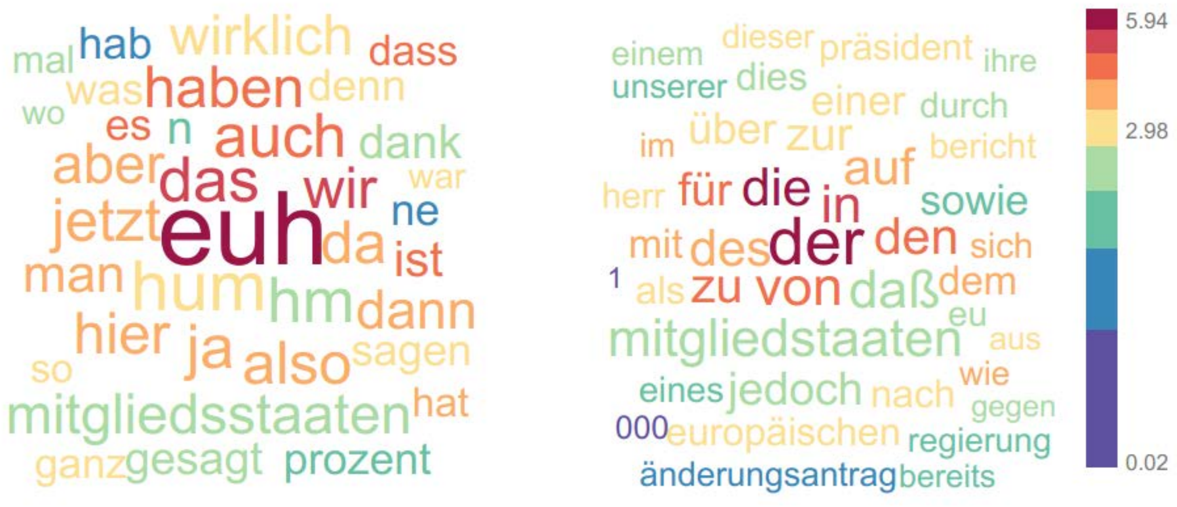
Detecting linguistic variation in translated vs. interpreted texts using relative entropy

Keywords — Corpus-based interpreting studies, translationese, relative entropy, linguistic variation

We propose a corpus-based, exploratory approach to detect typical linguistic features of interpreting vs. translation. The data sets we use are the Europarl-UdS corpus (Anonymous 2018), which contains originals and translations for the languages English, German and Spanish, and selected material from existing interpreting/combined interpreting-translation corpora (EPIC: Sandrelli and Bendazzoli 2005; TIC: Kajzer-Wietrzny 2012; EPICG: Defrancq 2015), complemented with German.

We build probabilistic language models (n-gram models) of source (EN) and target languages (ES, DE) for both interpreting and translation. To explore linguistic variation across translation and interpreting, we calculate the relative entropy (by Kullback-Leibler Divergence, KLD) between the translation and interpreting models, estimating the amount of additional bits (information) needed to model interpreting by translation (and vice versa). This gives us an indication not only of how different translation and interpreting outputs are overall but also of the linguistic features (here: words) that contribute most to the difference (Fankhauser et al., 2014). On the basis of a word-cloud visualization, we then explore the words that are the strongest signals of variation by relative frequency (shown by colour: high relative frequency red - low relative frequency blue) and highest distinctivity (shown by size). For an example see Figure 1.



a. **Interpreting** (b) **Translation**

Figure 1: Variation in translation mode in German texts with English as source language

We can observe that overall, interpreting exhibits more highly distinctive items as well as high frequency items than translation. Closer inspection reveals that well-known features of spoken discourse appear as typical for German interpreted texts, such as hesitation markers (*euh, hum*), particles, discourse markers and intensifiers (*so, mal, ja, ganz, wirklich*), deictics (*jetzt, hier*), reduced forms (*hab, ne, n*) and highly frequent general verbs (*haben, sein*). Some features seem to be a result of the interpretation process, some are related to this specific translation direction or are due to the fact that the original is also a spoken text. Written translations, instead, are characterised by a more nominal style indicated by various determiners and pronouns (*dem, dieser, ihre, unsere* etc.) and prepositions *(in, mit, für).* Additionally, more content words are shown to be distinctive for translations (e.g. *Änderungsantrag, Regierung*) than in interpreting and the mean word length is higher.

In our ongoing work, we combine the proposed method for detecting typical features with analysis of interacting constraints, e.g. source language shining-through, and explore additional information-theoretic methods to capture additional variables, e.g. translation difficulty measured by entropy (Anonymous, to appear).

Word count: 406

References:

Defrancq, Bart (2015). Corpus-based research into the presumed effects of short EVS. In: *Interpreting* 17.1: 26–45.

Fankhauser, Peter, Jörg Knappen and Elke Teich (2014). Exploring and Visualizing Variation in Language Resources. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC).*

Kajzer-Wietrzny, Marta (2012). *Interpreting universals and interpreting style.* PhD thesis. Adam Mickiewicz University, Poznań, Poland.

Sandrelli, Annalisa and Claudio Bendazzoli (2005). Lexical Patterns in Simultaneous Interpreting: A Preliminary Investigation of EPIC (European Parliament Interpreting Corpus). *Proceedings from the Corpus Linguistics Conference.*