Coherence relations in the Penn Discourse Tree Bank

The issue The primary research objective was to determine whether the distribution of relations in the PDTB reflects results from cognition and psycholinguistic studies, and whether this can shed further light on the discourse relations as a cognitive entity. Recent research has shown that coherence relations do not exist solely at a text level, but are also a key contributor to cognitive representations of text and to the processing of text (Sanders and Noordman, 2000; Sanders, Land and Mulder, 2007; Sanders and Spooren, 2009). It has been shown that the effect of coherence on text comprehension depends on a variety of factors: type of relation (Sanders and Noordman, 2000), whether the relation is explicit or implicit (Degand and Sanders, 2002; Sanders and Noordman, 2000), and genre (Taboada 2006). The data demonstrate that results of some processing studies are fully supported (causal relations as a cognitive entity, for example), however, for other relations classified in the PDTB, results are far less clear-cut.

Methodology and Data This paper shows the results of a corpus study of the Penn Discourse Treebank (PDTB). From the PDTB we extracted all discourse relations, their semantic classes, frequency of occurrence, and whether the connective was implicit or explicit (explicit (E) refers to relations signaled with a discourse connective, and implicit relations (I) are inferred, but not overtly signaled). Explicit relations outnumbered implicit by 18420 (E) to 16029 (I). Using Prasad et al.’s (2007) hierarchy of relations, results were analyzed according to semantic class of relation: Comparison, Contingency, Expansion and Temporal. Results show that comparison and temporal relations were significantly more likely to be explicit (7894 and 4254 respectively), whereas contingency and expansion relations were significantly more likely to be implicit (7422 and 14879 respectively). However, it is at the second semantic class level that results have implications for theoretical categories of relation. A detailed analysis of Contingency class, which comprises causal relations (contingency.cause) and conditionals (contingency, condition) shows that they pattern very differently. Causal relations are most frequently implicit (4111 E :1816 I), whereas conditionals are most frequently explicit and predominantly using the connective if. Temporal relations occurred overwhelmingly with an explicit connective (3430 E: 824 I). For expansion relations it was found that list, restatement and instantiation relations (representing the connectives and, for example, and specifically) were overwhelmingly implicit.

Discussion Firstly, the tendency of conditionals to be explicit was not surprising, however, the almost 100% explicit marking was surprising. The preference of causal relations to be implicit in the data here supports the literature that we expect causality (Asher and Lascarides 1998; Sanders and Spooren 2009; Levinson 2000). Therefore, there is no need for this link to be explicit, as the reader/hearer is already expecting this relationship to exist. For temporal relations, Sanders and Spooren (2009) also assert that relations are processed as temporally successive unless otherwise indicated. This would suggest that at least the temporal.succession relations should not require marking. However, here we show that successive relations are overwhelmingly explicit (1069 E:151 I). Similarly, the preference of temporal.precedence for explicit marking (though at a significantly lower ratio, 948:499) draws into question succession as the default temporal relation. Results for expansion reflects research that concludes that, as these relations do not have a nucleus-satellite relationship (Matthiesen and Thompson (1988)), and are less informative than other relations (especially causal) (Sanders and Spooren 2009), there is less need for a marker that would be less informative. However, interestingly, conjunction and disjunction prefer explicit connectives (and, instead, or and unless).

Conclusion This corpus study found that the distribution of relations varies highly according to the second level of relation. While it supports the findings of causality as a cognitive category requiring little additional, i.e. explicit, marking, findings here for temporal relations contradict assertions in the literature regarding the default processing order. Further contributions to discourse studies include verification of the high proportion of implicitly expressed discourse relations, here 46%) (reported between 60 and 70% depending on genre in Taboada (2006)), and supports Sanders and Spooren’s (2009) recommendation that discourse studies moves away from an assumption of the dominance of explicit markers. Also, it shows that corpus studies can contribute converging evidence to support cognitive and processing research.
References


