TransBank: A Meta-Corpus for Translation Research

Abstract

Corpora, or large text collections, have become a major staple of linguistic research in recent years, ranging from grammar topics, to lexicology, socio-linguistics, and applied linguistics. Corpus-based approaches have had an impact on the field of translation as well, however, mainly from a computer science perspective and with regard to the development of translation tools, such as translation memory, terminology management and machine translation systems. A rather large piece of the puzzle that still appears to be missing in this picture are data-driven methodologies that help make explicit real-world translation phenomena and make possible sound theoretical models capable of explaining them.

Against this background, our project aims to provide a large, open and expandable collection of translated texts and their original texts, which are to be stored in a bank – hence the name TransBank. In contrast to many other bi- or multilingual collections, the texts and translations are aligned, i.e., paired at the sentence level. The main innovative feature, however, is the ability to compile and download parallel sub-corpora on demand, tailored to the requirements regarding specific questions of translation research. Possible studies may include, for example, question such as if and how language change is caused by translation, what the impact of translation technology on the style of written text is, what differences there are between the same text-type in different languages, what stylistic differences there are between translations into different languages, or cognitive research interests in connection with the differences between texts produced by trainees and experienced practitioners. The sub-corpora, i.e., smaller, partial text collections for specific purposes, can be generated with the help of a search interface that allows for very complex queries and is at the same time very user-friendly. This is made possible by using faceted search technology, i.e., clickable labels that can be combined, excluded and/or filtered.

The key to such flexible corpora compilation is a precise set of metadata labels that help to capture the relation between translated texts and their originals, including the circumstances under which the translations were produced. This metadata set is the basis of a universal repository of empirical translation resources that combines the advantages of pre-existing, separate data collections.

In summary, what we are aiming to provide is re-usable, open, empirical data for translation research.