# Feature-rich NMT and SMT post-edited corpora for productivity and evaluation tasks with a subset of MQM-annotated data

**Kim Harris**
**Lucia Specia and Aljoscha Burchardt**

**Abstract**

This presentation will discuss the creation and practical use of a large data set created through an unprecedented large-scale collaboration between MT R&D and translation experts. It contains post-edited and annotated industry data for four morphologically rich language pairs ( EN-DDe, EN-CS, DE-EN, EN-LV ). A subset of "almost perfect" sentences also contains MQM error annotations for further detailed analysis and profiling for recurring error patterns. The post edits were performed by professional translators and the data is freely available for further use. The data used for post-editing comprised 20,000 to 45,000 sentences of industry data (IT, life sciences) depending on the language pair. The post-editing of all four language pairs was performed using PET (Aziz, W. et al ). Several crucial and novel data points were taken during the post-editing: time logging, keystroke logging quality evaluation of the post-editing effort by the translator upon completion of the post-editing. The recording of this information during the post-editing phase allows for specific features and novel combinations of features to be used for a variety of research- and user-oriented purposes, including establishing the actual post-editing effort by translators based on time and keystrokes and comparing these results to the perceived level of quality of the post-edited sentence, establishing correlations between certain characteristics such as sentence length and post-edit time, or post-edit time and human quality evaluation. The datasets also measure post-editing productivity and are used to detect error patterns in the MT output. This would allow users of MT to adequately assess a) the use of MT in general, b) the actual productivity gains achieved in two different systems or across languages, domains and other data subsets such as long sentences or sentences containing certain grammatical constructs or terminology. For two language pairs identical sets of source sentences comprising 30,000 sentences respectively were post-edited for NMT and SMT output, allowing for a variety of innovative comparisons to be done on the results of the two given the unique data points that were collected during post-editing In addition, the creation of MQM-annotated subsets of these post-edits for typical industry domains provide

information about error patterns and support feature-oriented quality estimation and evaluation currently unknown to MT quality evaluation and estimation and can be used to improve the MT output