Connective ambiguity and compensation by discourse signals in speech and writing

Background. Discourse, either written or spoken, is built upon coherence relations, which specify the nature of the link that connects clauses and other discourse segments. These coherence relations can be additive, contrastive, causal, conditional, etc. While their number and labels vary across frameworks, they have repeatedly been studied in connection with the category of discourse connectives (e.g. Knott & Dale, 1994). Connectives are the prototypical signals for coherence relations. Despite their famous ambiguity, no one would question that *whereas* in (1) expresses a contrastive relation.

(1) You cannot overdose with marijuana, whereas you can overdose with alcohol.

The subordinating conjunction in this example would typically be considered as the signal for the coherence relation of contrast between the two clauses, making the relation *explicit* in this context. In many approaches to coherence relations, the same example without the connective would be classified as an *implicit* relation, as in (2).

(2) You cannot overdose with marijuana, you can overdose with alcohol.

The contrastive interpretation is maintained despite the absence of *whereas*. The apparent dichotomy between implicit and explicit relations has been refined in recent works (Das & Taboada, 2018; Péry-Woodley et al., 2018), which also consider additional signals of coherence relations beyond connectives. These signals include, for instance, negative polarity (Webber, 2013) or focus markers (Carlson, 2014) in the context of contrastive relations. In examples (1-2) above, the different polarity between the two segments (negative-positive) and the syntactic parallelism between them act as signals for the contrastive relation, either reinforcing the connective in (1) or compensating for its absence in (2).

Despite the growing interest for these other discourse signals, they have not yet been systematically investigated in relation with the type of connective with which they co-occur, especially in terms of connective ambiguity. Not all connectives are equal in how strongly they signal a discourse relation (Asr & Demberg, 2012), depending on the number and frequency of other relations which they can express: a monosemous connective like *whereas* is stronger in expressing contrast than the additive conjunction *and*. Yet, in (3), a relation of contrast can still be easily interpreted.

(3) You cannot overdose with marijuana, and you can overdose with alcohol.

Here, the relation relies more on the negation and the parallelism than on the connective itself. Still, such less standard co-occurrences are not uncommon and deserve further investigation.

Research question. This study explores the combination of connectives with other discourse signals as a factor of connective ambiguity and genre variation. Ambiguity in discourse is pervasive, yet cognitive theories of production and processing suggest that it tends to be compensated in context. The paper adopts an innovative statistical approach to discourse signalling in order to uncover configurations and predictive signals which function as reinforcing cues for different relations and connective types.

Hypotheses. One could hypothesize that the "weaker" the connective for a given relation, the more frequently it will be compensated by other signals, following psycholinguistic models of language production such as the Uniform Information Density hypothesis (Levy & Jaeger 2007) or Rational Speech Act theory (Frank & Goodman 2012). This compensation is further expected to vary across genres, with more signals in formal and planned texts, under the influence of planning pressure and/or recipient design (Spooren, 1997). Thirdly, preferences for signalling may also vary with the cognitive complexity of the coherence relation: negative relations are expected to require strong signals (Hoek et al. 2017).

Method. These hypotheses are tested on a corpus of English conversations, chat and essays (Loyola CMC corpus, Goldstein-Stewart et al., 2008), where connectives and other discourse signals have been manually annotated. More specifically, all connectives (12,710 tokens) have been identified and their meaning-in-context disambiguated, using Crible & Degand's (in press) sense taxonomy. The distribution of connective types and functions across genres is discussed. Moreover, a sample of 2,000

of these connectives have further been analyzed in terms of their contextual features, including sentence mood and polarity, verb tense, unit type, syntactic construction, semantic relation, and other semantic and co-occurring elements in the context of the connective. The presence and nature of such signals is associated through statistical multivariate analysis with the three variables under scrutiny, namely connective ambiguity (measured as a score of marking strength, Asr & Demberg 2012), relation complexity and genre.

Results. Statistical modelling of the interaction between connective ambiguity, presence of compensating signals, and genre reveals that weak connectives are more often compensated than intermediate and strong connectives, and that this compensation is not affected by genre variation. Formality (and co-presence of the addressee) does explain the distribution of more or less ambiguous connectives in the corpus, but their combination (or not) with other discourse signals seems more dictated by the strength of the connectives and by the relation it expresses, with specification and contrast particularly prone to reinforced signalling, as opposed to consequence relations.

Conclusion. This study is a direct continuation of Das and Taboada's (2018) recent analysis of discourse signals beyond connectives. It is complementary to their endeavor and goes further by examining the interaction between discourse signals on the one hand, and connective ambiguity on the other, including genre as a potential factor of variation and resorting to robust statistical models. The results refine the divide between explicit and implicit relations by introducing weak, intermediate and strong connectives in the continuum, investigated in their interaction with other discourse signals.

References

Asr, F., & Demberg, V. (2012). Measuring the strength of linguistic cues for discousre relations. In *Proceedings of the COLING Workshop on Advances in Discourse Analysis and its Computational Aspects (ADACA)*, 33–42, Mumbai, India.

Carlson, K. (2014). Predicting contrasts in sentences with and without focus marking. *Lingua* 150, 78–91.

Crible, L. & Degand, L. (in press). Domains and functions: a two-dimensional account of discourse markers. *Discours*.

Das, D., & Taboada, M. (2018). Signalling of coherence relations in discourse, beyond discourse markers. *Discourse Processes* 55(8), 743–770.

Frank, A., & Goodman, N. (2012). Predicting pragmatic reasoning in language games. *Science* 336(6084), 998.

Goldstein-Stewart, J., Goodwin, K. A., Sabin, R. E., & Winder, R. K. (2008). Creating and using a correlated corpora to glean communicative commonalities. In *LREC2008 Proceedings*, Marrakech, Morocco.

Hoek, J., Zufferey, S., Evers-Vermeul, J., & Sanders, T. J. M. (2017). Cognitive complexity and the linguistic marking of coherence relations: A parallel corpus study. *Journal of Pragmatics* 121, 113–131.

Knott, A., & Dale, R. (1994). Using linguistic phenomena to motivate a set of coherence relations. *Discourse Processes* 18, 35–62.

Levy, R., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In B. Schölkopf, J. Platt & T. Hoffman (Eds.), *Advances in neural information processing systems (NIPS),* vol. 19, pp. 849–856. Cambridge, MA: MIT Press.

Péry-Woodley, M.-P., Ho-Dac, L.-M., Rebeyrolle, J., Tanguy, L., & Fabre, C. (2017). A corpus-driven approach to discourse organisation: from cues to complex markers. *Dialogue & Discourse* 8(1), 66–105.

Spooren, W. (1997). The processing of underspecified coherence relations. *Discourse Processes* 24, 149–168.

Webber, B. (2013). What excludes an alternative in coherence relations? *Proceedings of the 10th International Workshop on Computational Semantics (IWCS2013)*, 276–287.