Scheffler, T. (2020). Explicitness and implicitness of discourse relations across social media.
Paper presented at the Workshop "Explicit and implicit coherence relations: Different, but how exactly?",
Humboldt-Universität zu Berlin, Berlin, Germany. 17 - 18 January.

Explicitness and implicitness of discourse relations across social media

Research question. It is known that there are differences in the realization of discourse relations across text types, in particular between spoken conversations and written text (see e.g., Crible/Cuenca 2017). However, it is not clear which differences are due to topic or genre effects, to conversational/monologic use of language, to the spoken or written mode, or to individual stylistic differences (see Verhoeven/Daelemans 2018). In this presentation we attempt to tease these aspects apart to some extent by investigating discourse connective use in two social media, microblogs on Twitter and long-form blogs. We particularly address the following research questions: (i) Do blogs exhibit more explicit discourse relations than tweets? (ii) Which types of connectives and relations vary across the two media? (iii) Are individual author choices relevant for explicitation or implicitation of discourse relations?

Data. We study these questions wrt. a corpus of German blogs and tweets by 71 authors (~2.5mio tokens). For each author, 5-10 recent blog posts as well as the (up to 3248) most recent tweets have been automatically collected in February, 2017. Authors are linked between tweets and blogs. The complete overlap in authors allows us to compare the subcorpora while avoiding personal author style as a confounding factor. In addition, we are thus able to address research question (iii) by looking at a specific author's adaptation to the tweet or blog media. We selected authors from a list of German-speaking "parenting bloggers", further ensuring relative topic cohesion across the two subcorpora (topics are mainly personal life and parenting advice).

Analysis. We extract all German connectives listed in the German discourse marker lexicon DiMLex (Stede 2002, Scheffler/Stede 2016). The lexicon contains 274 connectives, as well as information on their spelling variants and the discourse relations they express. The connectives are automatically identified and disambiguated (Bourgonje/Stede 2018). We then analyse the connective and relation frequencies statistically along the axes of connective, author, and medium.

Results. Briefly, the general results confirm the expectations, and open up avenues for detailed analysis. (i) Overall, all explicit discourse connectives are relatively more frequent in the blogs than in the tweets (see Fig. 1). Possible explanations are that discourse relations are generally less frequent in the tweets, or that the relations are more often left implicit. Figure 2 shows frequent connectives for the cause-reason discourse relation – in particular "denn" and "weil" are much more frequent in blogs than on Twitter, indicating that causal relations are often left implicit in tweets. (ii) The difference in frequency of explicit discourse relations is significant only for certain types of connectives, notably formal and phrasal connectives, such as "in Anbetracht dessen, dass" (seeing as), "für den Fall, dass" (in case), "es sei denn" (unless). In addition, "und" (and) is used almost twice as often in blogs as in tweets, which may be due to longer and more complex sentences.
(iii) Even though all authors show the same tendency overall (to use connectives less frequently in tweets), this is not true for all relations/connectives for each author. In the presentation, we will present some individual case studies which analyse particular authors' blogs/tweets in detail.

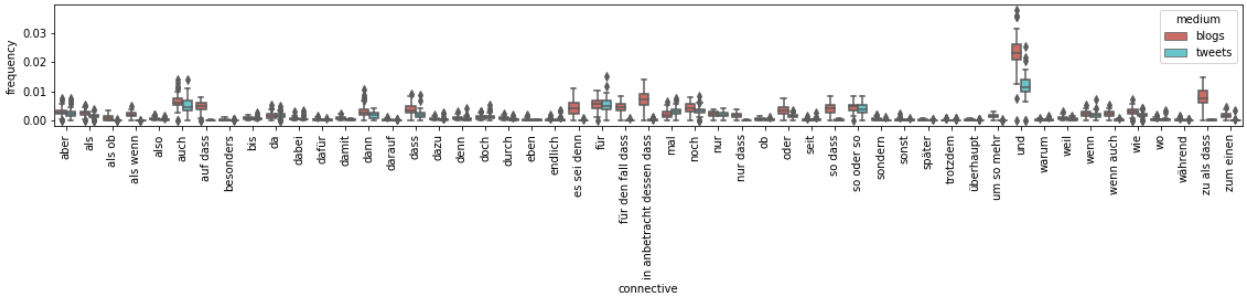

Fig. 1: Relative frequency (per token) of the 31 most frequent discourse connectives in the two media (blogs/tweets). Box plots show the frequency distribution of each connective over the 71 authors in the study.

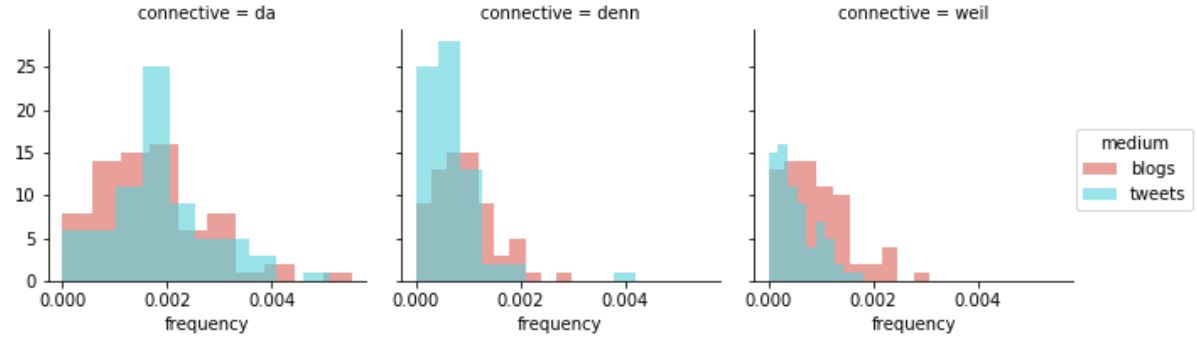

Fig. 2: Relative (per-token) frequency of causal connectives “da”, “denn”, and “weil” in German blogs and tweets, as a histogram over author-connective frequencies.

Selected References:

Peter Bourgonje and Manfred Stede. Identifying explicit discourse connectives in German. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, 327–331. Melbourne, Australia, 2018. Association for Computational Linguistics.

Ludivine Crible and Maria-Josep Cuenca. Discourse Markers in Speech: Characteristics and Challenges for Corpus Annotation. *Dialogue and Discourse* 8(2). 2017.

Manfred Stede. "DiMLex: A Lexical Approach to Discourse Markers" In: A. Lenci, V. Di Tomaso (eds.): *Exploring the Lexicon - Theory and Computation*. Alessandria (Italy): Edizioni dell'Orso, 2002.

Tatjana Scheffler and Manfred Stede. "Adding Semantic Relations to a Large-Coverage Connective Lexicon of German." In *Proceedings of LREC*. Portorož, Slovenia. 2016.

Ben Verhoeven and Walter Daelemans. "Discourse lexicon induction for multiple languages and its use for gender profiling." *Digital Scholarship in the Humanities* 34, no. 1. 2018.