Connective or no connective?

The influence of default expectations and the presence of segment-internal cues on the linguistic marking of coherence relations

Jet Hoek,[1] Sandrine Zufferey,[2] Jacqueline Evers-Vermeul,[1] & Ted J.M. Sanders[1]
[1] Utrecht University, Utrecht Institute of Linguistics OTS, [2]University of Bern

Coherence relations can be made linguistically explicit by means of connectives (e.g., *but, because*) or cue phrases (e.g., *on the other hand, which is why*), but can also be left implicit and conveyed through the simple juxtaposition of two clauses or sentences. In the absence of a connective, readers or listeners have to infer the relation between the discourse segments themselves using the linguistic context and their world knowledge (Kintsch, 1998; Zwaan & Radvansky, 1998). However, it seems that not all relations are equally easy to reconstruct when they are implicit. In (1), a fragment taken from the Europarl corpus (Koehn, 2005), for instance, the relation between the first and the second sentence has not been explicitly marked by a connective, but it can still be determined that the potential hate-inducing qualities of the remarks are the reason for the speaker to find them unworthy of being uttered in the European Parliament; an appropriate connective would have been *because*. The two clauses that make up the second sentence, on the other hand, are connected by *if*. Leaving out this connective would make the relation hard to reconstruct; although the remarks would still be understood to cause agitation with the public; them getting into the media would most likely be interpreted as a given, rather than as a hypothetical event.

(1) Those sort of remarks are unworthy of this Parliament. If [they get into the media] [it just stirs up hate.]

The intuition that some types of relations can be more easily left implicit than others is confirmed by analyses of discourse-annotated corpora. Studies by for instance Asr and Demberg (2012) on the Penn Discourse Treebank (PDTB Research Group, 2008) and Taboada (2006) on a corpus annotated using Rhetorical Structure Theory (Mann & Thompson, 1988) show comparable patterns in the marking of coherence relations. Causal relations and relations of general addition, for instance, are often expressed implicitly. Conditional relations and relations involving concession, on the other hand, tend to be explicitly marked. While the existence of asymmetries in the marking of coherence relations has been clearly established, the exact mechanisms that cause these asymmetries are not yet fully understood. In this presentation, we use parallel corpora to explore two mechanisms that may influence whether or not a coherence relation is marked by a connective: default expectations and the presence of segment-internal cues that can function as signals for coherence relations.

Expectations and the presence of connectives

Asr and Demberg (2012) propose that the linguistic marking of a coherence relation is strongly influenced by a relation's expectedness, with expected relations being more often left implicit. This assumption finds its roots in the Uniform Information Density (UID) hypothesis (Frank & Jaeger, 2008; Levy & Jaeger, 2007), which proposes that speakers "structure their utterances so as to avoid peaks or troughs in information density" (Levy & Jaeger, 2007:1). A linguistic element that marks something that was already expected by the reader hardly adds any information to the discourse, and therefore constitutes a trough in information density. Conversely, leaving implicit something that was not already projected causes an overload of information to be extracted from the linguistic elements that are present, thus constituting a peak in information density. The idea that expected relations can be left unmarked can also be thought of in terms of effort versus effect, key notions from Relevance Theory (Sperber &

Wilson, 1985; Wilson & Sperber, 2005). If an unexpected relation is not marked, its inference requires too much effort for the resulting cognitive effect. As a result, an easier, more expected coherence relation will be inferred. Explicitly marking unexpected relations therefore ensures that the right relation is established. For example, not explicitly signaling a conditional relation, such as the one in (1), makes it hard or even impossible to recover that relation, and the fragment will most likely receive a different, non-hypothetical interpretation.

If we assume that the linguistic marking of coherence relations is to a large extent governed by expectedness, it is key to determine which kinds of relations are expected to occur in a discourse. In this presentation, we look into default expectations in Study 1, and expectations generated on the basis of linguistic elements inside the discourse segments in Study 2.

Study 1: Default expectations

In our first study we investigate the influence of default expectations on the presence of connectives. We propose to determine a relation's expectedness using the notion of cognitive complexity. Traxler et al. (1997a) find that simple relations are processed faster than complex relations. They argue that readers construct the simplest possible coherence relation and adapt their representation of the discourse if this relation is not consistent with the context (see also Traxler et al., 1997b). If cognitively simple relations are expected, they should occur implicitly more often than cognitively more complex relations.

We test our hypothesis by means of a parallel corpus study, in which we analyze the translations of explicit English coherence relations from the Europarl Direct corpus into four target languages: Dutch, German, French, and Spanish. We find that cognitive complexity indeed seems to influence the linguistic marking of coherence relations, and that this does not vary between the languages in our corpus.

Study 2: The presence of segment-internal elements

Expectation-based accounts of the marking of coherence relations, as briefly discussed above, predict that speakers use a connective when it contributes essential information to the discourse. If the connective is barely informative or even entirely redundant, speakers will be more inclined to leave it out, in which case the relation will be implicit, or use a more general connective, in which case the relation will be underspecified. If another element within a discourse segment already signals or partly signals how that segment should be related to another segment from the discourse, this would eliminate or reduce the amount of information a connective would contribute. Other linguistic elements that convey information or raise expectations about the type of coherence relation that should be constructed are thus expected to influence the marking of coherence relations by connectives.

There are several segment-internal features that have been linked to particular types of coherence relations. These segment-specific elements include a wide range of linguistic categories, such as complex phrases, lexical items, modal markers, and verbal inflection. It seems, however, that not all linguistic elements that have been associated with a specific type of coherence relation signal the relation in the same way, and there appear to be differences in the way in which the presence of a specific linguistic element in the segments of a relation can impact the marking of that relation by means of a connective. In this study, we argue that there are three distinct ways in which segment-internal elements interact with the connective that marks a coherence relation: *division of labor*, *agreement*, and *general collocation*. We illustrate the existence of this three-way categorization using parallel corpus data. For each type of interaction, the reason why the segment-internal element functions as a cue for a specific type of relation is slightly different. In addition, the likelihood of the relation being marked by a connective in the presence of a segment-internal cue varies between the three interaction types.

References

Asr, Fatimeh T. & Demberg, Vera, 2012. Implicitness of discourse relations. *Proceedings of COLING 2012*. Mumbai, India, 2669-2684.

Frank, Austin F. & Jaeger, T. Florian, 2008. Speaking rationally: Uniform information density as an optimal strategy for language production. *Proceedings of the 28th meeting of the Cognitive Science Society*. Washington DC, USA, 939-944.

Kintsch, Walter, 1998. *Comprehension: A Paradigm for Cognition*. Cambridge: Cambridge University Press.

Koehn, Phillip, 2005. Europarl: A parallel corpus for statistical machine translation. Tenth Machine Translation Summit (MT Summit X), Phuket, Thailand. http://homepages. inf.ed.ac.uk/ pkoehn/publications/europarl-mtsummit05.pdf.

Levy, Roger & Jaeger, T. Florian, 2007. Speakers optimize information density through syntactic reduction. In: Schlökopf, B., Platt, J., & Hoffman, T. (Eds.), *Advances in Neural Information Processing Systems (NIPS)* (Vol. 19). Cambridge, MA: MIT Press, 849-856.

Mann, William C. & Thompson, Sandra A., 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text* 8(3), 243-281.

PDTB Research Group, 2008. The Penn Discourse TreeBank 2.0. *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*. European Language Resources Association (ELRA), 2961-2968.

Sperber, Dan & Wilson, Deirdre, 1985. Loose talk. *Proceedings of the Aristotelian Society LXXXVI*, 540-549.

Taboada, Maite, 2006. Discourse markers as signals (or not) of rhetorical relations. *Journal of Pragmatics* 38(4), 567-592.

Traxler, Matthew, J., Bybee, Michael D., & Pickering, Martin J., 1997a. Influence of connectives on language comprehension: Eye-tracking evidence for incremental interpretation. *The Quarterly Journal of Experimental Psychology: Section A 50* (3), 481-497.

Traxler, Matthew J., Sanford, Anthony J., Aked, Joy P., & Moxey, Linda M., 1997b. Processing causal and diagnostic statements in discourse. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 23(1), 88-101.

Wilson, Deirde & Sperber, Dan, 2005. Relevance Theory. In: Horn, L. & Ward, G. (Eds.), *The Handbook of Pragmatics*. New York: Wiley, pp. 607-632.

Zwaan, Rolf A. & Radvansky, Gabriel, 1998. Situation models in language comprehension and memory. *Psychological Bulletin* 123, 162-185.