The Role of Expectedness in the Implicitation and Explicitation of Discourse Relations

Jet Hoek
UiL-OTS, Utrecht University
Trans 10, NL-3512 JK
Utrecht, The Netherlands
j.hoek@uu.nl

Jacqueline Evers-Vermeul
UiL-OTS, Utrecht University
Trans 10, NL-3512 JK
Utrecht, The Netherlands
j.evers@uu.nl

Ted J.M. Sanders
UiL-OTS, Utrecht University
Trans 10, NL-3512 JK
Utrecht, The Netherlands
t.j.m.sanders@uu.nl

Abstract

Translation of discourse connectives varies more in human translations than in machine translations. Building on Murray's (1997) continuity hypothesis and Sanders' (2005) causality-by-default hypothesis we investigate whether expectedness influences the degree of implicitation and explicitation of discourse relations. We manually analyze how source text connectives are translated, and where connectives in target texts come from. We establish whether relations are explicitly signaled in the other language as well, or whether they have to be reconstructed by inference. We demonstrate that the amount of implicitation and explicitation of connectives in translation is influenced by the expectedness of the relation a connective signals. In addition, we show that the types of connectives most often added in translation are also the ones most often deleted.

1 Introduction

Discourse relations that hold between text segments can be explicitly signaled through connectives, but can also remain unmarked. For example, the causal relation in (1a) is explicitly encoded by the connective *because*. In its implicit counterpart in (1b), this causal relation has to be reconstructed by inference.

(1) a. Mike opened his umbrella *because* it was raining.
b. Mike opened his umbrella. It was raining.

In translation, connectives are very volatile items and can be added or removed between source text (ST) and target text (TT) (Halverson, 2004; Zufferey and Cartoni, 2014). Human translators more often leave out or reformulate a connective (up to 18%) than statistical machine translation models (up to 8%) (Meyer and Webber, 2013). In addition, when connectives are left out of machine translation (MT) output, this is not always justified and can result in translations that do not correspond to the original texts (cf. Li et al., 2014; Steele and Specia, 2014).

Specific deletions or additions of connectives in human translations have often been attributed to differences in linguistic resources between the languages in a translation pair (e.g. Becher, 2011; Hansen-Schirra et al., 2007). Other studies, however, have proposed that the deletion or addition of a connective is (also) dependent on the type of discourse relation a connective signals (e.g. Halverson 1996; Hoek and Zufferey, 2015). This study represents a first step in an effort to identify the factors that influence whether a connective can be left out of a translation without changing the interpretation of a fragment, or whether a connective should be translated into a target text by means of a comparable target language connective or another linguistic construction that expresses the same meaning. This knowledge can eventually be used to create MT systems that can translate explicit relations into implicit relations and vice versa in an idiomatic and fluent way that approaches the output of human translators.

Discourse-annotated corpora that include both implicit and explicit relations reveal that certain types of relations are easier to convey implicitly than others (Asr and Demberg, 2012; Das and Taboada, 2013; Versley, 2013). Causal relations, as in (1), for instance, appear more often without a connective or a cue phrase than negative relations, as in (2), or conditional relations, as in (3). The question marks in the b-sentences indicate that it is difficult to arrive at the negative or conditional interpretation, respectively, of the relations in the a-sentences.

(2) a. Ann is happy, *although* she lost the race.
b. $^{??}$Ann is happy. She lost the race.

(3) a. *If* he wants to be rich someday, he should get off the couch.
b. $^{??}$He wants to be rich someday. He should get off the couch.

Proceedings of the Second Workshop on Discourse in Machine Translation (DiscoMT), pages 41–46,
Lisbon, Portugal, 17 September 2015. ©2015 Association for Computational Linguistics.

In this paper, we pursue the idea that the types of discourse relations that are often implicit correspond to the types of relations people expect in a discourse. According to the continuity hypothesis (Murray, 1997) and the causality-by-default hypothesis (Sanders, 2005), continuous and causal relations are generally the expected types. These hypotheses are corroborated by processing studies (e.g. Koornneef and Sanders, 2013; Kuperberg et al., 2011; Mak and Sanders, 2013; Sanders and Noordman, 2000) and corpus-based research. Asr and Demberg (2012) for instance demonstrate that the implicit relations in the Penn Discourse Treebank (PDTB, Prasad et al. 2008) are often continuous and/or causal.

If types of discourse relations differ in their degree of expectedness, and thereby in their degree of implicitness in monolingual texts, this should affect translation. In other words: we hypothesize that a discourse relation's potential to remain implicit (because of its expectedness) influences how often that type of relation is implicitated or explicitated in translation. For expected types of relations, which are often implicit in the ST, there are many instances at which translators can choose (either deliberately or subconsciously) to add a connective. Conversely, when an expected relation is explicitly marked in the ST, there will often be the option of leaving out the connective in the TT. What this predicts, then, is that markers of the types of relations that are most often added in translation will also be the ones most often deleted, regardless of language pair or translation direction. In this study, we test these predictions by comparing additions and deletions of connectives in two language pairs (English-Dutch and English-German) from the Europarl Direct corpus[1] (Koehn, 2005; Cartoni et al., 2013), and determining how the (interpretation of the) discourse relation in the ST or TT is conveyed in the other language.

2 Method

We define ***implicitness*** and ***explicitness*** as monolingual concepts that refer to whether the interpretation of, in this case, a discourse relation is explicitly encoded, as in (1a), or if it has to be reconstructed by inference, as in (1b). We use ***implicitation*** and ***explicitation*** to refer to shifts in implicitness or explicitness between ST and TT. In case of implicitation, the TT is more implicit than the ST. In case of explicitation, the TT is more explicit than the ST.

For this study, we compared three types of discourse relations: causal, negative, and conditional relations. Causal relations are among the expected types of relations, while negative and conditional relations are not. We therefore expect more implicitations and explicitations of causal relations than of negative or conditional relations. We selected prototypical connectives signaling these relation types in all three languages in our corpus, see Table 1.

	English	**Dutch**	**German**
Causal	*because*	*omdat*	*weil*
Negative	*although*	*hoewel*	*obwohl*
Conditional	*if*	*als*	*wenn*

Table 1. Connective selection per language and type of relation

We automatically extracted English ST fragments containing *because*, *although*, and *if* from the Europarl Direct corpus, along with their translations in Dutch and German. We also extracted Dutch and German TT fragments containing *omdat*, *hoewel*, and *als*, and *weil*, *obwohl*, and *wenn*, respectively, along with the corresponding English ST fragments. We randomly selected 250 instances of each connective and made sure these were used to mark a discourse relation. In total, we had 3000 ST-TT fragment pairs.

2.1 Annotation

For all connectives we determined how they were translated, or what they were a translation of. In the analysis we used the categories *explicit*, *paraphrase*, *underspecified connective*, *syntax*, and *implicit*.

In *explicit* cases, the connective corresponds to a similar connective or cue phrase in the other language. In the *paraphrase* category the type of relation is still explicitly encoded in the text, but with different linguistic means, as in (4). We coded a fragment as *implicit* if it contained a relation not marked by means of any connective or cue phrase. (5) is an example of implicitation, since the relation is explicitly encoded in the ST, but implicit in the TT. The implicitness is indicated with the Ø symbol.

[1] The Europarl Corpus is a version of the original Europarl corpus that only includes ST fragments that were originally uttered in that language, e.g. all fragments in the EN-DU part of the corpus were originally uttered in English. The corpus is aligned per language pair.

(4) (ep-96-07-03)
TT ***Hoewel*** “*although*” wij wegens de politieke situatie in Italië zelf aanvankelijk twijfels hadden, heeft het voorzitterschap toch opmerkelijke resultaten geboekt.
ST ***Despite*** the initial doubts we had due to the domestic political situation, there were some significant achievements …

(5) (ep-98-02-20)
ST Insofar as the POSEIMA programme is concerned, we have to admit that you could not think of a more complicated or indirect or inefficient way to aid islands or remote regions, ***because*** in the first place there is no guarantee whatsoever that this money is going to the aid of the people who need it or for whom it was intended.
TT Voor wat betreft het POSEIMA-programma kunnen we alleen maar toegeven dat dit wel de ingewikkeldste, minst directe en meest inefficiënte manier is die je kunt bedenken om hulp te bieden aan eilandregio’s en plattelandsgebieden. **Ø** Ten eerste bestaat er geen enkele garantie dat het geld inderdaad bij de mensen terechtkomt die het nodig hebben en voor wie het ook is bedoeld.

In the Dutch TT in (4), the connective *hoewel* “although” expresses a negative relation. The English ST does not contain this negative relation, but uses *despite* plus a noun phrase to explicitly indicate contrast.

In addition, connectives in a ST or TT that were less specific than the corresponding connectives in the other text were considered *underspecified connectives*. In these cases, neither the original nor the translation contains an implicit discourse relation. See for example (6), where the original temporal relation is marked with a more specific causal connective in the German translation. Hence, the translation can be seen as a case of explicitation.

(6) (ep-98-05-27)
TT Wir haben diesen Änderungsantrag im Namen von Herrn Wynn vorgelegt, um die Frage der Personalplanung für den Bürgerbeauftragten noch bis zur ersten Lesung offen zu lassen, ***weil*** “*because*” wir dann den gesamten Personalbedarf genauer einschätzen können.
ST We have tabled this amendment in Mr Wynn’s name in order to leave the matter of staffing for the Ombudsman open until the first reading ***when*** one will have a clearer view of the overall need concerning staff.

Furthermore, we distinguish a *syntax* category, in which the syntax of the fragment is dramatically different from the corresponding fragment containing the connective and the relation disappears altogether, as in (7).

(7) (ep-00-03-14)
TT Een aantal Britse leden van het Europees Parlement zijn benaderd door belangengroepen van landbouwers, ***omdat*** “*because*” deze bang zijn dat de verbrandingsrichtlijn ook van toepassing zal zijn op alle verbrandingsinstallaties op boerenbedrijven.
ST A number of United Kingdom MEPs have been contacted by farming interests, ***who*** are very worried that the incineration directive will apply to all on-farm incinerators in the United Kingdom.

In (7) the causal relation signaled by *omdat* “because” in the Dutch TT is absent in the English ST. The second clause in the Dutch causal relation corresponds to a relative clause in the English ST, which does not explicitly signal causality. Instead, it has to be inferred by readers or listeners that the content of the relative clause presents the reason why farming interest groups have been contacting MEPs.

Two trained annotators, the first and second author of this paper, annotated the first 50 fragments for each connective for each language pair and translation direction (6x50 fragments). After establishing that there was a good inter-annotator agreement (κ = 0.84) and discussing the fragments that were disagreed on, one annotator finished the annotation of the remaining fragments.

On the basis of the annotations, we established for each ST-TT fragment pair whether it constituted a case of implicitation or explicitation. The categories *underspecified connective*, *syntax*, and *implicit* were considered to be instances of implicitation if they showed up in the TT equivalents of ST connectives, and instances of explicitation if they showed up in the ST equivalents of TT connectives. The categories *explicit* and *paraphrase* were grouped together as explicit-to-explicit translations. Statistical analysis was thus conducted on two categories instead of five.

2.2 Data analysis

Log-linear analysis was used to estimate the probability of occurrence of implicitations/

explicitations. The null model estimates the average probability. This model was compared to more complex models in which the probability was estimated as a function of our variables and the interactions between them: *relation type* (causal vs. negative vs. conditional), *marking* (implicit in the other language vs. explicit in the other language), *language pair* (EN-DU vs. EN-GE), and *direction* (ST→TT vs. TT→ST).

3 Results

The model in which all variables and several interactions were included was the best model. It retained a main effect of *marking* (χ^2 (1) = 3051.65, p < .001), two-way interactions of *relation type* and *marking* (χ^2 (2) = 82.91, p < .001), and of *marking* and *direction* (χ^2 (1) = 6.23, p = .01), plus a three-way interaction of *language pair*, *marking*, and *direction* (χ^2 (1) = 10.38, p = .001).

The two-way interaction between *relation type* and *marking* indicates that the amount of implicitation and explicitation of connectives in translation is influenced by the type of relation they signal. This relationship is visualized in Figure 1. As we hypothesized, causal relations were more often implicit than negative relations (z = 6.21, p < .001), which in turn showed more implicitation than conditional relations (z = 4.72, p < .001).

Taken together, the three-way interaction between *language pair*, *marking*, and *direction*, and the two-way interaction between *marking* and *direction* indicate the following. The English-German pairs adhere to the two-way interaction: the number of explicitations (explicit in TT, implicit in ST) was higher than the number of implicitations (explicit in ST, implicit in TT). This implies that connectives in German translations stem relatively frequently from an *underspecified connective*, another *syntax* or an *implicit* relation, while English ST connectives are hardly implicitated when translated into German TT.

For English-Dutch, this directional difference does not hold: the number of implications from ST to TT is higher than in German (z = 2.53, p = .01). This can also be derived from Figure 1, which illustrates that for EN-DU the overall number of implicitations is comparable to the overall number of explicitations. Crucially, the three-way interaction does not involve *relation type*, which means that the difference between EN-DU and EN-GE was not affected by the type of relation.

4 Discussion and conclusion

Our results show that the expectedness of discourse relations, as defined on the basis of the continuity hypothesis and the causality-by-default hypothesis, affects translation. Causal connectives, which are expected in discourse, are both more often added and deleted in translation than relations that are not expected, in this case negative and conditional connectives. We also found that negative connectives were more often added and deleted than conditional connectives.

Since this study included only English-Dutch and English-German translations, and Dutch and German are closely related languages, it may be possible that the implicitation and explicitation patterns we found are generalizable only within the language family. However, in an earlier study in which we only looked at the translations of ST connectives we also included English-French and English-Spanish translations (Hoek and Zufferey, 2015). Here we found identical implicitation patterns for French and Spanish (both of which belong to a different language family) as for Dutch and German. This suggests that our results are also generalizable across language families.

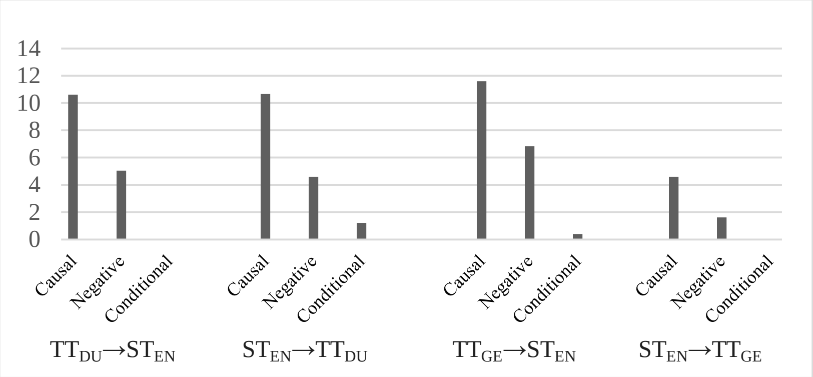

Figure 1. Percentage of implicit translations/originals per type of relation, per language pair

Our finding that negative connectives were more often added and deleted than conditional connectives is not predicted by the continuity hypothesis or the causality-by-default hypothesis, but it seems to be corroborated by corpus studies. Conditional relations hardly ever seem to be implicit in monolingual corpora, while this is less rare for negative relations (e.g. Asr and Demberg, 2012; Das and Taboada, 2013). We will address the difference between negative and conditional relations in further research.

We found more explicitations than implicitations for English-German translations, but not for English-Dutch translations. The observation that translation pairs and translation directions can differ in the overall number of connectives that are added or deleted has also been made in corpus-based studies (e.g. Becher, 2011; Cartoni et al., 2011). This effect did not, however, interact with the relative frequencies of implicitation or explicitation of relation types.

It should be noted the frequency of implicitations (3.6%) that we found was much lower than the frequency reported by Meyer and Webber (2013) (up to 18%). This can probably be attributed to our relatively broad definition of the explicit category *paraphrase*, which for instance included verbs expressing causality (e.g. *make*, *cause*) and the subjunctive in German, since this explicitly encodes conditionality. If we were to include all paraphrases in our implicitations, we would arrive at a higher percentage of 11.2%.

The potential to remain implicit appears to influence how often a relation is implicitated or explicitated in translation. To improve the quality and naturalness of machine translation, it therefore seems crucial to distinguish between deletions and additions of connectives in which the relation in the other language is implicit and those in which the relation is marked by different linguistic means, and to incorporate factors that influence whether a relation can be left implicit or whether it should be explicitly signaled into a machine translation model.

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/).

References

Fatemeh T. Asr and Vera Demberg. 2012. Implicitness of discourse relations. *Proceedings of COLING*. Mumbai, India.

Viktor Becher. 2011. When and why do translators add connectives? *Target*, 23(1):26–47.

Bruno Cartoni, Sandrine Zufferey and Thomas Meyer. 2013. Using the Europarl corpus for linguistics research. *Belgian Journal of Linguistics*, 27:23–42.

Bruno Cartoni, Sandrine Zufferey, Thomas Meyer and Andrei Popescu-Belis. 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.

Debopam Das and Maite Taboada. 2013. Explicit and implicit coherence relations: A corpus study. *Proceedings of the 2013 Annual Conference of the Canadian Linguistic Association.* Victoria, Canada.

Sandra Halverson. 1996. Norwegian-English translation and the role of certain connectives. *Translation and Meaning, part 3: Proceedings of the Maastricht Session of the 2nd International Maastricht-Lodź Duo Colloquium on 'Translation and Meaning'*, 128–139. Maastricht, the Netherlands.

Sandra Halverson. 2004. Connectives as a translation problem. In Harald Kittel, Armin Paul Frank, Norbert Greiner, Theo Hermans, Werner Koller, José Lambert and Fritz Paul. (Eds.), *An International Encyclopedia of Translation Studies*, 562–572. Berlin/New York: Walter de Gruyter.

Silvia Hansen-Schirra, Stella Neuman and Erich Steiner. 2007. Cohesive explicitness and explicitation in an English-German translation corpus. *Languages in Contrast* 7(2):241–265.

Jet Hoek and Sandrine Zufferey. 2015. Factors influencing the implicitation of discourse relations across languages. *Proceedings 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation* (*ISA-11*), 39–45. London, United Kingdom.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. *Proceedings* *of the 10th Machine Translation Summit*, 79–86. Phuket, Thailand.

Arnout W. Koornneef and Ted J. M. Sanders. 2013. Establishing coherence relations in discourse: The influence of implicit causality and connectives on pronoun resolution. *Language and Cognitive Processes*, 28(8):1169–1206.

Gina R. Kuperberg, Martin Paczynski and Tali Ditman. 2011. Establishing causal coherence across sentences: An ERP study. *Journal of Cognitive Neuroscience*, 23:1230–1246.

Junyi Jessy Li, Marine Carpuat and Ani Nenkova. 2014. Assessing the discourse factors that influence the quality of machine translation. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 283–288. Baltimore, MD, USA.

Willem M. Mak and Ted J. M. Sanders. 2013. The role of causality in discourse processing: Effects of expectation and coherence relations. *Language and Discourse Processes*, 28(9):1414–1437.

Thomas Meyer and Bonnie Webber. 2013. Implicitation of discourse connectives in (machine) translation. *Proceedings of the 1st DiscoMT Workshop at ACL 2013*, 33–42. Sofia, Bulgaria.

John D. Murray. 1997. Connectives and narrative text: The role of continuity. *Memory and Cognition*, 25:227–236.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi and Bonnie Webber. 2008. The Penn Discourse Treebank 2.0. *Proceedings of the 6th International Conference on Language Resources and Evaluation* (LREC 2008), 2961–2968. Marrakech, Morocco.

Ted J. M. Sanders. 2005. Coherence, causality and cognitive complexity in discourse. *Proceedings/Actes SEM-05, First International Symposium on the Exploration and Modelling of Meaning*, 105–114.

Ted J. M. Sanders and Leo G. M. Noordman. 2000. The role of coherence relations and their linguistic markers in text processing. *Discourse Processes*, 29:37–60.

David Steele and Lucia Specia. 2014. Divergences in the usage of discourse markers in English and Mandarin Chinese. *Proceedings of the 17th International Conference on Text, Speech and Dialogue*, 189–200. Brno, Czech Republic.

Yannick Versley. 2013. A graph-based approach for implicit discourse relations. *Computational Linguistics in the Netherlands Journal*, 3:148–173.

Sandrine Zufferey and Bruno Cartoni. 2014. A multifactorial analysis of explicitation in translations. *Target*, 26:361–384.