Analysis of simplification in coreference from two perspectives

In this paper, we analyse coreference features of the German language, focusing on the phenomenon of *simplification*, i.e. the tendency to use words and constructions that are assumed to be easier perceived, understood, or produced. Simplification is one of the means used by language users in order to optimise communication effectively. We are interested in how simplification is reflected in coreference in two different language products exposed to the phenomena of simplification: simultaneous interpreting and Easy German. As seen from example (1), the English source contains the chain *the practice of sandblasting – which – jeans sandblasted* with mentions filled with a relative pronoun and a full lexical phrase. At the same time, the interpreting into German contains a demonstrative pronoun (*das*) and an adverb (*so*) instead. From the lexical point of view, the means of referring are simpler in the interpreted output. In contrast, the coreference chain in the Easy German example in (2) contains no pro-forms, but lexical repetitions as a simplification strategy. In addition, the anaphors are highlighted by being positioned sentence-initially.

(1) **English original**: *In particular, I want to draw attention to the practice of sandblasting of jeans which happens more in Bangladesh than anywhere else in the world. Up to one hundred million pairs of jeans sandblasted a year being export from Bangladesh.* **German interpreting**: *Aber was dort in Bangladesch passiert, ist weiter eine Bedrohung für die Gesundheit der Arbeitnehmer, insbesondere die Sandstrahlmethode für Jeans. Das wird in Bangladesch vor allen Dingen durchgeführt. Einhundert Millionen Jeans werden so hergestellt und exportiert pro Jahr.*

(2) **Easy German**: *In Hamburg sind am Wochen·ende 2 große Veranstaltungen. Diese 2 großen Veranstaltungen sind: • Ein Musik·fest. • Und eine Sport·veranstaltung. Die 2 großen Veranstaltungen sind in St. Pauli. […] Und die 2 großen Veranstaltungen sind […] Zu diesen 2 großen Veranstaltungen kommen sehr viele Menschen.*

While both language products are known to be simplified, the driving forces of the optimisation process differ. Easy German is simplified to be better perceived and understood by the target audience, i.e. the receiver side. At the same time, simultaneous interpreting is simplified due to the production constraints on the producer side, i.e. the interpreter who optimises the output to reduce their own cognitive load.

We are interested in the differences and similarities of the simplified language products that are the results of these two varying optimisation reasons. For instance, shorter coreference chains, mentions used as subjects and fewer expression variants per chain indicate simplification, as well as the length of the expression measured in words: the shorter the mention expressions, the simpler the text. The formulated features are based on the studies in the area of automatic coreference resolution for German (see e.g. [5]) as well as accessibility analysis for German (see e.g. [2]).

We use two different sets of data. For the analysis of simultaneous interpreting, we use a sample of 137 texts of German interpreting from English extracted from EPIC-UdS ([3]), a multilingual parallel and comparable corpus of simultaneous interpreting of political speeches. For the analysis of Easy German, we use a sample of about 4,700 texts from DE-Lite v1 ([1]).To analyse coreference, we annotated the data with the state-of-the-art coreference resolver CorPipe ([4], [6]). In our presentation, we will compare the frequency distributions of the annotated coreference features across the texts in the two data sets and discuss them in relation to the different simplification strategies.

References

1. Jablotschkin, Sarah, Teich, Elke, & Heike Zinsmeister (2024). DE-Lite - a New Corpus of Easy German: Compilation, Exploration, Analysis. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 106–117. Association for Computational Linguistics, St. Julian's, Malta.

2. Kunz, Kerstin (2010). *Variation in English and German nominal Coreference. A study of political essays*. Peter Lang, Frankfurt am Main, Germany.

3. Przybyl, Heike, Lapshinova-Koltunski, Ekaterina, Menzel, Katrin, Fischer, Stefan, & Teich, Elke (2022). EPIC UdS - Creation and Applications of a Simultaneous Interpreting Corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1193–1200. European Language Resources Association, Marseille, France.

4. Straka, Milan (2023). ÚFAL CorPipe at CRAC 2023: Larger Context Improves Multilingual Coreference Resolution. In *Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution*, pages 41–51. Association for Computational Linguistics, Singapore.

5. Strube, Michael, & Hahn, Udo (1999). Functional Centering: Grounding Referential Coherence in Information Structure. *Computational Linguistics* 25(3). 309–344.

6. Žabokrtský, Zdeněk, & Ogrodniczuk, Maciej (2023). *Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution*. Association for Computational Linguistics, Singapore.