

Factors Influencing the Implication of Discourse Relations across Languages

Jet Hoek*

*Utrecht Institute of Linguistics
Utrecht University, The Netherlands
j.hoek@uu.nl

Sandrine Zufferey**

**Université de Fribourg,
Switzerland
sandrine.zufferey@unifr.ch

Abstract

Relations that hold between discourse segments can, but need not, be made explicit by means of discourse connectives. Even though the explicit signaling of discourse relations is optional, not all relations can be easily conveyed implicitly. It has been proposed that readers and listeners have certain expectations about discourse and that discourse relations that are in line with these expectations (default) are more often implicit than the ones that are not (non-default). In this paper, we analyze the implication of discourse relations from a multilingual perspective. Using an annotation scheme for discourse relations based on Sanders, Spooren, & Noordman (1992), we distinguish between default and non-default discourse relations to predict the amount of implicit translations per relation in parallel corpora from four language pairs. We argue that the existing hypotheses about reader expectations are not sufficient to explain default discourse relations and propose that the rate of implication of discourse relations is governed by cognitive complexity: default discourse relations are cognitively simple within the framework of basic categories of discourse relations.

1. Introduction

Discourse connectives like *but* and *because* in English are often used to explicitly mark discourse relations such as ‘cause’ and ‘concession’ that hold between two discourse segments (Halliday & Hasan, 1976; Mann & Thompson, 1988; Sanders, Spooren, & Noordman, 1992; Knott & Dale, 1994). In addition, connectives are important for text processing, comprehension and memorization (e.g. Britton et al., 1982; Caron, Micko, & Thüring, 1988; Haberlandt, 1982; Millis, Golding, & Barker, 1995; Sanders & Noordman, 2000). Despite their usefulness, connectives are not indispensable for the communication of discourse relations, as they can often be left implicit, in which case the relation can be reconstructed by inference. The causal relation conveyed by the connective *because* in (1) can for instance still be inferred in the absence of this connective, as in (2).

- (1) John is happy because he won the race.
- (2) John is happy. He won the race.

Not all discourse relations, however, are equally easy to infer in the absence of a connective. For example, if the concessive connective *although* in (3) is removed, as in (4), the original coherence relation between the two segments is lost. In (4), the second segment is expected to be explaining the first one, but the semantic content of the relation clashes with this expectation, as the fact of losing the race is not a likely reason for being happy.

- (3) John is happy although he lost the race.
- (4) ? John is happy. He lost the race.

Sanders (2005) proposed the “causality-by-default hypothesis” for the interpretation of implicit relations, which states that hearers by default expect two segments in a discourse to be causally related. This may explain the causal interpretation triggered by the implicit relation in (4). There are, however, restrictions to this causality-by-default principle. Most importantly, the propositional content of the two segments (clauses) has to allow for a causal interpretation. Murray’s (1995; 1997)

“continuity hypothesis” suggests that readers by default expect each discourse segment to be both causally and temporally continuous with the preceding context. More specifically, by default, hearers expect events in discourse to correspond to the order in which they occurred in the world.

The roles of continuity and causality for discourse processing have been confirmed in a number of experimental studies. Murray (1997) found that when subjects are asked to continue a sentence ending with a period, their answers are often causally related to the first segment. Sanders & Noordman (2000) found that segments that are causally related to the preceding discourse are processed faster than when identical segments hold an additive relation to the preceding context. The causally related information was also recalled better. More recently, Kuperberg, Paczynski and Ditman (2011) demonstrated that causal inference influenced the processing of upcoming words in a sentence even in the absence of a connective. In addition, Koornneef and Sanders (2013) and Mak and Sanders (2013) show that causal expectations influence the processing of implicit relations and relations signaled by *because*, but not the processing of relations signaled by *but* or *and*.

Another line of evidence for these principles comes from corpus data. Asr and Demberg (2012; 2013) used the annotation of explicit and implicit relations provided in the Penn Discourse Tree Bank corpus (Prasad et al., 2008), starting from the assumption that connectives can be absent in expected relations. This assumption is related to the Uniform Information Density Hypothesis (Frank & Jaeger, 2008), which states that information is evenly spread across sentences within a text and that redundant markers tend to be omitted. Asr and Demberg calculated the ratio of implicitness for each coherence relation by dividing the number of implicit occurrences of a relation by its total number of occurrences. They reported that discontinuous relations such as ‘concession’ had indeed a lower ratio of implicitness than continuous relations like ‘cause’ and ‘addition’. They also reported more fine-grained distinctions within categories of discourse relations. More specifically, they found that temporal relations following the order of events in the world had a higher ratio of implicitness than temporal relations reversing the order of events in the world. In short, their corpus study suggests that continuous and causal relations are expected by default, leading to a higher number of implicit relations.

Asr and Demberg’s corpus studies were conducted from a monolingual perspective. An important question for the study of explicit and implicit communication of discourse relations that we investigate in this paper is whether the same principles apply cross-linguistically. We argue that if the principles that influence expectations about discourse are cognitively motivated, as they are hypothesized to be, then they should apply in the same way across languages. In order to investigate this question, we counted the number of explicit and implicit translations of several connectives conveying expected, or default, relations and connectives conveying unexpected, or non-default, relations in a large multilingual corpus. Although connectives are well known to be volatile items in translation (Halverson, 2004; Zufferey & Cartoni, 2014) and can be added, removed, or rephrased by translators, this variability should be limited by the potential of implicitness of each discourse relation. More specifically, we hypothesize that relations expressing default interpretations (i.e. continuous and causal relations) should be implicitated in translation more often than non-default relations, independently of the range of translation equivalents provided by the target language for each connective. In Section 2, we present the corpus study conducted to assess this hypothesis and we discuss its results and implications for future research in Section 3.

2. Corpus study

To test whether default relations more often receive an implicit translation than non-default relations, we extracted a set of discourse relations from the directional version of the Europarl corpus: Europarl Direct (Koehn, 2005; Cartoni, Zufferey & Meyer, 2013). Unlike the original Europarl corpus, Europarl Direct is a set of parallel corpora that only contains fragments that were originally uttered in the relevant source language, along with its corresponding translations. We selected English as our source language, and Dutch, German, French, and Spanish as our target languages: our final selection consisted of English discourse relations, with their translations in all four target languages.

As Table 1 shows, we selected discourse relations that were expressed by means of an English connective representative of certain basic features from a taxonomy of coherence relations (Sanders et al., 1992; Scholman, Evers-Vermeul, Sanders, submitted):

| Connective | Relation | Continuity |
|-----------------|---------------------------|---------------|
| <i>Also</i> | Positive, additive | Continuous |
| <i>Because</i> | Positive, causal | Continuous |
| <i>Although</i> | Negative, additive/causal | Discontinuous |
| <i>If</i> | Positive, conditional | ? |

Table 1. Basic features the English connectives prototypically convey and whether these are continuous or discontinuous.

The discourse annotation method proposed by Sanders et al. (1992) distinguishes different features of discourse relations (in addition, ‘end labels,’ or relations names, are provided for the relation as a whole; most discourse annotation schemes employ only relation names). One of these features is ‘polarity.’ A discourse relation consists of an antecedent (P) and a consequent (Q). A relation has positive polarity if the two segments, S_1 and S_2 , function as P and Q , as in (1): winning a race is a plausible reason for being happy. A relation has negative polarity if P or Q is expressed by a negative counterpart of S_1 or S_2 (not- S_1 or not- S_2), as in (3): not winning a race is not likely to result in a happy contestant. Other features of discourse relations are ‘basic operation’ (causal, additive, temporal, conditional), ‘source of coherence’ (objective, subjective, speech act), and ‘order of the segments’ (basic (P - Q) non-basic(Q - P)). The current study only discusses the polarity and the basic operation of discourse relations.

Also and *because* both signal relations with positive polarity. *Because* is used to convey causal relations, whereas *also* signals additive, non-causal, relations. The continuity hypothesis does not predict any differences in implicitation between additive and causal relations, since both can be considered continuous relations, but based on the causality-by-default hypothesis, which poses that the default interpretation of implicit relations is a causal one, we suspect that causal relations are more often implicated than additive relations. *Although* signals relations with negative polarity. Negative relations can be considered discontinuous: the discourse segments do not follow logically from each other. Instead, one of the segments functions as a negative counterpart to the other segment (e.g. contrastive cause – consequence). Since negative relations do not constitute a default interpretation by readers or listeners, we suspect that they will be less often translated implicitly than positive relations. Finally, we selected conditional relations, prototypically signaled by *if* in English. Although conditional relations cannot be categorized as either continuous or discontinuous, corpus-based studies demonstrate that they are almost always signaled by means of a connective (Asr & Demberg, 2012; Das & Taboada, 2013; Taboada & Das, 2013). We therefore expect to find a limited amount of implicit translations of relations signaled by *if*.

After randomly extracting fragments from the parallel corpora based on the presence of *although*, *because*, *also*, or *if*, we selected only those fragments in which the connective was used to signal a discourse relation. We then manually annotated the way in which the relations were translated: explicitly, implicitly, or by means of a paraphrase or syntactic construction. Translated fragments were only considered to be implicit discourse relations if they still contained a discourse relation. Although the meaning of (5) is very similar to (1) or (2), it cannot be considered an implicit discourse relation:

- (5) His victory made John very happy.

Examples such as (5) were therefore categorized as paraphrases. Fragments categorized as syntactic constructions were for instance translations of conditional relations by means of a subjunctive in German.

3. Results and discussion

The number of implicitations per relation and target language can be found in Table 2. This includes only those translations that contained a discourse relation: all instances in which the target text used a paraphrase or syntactic construction to convey the meaning of the discourse relation in the source text were excluded from the analysis. Translations in which the meaning of the original discourse relation was lost were also left out of consideration.

| | Dutch (%) | German (%) | French (%) | Spanish (%) |
|-----------------|------------------|-------------------|-------------------|--------------------|
| <i>Also</i> | 19/194 (9.50) | 15/181 (7.50) | 14/190 (7.00) | 7/195 (3.50) |
| <i>Because</i> | 27/383 (6.37) | 8/391 (1.89) | 19/389 (4.48) | 4/393 (0.94) |
| <i>Although</i> | 7/248 (2.68) | 2/248 (0.77) | 5/256 (1.91) | 1/261 (0.38) |
| <i>If</i> | 0/226 (0.00) | 0/201 (0.00) | 0/222 (0.00) | 0/241 (0.00) |

Table 2. Number of implications per English connective per target language compared to the total of all discourse relations in the target text that corresponded to the original relation in the source text.

The results from Table 2 indicate that the overall rate of implicitation appears to differ between language pairs. In English-Dutch and English-French translations, for instance, a lot more relations appear implicitly in the target texts than in English-Spanish translations. This corresponds to observations from both quantitative and qualitative analyses of translations (e.g. Becher, 2011; Cartoni, Zufferey, Meyer, & Popescu-Belis, 2011). Despite the overall difference in the level of implicitation, similar relative differences between the implicitation of relations can be observed, see Tables 3a-d in the appendix for details.

As we hypothesized, in all languages positive additive relations (signaled by *also*) are translated implicitly significantly more often than negative (*although*) (all $p < 0.05$) and conditional (*if*) relations (all $p < 0.001$). In both Dutch and French positive causal relations (*because*) are also translated implicitly significantly more often than negative and conditional relations (all $p < 0.05$), but these differences were not found for Spanish (causal vs. negative $p = 0.338$, causal vs. conditional $p = 0.147$). In German, we found a significant difference between causal relations and conditional relations ($p < 0.05$), but not between causal relations and negative relations ($p = 0.183$). A comparison between additive and causal relations indicated that there was a significant difference in implicitation in the other direction than we initially hypothesized on the basis of the causality-by-default hypothesis: in all target languages except for Dutch ($p = 0.16$), we found more implicit additive relations than implicit causal relations (all $p < 0.05$). Finally, we found more implicit negative relations than conditional relations in both Dutch and French (both $p < 0.05$). In fact, we did not find instances of implicated conditional relations in any of the target languages.

Although these results partly confirm our hypotheses, they seem to call for a reconsideration of the role of default interpretations, or expectations, in the implicitness of discourse relations. The causality-by-default hypothesis poses that readers expect relations to be causally related. However, it also points out that “readers will ... arrive at an additive relation if no causal relation can be established” (Sanders, 2005, p. 9). The causality-by-default hypothesis rightly predicts that causal relations can often be implicit, but when applied to the implicitness of discourse relations it can conversely be interpreted as blocking the possibility of expressing an additive relation without a connective when the resulting implicit relation can be interpreted as a causal one (additive connectives have been claimed to block a causal interpretation of discourse relations, e.g. Levinson (2000), Koornneef and Sanders (2013), and Mak and Sanders (2013)). Indeed, none of the implicit additive relations in our study allow for a causal relation. Our finding that there were more implicit additive relations than causal relations in three out of four target languages might thus largely be influenced by the number of additive relations that would, when implicit, allow for a causal interpretation. We will address this question in the continuation of this study.

The hypothesis that default interpretations cannot only facilitate implicitation but also block it can be extended to the continuity hypothesis: when a relation does not constitute a default relation (either positive causal or positive additive), it can only be implicit if the content of the two segments blocks the default interpretation, or when there is enough evidence in favor of a non-default relation, for instance the presence of word pairs (e.g. Halliday and Hasan’s (1976) ‘semantic relations’ or Taboada and Das’s (2013) ‘entity features’). Corroborating this hypothesis are the few conditional relations we found in our study from which the conditional connective had been removed: these can no longer be interpreted as conditional relations and instead receive a causal or additive relation. Because the relation in the target text did not correspond to the relation in the source text, we did not consider these relations to be examples of implicit conditional relations.

Our results support findings from monolingual corpus studies that conditional relations are usually explicit. Conditional relations can therefore be supposed to constitute non-default

interpretations, but neither the causality-by-default hypothesis nor the continuity hypothesis can account for this. We therefore propose that default interpretations are governed by cognitive complexity. In the framework of basic features of discourse relations we employ, there are cognitively more complex members on each level: relations with negative polarity are for instance more complex than relations with positive polarity, relations with a non-basic order of segments are more complex than basic order relations, and conditional relations are more complex than non-conditional relations (see also Evers-Vermeul & Sanders, 2009). The cognitively simple alternatives then constitute the default interpretations. Note that this hypothesis makes rather fine-grained predictions, as a relation can be relatively complex or simple, default of non-default, on multiple levels. To test this hypothesis, we will continue to extend the current study by manually annotating each relation in the source language and determining the specific relation they signal within the framework of basic features.

Acknowledgments

We are grateful to the Swiss National Science Foundation (SNSF) for the funding of this work under its Sinergia program, grant n. CRSII2_147653 (MODERN project, see www.idiap.ch/project/modern/).

Appendix

Dutch

| | <i>also</i> | <i>although</i> | <i>if</i> |
|-----------------|-------------------------|-------------------------|---------------------------|
| <i>because</i> | $\chi^2=0.97$ p=0.16 | $\chi^2=4.48$ p<0.05 | $\chi^2=15.05$ p<0.001 |
| <i>also</i> | | $\chi^2=8.34$ p<0.05 | $\chi^2=20.97$ p<0.001 |
| <i>although</i> | | | $\chi^2=4.68$ p<0.05 |

Table 3a English-Dutch data

German

| | <i>also</i> | <i>although</i> | <i>if</i> |
|-----------------|---------------------------|---------------------------|---------------------------|
| <i>because</i> | $\chi^2=10.92$ p<0.001 | $\chi^2=0.82$ p=0.183 | $\chi^2=9.63$ p<0.05 |
| <i>also</i> | | $\chi^2=13.48$ p<0.001 | $\chi^2=19.15$ p<0.001 |
| <i>although</i> | | | p=0.305* |

Table 3b English-German data

French

| | <i>also</i> | <i>although</i> | <i>if</i> |
|-----------------|-------------------------|--------------------------|---------------------------|
| <i>because</i> | $\chi^2=2.78$ p<0.05 | $\chi^2=2.93$ p<0.05 | $\chi^2=9.63$ p<0.001 |
| <i>also</i> | | $\chi^2=9.68$ p<0.001 | $\chi^2=18.21$ p<0.001 |
| <i>although</i> | | | p<0.05* |

Table 3c English-French data

Spanish

| | <i>also</i> | <i>although</i> | <i>if</i> |
|-----------------|-------------|-----------------|-----------|
| <i>because</i> | p=0.036* | p=0.338* | p=0.147* |
| <i>also</i> | | p<0.05* | p<0.001* |
| <i>although</i> | | | p=0.520* |

Table 3d English-Spanish

Tables 3a-d. Comparison of implicitation of discourse relations in the parallel corpora. All measures are Chi-square analyses (one-sided, df=1), unless marked with *: these are Fisher's exact tests (one-sided, df=1).

References

- Asr, F.T. & Demberg, V. (2012). Implicitness of discourse relations. *Proceedings of COLING*. Mumbai, India.
- Asr, F.T. & Demberg, V. (2013). On the information conveyed by discourse markers. *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics*, 84-93. Sofia: Bulgaria.

- Becher, V. (2011). When and why do translators add connectives? *Target* 23(1), 26-47.
- Britton, B.K., Glynn, S.M., Meyer, B.J.F., & Penland, M.J. (1982). Effects of text structure on the use of cognitive capacity during reading. *Journal of Educational Psychology* 74, 51–61.
- Caron, J., Micko, H.C., & Thuring, M. (1988). Conjunctions and the recall of composite sentences. *Journal of Memory and Language* 27, 309–323.
- Cartoni, B., Zufferey, S., & Meyer, T. (2013). Using the Europarl corpus for linguistics research. *Belgian Journal of Linguistics* 27, 23-42.
- Cartoni, B., Zufferey, S., Meyer, T., & Popescu-Belis, A. (2011). How comparable are parallel corpora? Measuring the distribution of general vocabulary and connectives. *Proceedings of the 4th Workshop on Building and Using Comparable Corpora*, 78-86. Portland, Oregon.
- Das, D. & Taboada, M. (2013). Explicit and implicit coherence relations: A corpus study. *Proceedings of the 2013 annual conference of the Canadian Linguistic Association*.
- Evers-Vermeul, J. & Sanders, T.J.M. (2009). The emergence of Dutch connectives; how cumulative cognitive complexity explains the order of acquisition. *Journal of Child Language* 36(4), 829-854.
- Frank, A. & Jaeger, T. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. *Proceedings of the 28th meeting of the Cognitive Science Society*.
- Haberlandt, K. (1982). Reader expectations in text comprehension. In: J-F Le Ny & W Kintsch (Eds.), *Language and Comprehension*. North-Holland, Amsterdam, 239-249.
- Halliday, M.A.K. & Hasan, R. (1976). *Cohesion in English*. Longman, London.
- Halverson, S. (2004). Connectives as a translation problem. In: H. Kittel et al. (Eds.), *An International Encyclopedia of Translation Studies*, 562–572. Berlin/New York: Walter de Gruyter.
- Knott, A. & Dale, R. (1994). Using linguistic phenomena to motivate a set of coherence relations. *Discourse Processes* 18, 35–62.
- Koehn, P. (2005) Europarl: A parallel corpus for statistical machine translation. *Proceedings of the 10th Machine Translation Summit*, 79-86. Phuket, Thailand.
- Koornneef, A.W., & Sanders, T.J.M. (2013). Establishing coherence relations in discourse: The influence of implicit causality and connectives on pronoun resolution. *Language and Cognitive Processes* 28(8), 1169-1206.
- Kuperberg, G., Paczynski, M., and Ditman, T. (2011). Establishing causal coherence across sentences: An ERP study. *Journal of Cognitive Neuroscience* 23, 1230–1246.
- Mak, W.M., Sanders, T.J.M. (2013). The role of causality in discourse processing: Effects of expectation and coherence relations. *Language and Discourse Processes* 28(9), 1414-1437.
- Mann, W.C., & Thompson, S.A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text* 8, 243–281.

- Millis, K.K., Golding, J.M., & Barker, G. (1995). Causal connectives increase inference generation. *Discourse Processes* 20(1), 29–49.
- Murray, J.D. (1995). Logical connectives and local coherence. In R. Lorch & E. O'Brien (Eds.), *Sources of cohesion in text comprehension*, 107-125. Hillsdale, NJ: Erlbaum.
- Murray, J.D. (1997). Connectives and narrative text: The role of continuity. *Memory and Cognition* 25, 227–236.
- Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., & Webber, B. (2008). The Penn Discourse Treebank 2.0. *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, 2961–2968.
- Sanders, T.J.M. (2005). Coherence, causality and cognitive complexity in discourse. *Proceedings/Actes SEM-05, First International Symposium on the Exploration and Modelling of Meaning*, 105–114.
- Sanders, T.J.M., & Noordman, L.G.M. (2000). The role of coherence relations and their linguistic markers in text processing. *Discourse Processes* 29, 37–60.
- Sanders T.J.M., Spooren W.P.M.S. & Noordman L.G.M. (1992). Towards a Taxonomy of Coherence Relations. *Discourse Processes* 15, 1–36.
- Scholman, M.C.J., Evers-Vermeul, J., & Sanders, T.J.M. (subm.). *Categories of coherence relations in discourse annotation*. Submitted for publication.
- Taboada, M. & Das, D. (2013). Annotation upon annotation: Adding signalling information to a corpus of discourse relations. *Dialogue and Discourse* 4(2), 249-281.
- Zufferey, S. & Cartoni, B. (2014). A multifactorial analysis of explicitation in translation. *Target* 26, 361–384.