How easy is Easy German?
From word complexity to sentence simplicity

Umesh Patil, Jesus Calvillo, Anne-Kathrin Schumann, Felix Dittrich, Martin Rosenbusch, Ian List, Dennis Engel & Tobias Wittig
t2k GmbH
umesh.patil@text2knowledge.de

We propose an automatic metric for quality estimation (QE) of the simplicity of text and show that: (i) it correlates well with human judgments of simplicity, (ii) it can be used to highlight the variability in the simplicity of different Easy German corpora, and (iii) it can be used to clean and improve the quality of training corpora. We plan to make the metric available for academic purposes.

SIMPLICITY METRIC. We trained an *XGBoost* classifier for a Complex Word Identification task using a dataset of 8K German words and phrases annotated for complexity. The classifier was trained using word level features such as word length, frequency, part-of-speech, among others. It can also output a word complexity score. We add the complexity scores of all the words in a sentence to define the simplicity score of a sentence. We evaluated the simplicity metric using the *TextComplexityDE* dataset containing 1K sentences rated by L2 learners. The correlation analysis between the predicted simplicity scores and human ratings showed a strong correlation (Pearsons' r of 0.78 for complexity, 0.67 for understandability and 0.72 for lexical difficulty; all p-values <0.0001).

APPLICATIONS. The correlation analysis demonstrates that the simplicity metric can be straightforwardly used for QE of text simplification (see Fig.). Further, we propose that the metric can also be used for cleaning corpora by defining simplicity levels for sentences. Quality training data is crucial for accurate simplification by language models. We noted complexity variation in our Easy German dataset and conducted a preliminary manual annotation, classifying texts into four complexity levels. Applying the simplicity metric confirmed these distinctions, validating its sensitivity to complexity and helping us refine our classifications, thus enhancing dataset quality.

FUTURE WORK. We are currently exploring ways to extend the simplicity metric by incorporating psycholinguistic features, such as incremental processing costs based on *surprisal* and working memory principles.

Figure. Example of word- and sentence-level simplicity scores for the input and output of the *t2k* model. Word-level complexity ranges from blue (easy) to red (complex), with sentence-level simplicity scores displayed at the top.