Metadata First! Designing a Label Set for a Translation Meta-Corpus

In the translation industry, vast amounts of textual data are being produced, be it by language service providers, professional freelance translators, amateur translators in the crowd, trainee translators, or AI machine translation systems. Driven by the irreversible trend towards translation automation, several repositories for the large-scale collection of translation data have been created. These repositories are mainly being used as a resource of training data for sophisticated translation tools that help boost productivity in the language industry. They are, however, of limited value for data-driven research into the complex nature of translation phenomena. The reason for this is that due to the primacy of size in the era of big data, less attention is paid to metadata about the texts in the repositories. After all, the object of study in Translation Studies is not limited to translation products but extends to translation processes and the linguistic, cognitive, socio-cultural, socio-economic and technological factors that influence translation. Neither of these factors can be investigated in a meaningful manner without metadata that describe how, when and under what circumstances a given of translation data has been produced.

If industry and academia in translation are to fully exploit the benefits of cutting-edge data-driven techniques, the development of a precise set of translation-related metadata labels is a prerequisite. The development of such a label set that is equally useful for all stakeholders in translation faces the challenge of the need to accommodate in a generally valid manner a large variety of different forms of translation-related text production processes (e.g. technical translation vs. literary translation vs. movie subtitling vs. software localization, translation of various text types and genres etc.), each of which exhibits its own characteristic features. An overly exhaustive description of the parameters that govern the production of translations and their respective originals may be tempting from an epistemological perspective but lacks practicality and usability in research and industry settings alike.

In this contribution, we describe the principles that apply to the design of the label set used to capture the metadata of texts included in a novel repository of real-world translation data to be gathered from various sources. Given its broad scope of any type of translation, the construction of the repository walks a tightrope between general validity and usability. The repository's function is to be a meta-corpus that allows translation researchers to compile and download just the specific sub-corpus that they need, which can be defined by the metadata provided. Thus the meta-corpus aims to be a key element of infrastructure for modern, empiric, data-driven translation research.