian, English⇨Lithuanian, English⇨Estonian, English⇨Greek, English⇨Croatian, Croatian⇨English, English⇨Romanian, English⇨Slovenian, Slovenian⇨English, German⇨English, German⇨Romanian, Romanian⇨German, Lithuanian⇨Romanian, Romanian⇨Greek, Greek⇨Romanian, Romanian⇨English, Latvian⇨Lithuanian. The SMT systems are created using existing SMT techniques – Moses decoder, language models and translation models trained on available parallel corpora e.g. JRC-ACQUIS Multilingual Parallel corpus, SETimes corpus. Translations on 17 systems are evaluated in three domains: a mixed domain, consisting test data from SE-times and Acquis; a balanced domain (consisting data from EU, news, legalism, technology etc.) and news domain.

A software infrastructure, MT-serverland, is developed in order to facilitate the access to MT functionality both for research and applications. This software infrastructure provides a client-server architecture in which MT engines for many language pairs, many users, and many domains can be easily accessed through a uniform framework. All 17 ACCURAT statistical machine translation systems can be accessed in the MT-serverland and used by ACCURAT partners.

3. The expected final results

The ACCURAT project will provide researchers and developers with a fully functional model, methodology and tools for exploiting comparable corpora in MT. The key outputs of the ACCURAT project will be:

- Criteria and metrics of comparability and parallelism;
- Toolkit for multi-level alignment and information extraction from comparable corpora;
- Tools for building comparable corpora from the Web;
- Multilingual comparable corpora for under-resourced languages and narrow domains for the project languages;
- Methodology for application of data extracted from comparable corpora in statistical and rule-based machine translation.

The ACCURAT methodology and tools will be demonstrated through application scenarios developed in the project: translation solutions for professional translators in localization services, MT for web authoring – blog writing, and for translation in narrow domains.

4. Impact: potential impact and use

The ACCURAT project addresses the widely recognized bottleneck of insufficient parallel corpora for under-resourced languages and narrow domains, i.e., the lack of sufficient linguistic data for machine translation systems. The ACCURAT will provide methods for automatic acquisition and annotation of language resources, removing gaps in language coverage and increasing quality of translation and providing methods for automated translation to make it more adaptive. The technological advances brought about by the ACCURAT project will advance the overall theory and practice of MT, corpus linguistics, information extraction and natural language processing on the whole.

ACCURAT will have a positive impact on further European integration of research and ICT industry from countries which have recently acceded to the EU (Latvia, Lithuania, Estonia, Romania, Slovenia) and candidate country Croatia.

Project has created significant interest in research community. Its first results have been published in 9 scientific papers prepared by project team members.

5. The address of the project public website

The ACCURAT website (http://www.accurat-project.eu/) was launched at the start of the project. This is the place where all information related to ACCURAT is stored and made accessible to the community. Beside classic dissemination website content, we also provide some newer means of dissemination information about the project such as video recorded lectures.