Parallel Corpora and Machine Translation

Intervention au séminaire du Laboratoire CLILLAC-ARP du lundi 16 octobre 2017

Martin Volk
Universität Zürich
Institut für Computerlinguistik Laboratoire DILTEC EA2288

Abstract: Large parallel corpora are a valuable resource for linguistics, translation studies, and language technology. In the first part of this talk we present our work on collecting and annotating large parallel corpora (DE, EN, FR, IT) that cover long time spans. We discuss crowd-correction of OCR errors, automatic language identification and code-switch detection, Part-of-Speech tagging and lemmatization, and cross-language alignment. We demonstrate Multilingwis, our search tool for multi-parallel corpora.

In the second part we show how these parallel corpora can be used to build machine translation systems. We explain the basic ideas behind statistical and neural machine translation and give examples from our application of these technologies. We discuss the state-of-the-art of current neural translation systems (Google Translate, DeepL), their merits and short-comings.