Contact: Jörg Tiedemann
Dept of Linguistics and Philology
jorg.tiedemann@lingfil.uu.se
tel +4618 4711412

1) Trainable Tokenizers and Sentence Segmenters for Let's MT!

Tokenization and the detection of sentence boundaries are important for most NLP applications. However, the complexity and the language-specificity of these problems are often neglected and generic tools are used. The goal of this project is to integrate trainable tokenizers and sentence segmenters in an existing platform for corpus preparation. Classification-based tokenizers and segmenters exist and will be used in this project (for example, OpenNLP). The task is to develop interfaces and tools to support training and testing of language specific segmentation models. A prototype exists which can be used as the starting point.

2) Integration of interactive sentence alignment in Let's MT!

Let's MT! is a collaborative on-line platform for statistical machine translation. It provides the possiblity to upload training data for building domain-specific customized translation models. Previously translated documents are the most important resource to create statistical translation models. Such documents are aligned sentence-by-sentence to create so-called "parallel corpora". Let's MT does this job automatically when documents are uploaded. So goal of this project is to integrate a module that allows to inspect conversion and alignment results and that enables simple manual corrections to erroneous alignments. A conceptual prototype for interactive sentence alignment exists. The functionality should be similar to another existing software: http://wanthalf.saga.cz/intertext (which could even be used within this project)

3) Detection of parallel documents

Statistical machine translation systems are trained on large sets of parallel (previously translated) documents. The goal of this project is to develop software that automatically detects parallel (translated) documents with large collections of documents using name/path heuristics and language footprints. This tool also needs to be integrated in an existing web-based platform for collaborative machine translation (Let's MT!)

4) From WikiSource to Parallel Corpora

Many copyright-free books are published at WikiSource.org in a variety of languages. This project aims at creating a parallel corpus out of the on-line material in this collection. Challenges in this project include the automatic discovery of parallel documents in the collection, the detection of omissions and insertions, and the conversion and alignment of translated documents. Various tools are available for extraction and alignment of texts. However, it will be necessary to implement dedicated scripts and tools for handling WikiSource data.

5) Implementing an iPhone or/and Android App for Let'sMT!

Let's MT! is a collaborative on-line platform for statistical machine translation. It provides the possiblity to upload training data for building domain-specific customized translation models and to run translation engines on-line. The task of this project is to use the public translation API to create an Android or iPhone App that can be used together with existing translation services provided by Let's MT!

6) Visualization and Annotation of Parallel Treebanks

OPUS (http://opus.lingfil.uu.se/) is a large collection of parallel corpora in various languages. Part of the data is parsed and word-aligned using existing tools. The aim of this project is to develop an interface for the visualization of parsed and word-aligned parallel corpora. This interface should also include possibilities for editing annotation and alignment. A conceptual prototype can be seen here: http://opus.lingfil.uu.se/svg/parallel.php.