Translation Effect on Discourse Connective Choice

Frances Yung, Merel Scholman, Ekaterina Lapshinova-Koltunski,
Christina Pollkläsener & Vera Demberg*

Abstract. Discourse connectives are often added, omitted, or rephrased in translation. Prior work has shown a tendency for explicitation of discourse connectives, but such work was conducted using restricted sample sizes due to difficulty of connective identification and alignment. The current study exploits automatic methods to facilitate a large-scale study of connectives in English and German parallel texts. Our results based on over 300 types and 18000 instances of aligned connectives and an empirical approach to compare the cross-lingual specificity gap provide strong evidence of the *Explicitation Hypothesis*. We conclude that using relative entropy to study the specificity of connectives can provide more fine-grained insights into translation patterns.

Keywords. discourse relations; human translation; entropy; word alignment; explicitation; implicitation

1. Introduction. Discourse connectives such as *because* and *however* are often described as "volatile items" in translation: translators often add, rephrase or remove them (e.g. Zufferey & Cartoni 2014). Prior studies have focused specifically on whether connectives are added (i.e. the relation sense is *explicitated*) or removed (i.e. *implicitated*), and have shown that there is a tendency for explicitation in translation (but this also depends on various other factors, see e.g., Hoek et al. 2015, 2017, Lapshinova-Koltunski et al. 2022, Zufferey 2016). The current work focuses on a less studied aspect of connectives in translation, namely when they are underspecified (e.g. connectives like "and" or "but" are compatible with many different types of discourse relations) or highly specific (e.g. the connective "nevertheless" can only mark concessive relations). The question we address is whether we can see a similar pattern of explicitation of connectives in translation for connectives that are already explicit (but possibly unspecific) in the source text.

One factor that impedes a comprehensive study of DCs in translation is the (manual) annotation effort that is required for this task. Consequently, many studies are restricted to limited samples and a subset of DCs. To facilitate a more comprehensive investigation, we explore an automatic approach to identify and align connectives. Specifically, we use language-specific discourse parsers (Bourgonje 2021, Knaebel 2021) and a neural word alignment model (Dou & Neubig 2021) to link a large range of connectives and their translations in English and German parallel texts. We test the feasibility of this approach by replicating the well-established explicitation results in our newly created dataset. Using an empirical measure of cross-lingual specificity gap, we identify all

*This work was supported by the Deutsche Forschungsgemeinschaft, Funder Id: http://dx.doi.org/10.13039/501100001659, Grant Number: SFB1102: Information Density and Linguistic Encoding, by the the European Research Council, ERC-StG Grant no. 677352. Authors: Frances Yung, Saarland University (frances@coli.uni-saarland.de), Merel Scholman, Saarland University & Utrecht University (m.c.j.scholman@coli.uni-saarland.de), Ekaterina Lapshinova-Koltunski, University of Hildesheim (lapshinovakoltun@uni-hildesheim.de), Christina Pollkläsener, Saarland University (s8pochri@teams.uni-saarland.de), & Vera Demberg, Saarland University (vera@coli.uni-saarland.de).

the cases of (under)specifications instead of a subjectively defined subset and demonstrate evidence for explicitation in translation, in terms of both insertion and specification of DCs.

2. Background.

2.1. EXPLICITATION HYPOTHESIS. One of the most well-known accounts of translation effects, the Explicitation Hypothesis, suggests that translations tend to be more explicit than the source texts (Blum-Kulka 1986). Klaudy (1998) more specifically distinguishes between *obligatory explicitations* and *translation-inherent explicitations*. Obligatory explicitation results from grammatical and stylistic differences between the source and target languages, as well as pragmatic and cultural preferences of the source and target readers. For example, Becher (2010) found that over 50% of *damit* instances in German translated texts are the result of explicitation, but all except a few are explicitations that address the cross-lingual contrast.

By contrast, translation-inherent explicitations are language-independent and depend on the nature of the translation process. In order to identify any translation-inherent explicitations, corresponding *implicitation in the opposite translation direction* should be taken into account (Klaudy 2009). That is, explicitation due to the contrast in the explicitness of the source and target languages (with some languages being more prone to expressing discourse relations through explicit connectives than others), should be counter-balanced by the degree of implicitation when translating in the other direction. Becher (2011b) found that the insertions of discourse connectives in English to German translation are in fact more than the number of omissions in German to English translation, but still, most of the insertions can be qualitatively explained by the known observation that German is more explicit than English (Hawkins 1986, House 2014, Becher 2011a).

Various other factors have also been found to affect the explicitation of connectives, such as the type of the coherence relations and the connectives involved (Zufferey & Cartoni 2014, Crible et al. 2019), the identity of the source and target languages (Zufferey 2016), register and translator expertise (Dupont & Zufferey 2017), contrast between the constraints and communicative norms of the source and target languages (Marco 2018), the cognitive interpretability and expectedness of the relations in context (Hoek et al. 2015, 2017), information density and the mode of translation (Lapshinova-Koltunski et al. 2022).

2.2. EXPLICITATION OF DCS IN TRANSLATION. Much of the earlier work on explicitation of DCs focused largely on cases where connectives are inserted or omitted in translation or they provided qualitative estimations of specificity without basing it on a quantitative method (Crible et al. 2019, Lapshinova-Koltunski et al. 2022). In the current work, we quantify the specificity gap between a connective and its translation, to identify cases where, for example, a stronger connective is used in translation (e.g. English "*and*" translated as German "*außerdem*"). Our empirical approach allows us to objectively identify all cases where a more specified connective verbalizes the relation to a greater degree.

The specificity of connectives likely differs between languages due to the contrast between the connective lexicons and discourse marking of these languages. One connective could therefore appear to be more specific than another connective in a different language due to differences between the lexicons, even though both connectives express a similar range of relation senses. Moreover, previous studies found that the explicitation pattern of a given connective in a target language is

directly related to the alternative options available in that language (Becher 2011b, Zufferey & Cartoni 2014). To address the issue of cross-lingual correspondence, we derive estimates of a connective's specificity empirically by normalizing connectives' entropy value within a language.

3. Methodology. We analyze the parallel texts taken from the Europarl Direct Corpus (Cartoni & Meyer 2012), which are proceedings from the European Parliament. The data contains 171k tokens of English texts and their German translation from 18 proceedings, and 95k tokens of German texts and their English translation from 15 proceedings.

We use two language-specific parsers to identify and annotate the discourse relations in the English and German texts (Knaebel 2021, Bourgonje 2021) and align the identified connectives cross-lingually using the Awesome Align word alignment model (Dou & Neubig 2021). We analyze the alignments of the source/target English and German texts respectively, in order to identify explicitation and implicitation in both translation directions.

In addition, we determine the specificity level of each English and German connective based on their manual annotation in existing discourse-annotated resources. We extract the distribution of sense labels assigned to the *explicit* connectives in PDTB3.0 for English connectives and the PCC2.0 corpus for German connectives (Bourgonje & Stede 2020). We define the specificity of each connective by the entropy of its sense distribution in relation to the entropy of all explicit relations in the corresponding corpus, and round the values to 1 decimal place. We call this measure *relative entropy*. Overall, we assign *relative entropy* to 173 English and 126 German connective types. The average relative entropy of the English and German connectives are 0.122 and 0.065 respectively.

4. Results. We first look at how connectives are implicitated and explicitated in English and translations, and then we take a closer look at how the English and German connectives correspond to each other.

A total of 8058 English and 9739 German connectives have been identified and annotated by the discourse parsers and aligned. Table 1 shows the proportions of automatically identified connectives that are aligned to *"null"* or a DC of higher entropy in the other language, grouped by four categories of relations as identified by the discourse parsers.[1]

It can be observed that, when translating from English to German (top sub-table), more DCs are added than removed (26.1% vs 13.8%). The reverse is observed in German to English translation (bottom sub-table), where more DCs are removed than added (21.6% vs 12.3%). The same tendency is observed for under-specification and specification. This confirms Becher (2011a,b)'s qualitative findings that German is more explicit in terms of discourse relation marking.

Zufferey & Cartoni (2014) and Zufferey (2016) found that, based on the analysis of the translation of a subset of connectives, explicitation is not a general phenomenon. The roles of the source and target languages, the type of relations, and the specific DCs all have influences. We also see different patterns of explicitation depending on the translation directions and category of relations, e.g., CONTINGENCY relations are explicitated more often in English-to-German translations than

[1]The implicitation and explicitation proportions do not add up to 100%, because: 1) the proportions are normalized against the total connective counts of the each source/target language; and 2) overall, 58.0% of the connectives have been aligned to a connective of the same specificity level, and the specificity scores of 22.7% of the identified connectives or the aligned tokens is unknown (i.e. those tokens are not annotated in PDTB3.0 or PCC2.0).

EN →DE	EN original (171K tokens)				DE translation (164K tokens)			
	ttl. DC count	omission	under-specif.	impl. total	ttl. DC count	insertion	specification	expl. total
EXP	2329	13.1%	9.2%	22.4%	2821	**20.6%**	**3.1%**	**23.7%**
CONT	906	16.8%	6.8%	23.6%	1383	**33.0%**	**18.7%**	**51.8%**
COMP	978	7.5%	13.3%	20.8%	979	**24.9%**	**35.4%**	**60.4%**
TEMP	426	25.6%	13.8%	39.4%	505	**40.2%**	**16.6%**	**56.8%**
Total	4639	13.8%	10.0%	23.8%	5688	**26.1%**	**13.7%**	**39.8%**
DE →EN	DE original (95K tokens)				EN translation (107K)			
	ttl. DC count	omission	under-specif.	impl. total	ttl. DC count	insertion	specification	expl. total
EXP	1876	17.6%	3.0%	20.7%	1605	**13.8%**	**20.1%**	**33.9%**
CONT	1146	24.5%	16.8%	41.3%	831	10.5%	**7.8%**	18.3%
COMP	638	21.2%	32.1%	53.3%	673	**9.5%**	**15.9%**	**25.4%**
TEMP	391	32.7%	6.4%	39.1%	310	15.8%	**41.9%**	**57.7%**
Total	4051	21.6%	11.8%	33.4%	3419	12.3%	**18.3%**	**30.6%**

Table 1: Proportions of connectives that are not aligned to any words in the target text (*omission*) or the source text (*insertion*); and connectives that are aligned to a connective of higher relative entropy (rel. ent.) in the target text (*under-specification*) or the source text (*specification*). Bolded proportions refer to proportions of explicitation exceeding the proportions of implicitation of the same type in the opposite translation direction (compared against the sub-table in diagonal).

in the other direction.

Moreover, our analysis of connectives typically expressing all types of relation senses provides a more comprehensive picture. The results show that the explicitation strategy also differs across different relation senses and translation directions. For example, insertion seems to be more frequent than specification in German translations, except for COMPARISON relations, but the reverse is the case for English translations.

To find out whether these patterns can be explained by obligatory explicitations or translation-inherent explicitions, we look at the connectives that are most frequently omitted/inserted and (under-)specified. We found that connectives that are most frequently added in the translation, are also those that are most frequently omitted in the opposite translation direction, consistent with reports by Hoek et al. (2015) and supporting the findings of Becher (2011b) that most explicitations are obligatory due to the cross-lingual contrast of English and German.

Taking into account obligatory translation effects, we still find more explicitation in the translation than would have been expected (see bolded numbers in Table 1). In other words, the *Explicitation Hypothesis* is quantitatively confirmed for both explicitation strategies, translation directions and all categories of relations, save two exceptions: CONTINGENCY and TEMPORAL connectives are frequently dropped in English to German translation and they are not counter-balanced by the insertion in German to English translation. Our inspection of the results suggests that the high rate of these omissions could be attributed to the dropping of *when*, *if* and *so* in English to German translation. Previous work has found that CAUSAL DCs like *so* are often omitted due to processing ease (Hoek et al. 2017).

To summarize, results based on automatic cross-lingual DC annotation and an empirical ap-

proach to compare DC specificity reveal systematic patterns of implicitation and explicitation in English-German translation. We found evidence that explicitations counter-balance and exceed opposite implicitation.

5. Discussion and Conclusion. The current study investigated explicitation of discourse connectives in English-German parallel texts. To gain a comprehensive insight of the patterns underlying explicitation, we exploited an automatic approach to connective identification and alignment, which allowed us to study a large variety of connectives (173 English and 126 German connective types) and many samples per language (8058 English and 9739 German connectives were identified in our dataset).

We also propose a novel method of studying explicitation in translation, namely by considering the relative entropy of corresponding connectives in parallel text. Our results showed that the general pattern of explicitation in translation replicates to specification of connectives. The large-scale alignments provide additional insights, such as the fine-grained interaction between relation type and explicitation strategy across different languages. Such analyses would not have been possible without taking into account how all types of DCs are translated within the same span of text and a well-defined measure to identify cross-lingual specificity gap.

We conclude that the results from our empirical and automatic approach of identifying explicitation, both in terms of addition and specification of DCs support the *Explicitation Hypothesis* in translation between English and German. Future work will focus on applying a similar methodolgy to less studied language pairings to gain further insight into the generalizability of DC translation patterns.

References

Becher, Viktor. 2010. Towards a more rigorous treatment of the explicitation hypothesis in translation studies. *Trans-kom* 3(1). 1–25.

Becher, Viktor. 2011a. *Explicitation and implicitation in translation. a corpus-based study of english-german and german-english translations of business texts*: Staats-und Universitätsbibliothek Hamburg Carl von Ossietzky dissertation.

Becher, Viktor. 2011b. When and why do translators add connectives?: A corpus-based study. *Target. International Journal of Translation Studies* 23(1). 26–47.

Blum-Kulka, Sh. 1986. Shifts of cohesion and coherence in translation. *Interlingual and Intercultural Communication. Discourse and Cognition in Translation and Second Language Acquisition Studies* 17–35.

Bourgonje, Peter. 2021. *Shallow discourse parsing for german*, vol. 351. IOS Press.

Bourgonje, Peter & Manfred Stede. 2020. The potsdam commentary corpus 2.2: Extending annotations for shallow discourse parsing. In *Proceedings of the 12th language resources and evaluation conference*, 1061–1066.

Cartoni, Bruno & Thomas Meyer. 2012. Extracting directional and comparable corpora from a multilingual corpus for translation studies. In *Proceedings of the eighth international conference on language resources and evaluation (lrec'12)* CONF, 2132–2137.

Crible, Ludivine, Ágnes Abuczki, Nijolė Burkšaitienė, Péter Furkó, Anna Nedoluzhko, Sigita Rackevičienė, Giedrė Valūnaitė Oleškevičienė & Šárka Zikánová. 2019. Functions and trans-

lations of discourse markers in ted talks: A parallel corpus study of underspecification in five languages. *Journal of Pragmatics* 142. 139–155.

Dou, Zi-Yi & Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. *arXiv preprint arXiv:2101.08231* .

Dupont, Maïté & Sandrine Zufferey. 2017. Methodological issues in the use of directional parallel corpora: A case study of english and french concessive connectives. *International journal of corpus linguistics* 22(2). 270–297.

Hawkins, John A. 1986. *A comparative typology of english and german: Unifying the contrasts*. London Sydney: Croom Helm.

Hoek, Jet, Jacqueline Evers-Vermeul & Ted JM Sanders. 2015. The role of expectedness in the implicitation and explicitation of discourse relations. In *Proceedings of the second workshop on discourse in machine translation*, 41–46.

Hoek, Jet, Sandrine Zufferey, Jacqueline Evers-Vermeul & Ted JM Sanders. 2017. Cognitive complexity and the linguistic marking of coherence relations: A parallel corpus study. *Journal of pragmatics* 121. 113–131.

House, Juliane. 2014. *Translation quality assessment: Past and present*. Springer.

Klaudy, Kinga. 1998. Explicitation. *Routledge encyclopedia of translation studies* 80–84.

Klaudy, Kinga. 2009. The asymmetry hypothesis in translation research. *Translators and their readers. In Homage to Eugene A. Nida. Brussels: Les Editions du Hazard* 283. 303.

Knaebel, René. 2021. discopy: A neural system for shallow discourse parsing. In *Proceedings of the 2nd workshop on computational approaches to discourse*, 128–133.

Lapshinova-Koltunski, Ekaterina, Christina Pollkläsener & Heike Przybyl. 2022. Exploring explicitation and implicitation in parallel interpreting and translation corpora. *The Prague Bulletin of Mathematical Linguistics* (119). 5–22.

Marco, Josep. 2018. Connectives as indicators of explicitation in literary translation: A study based on a comparable and parallel corpus. *Target. International Journal of Translation Studies* 30(1). 87–111.

Zufferey, Sandrine. 2016. Discourse connectives across languages: Factors influencing their explicit or implicit translation. *Languages in Contrast. International Journal for Contrastive Linguistics* 16(2). 264–279.

Zufferey, Sandrine & Bruno Cartoni. 2014. A multifactorial analysis of explicitation in translation. *Target. International Journal of Translation Studies* 26(3). 361–384.