Investigating Explicitation of Discourse Connectives in Translation using Automatic Annotations

Frances Yung[1] **Merel C.J. Scholman**[1,2]
Ekaterina Lapshinova-Koltunski[3] **Christina Pollkläsener** [3] **Vera Demberg** [1]
[1]Saarland University, Saarbrücken, Germany
[2] Utrecht University, Utrecht, Netherlands
[3]University of Hildesheim, Hildesheim, Germany
{frances,m.c.j.scholman,vera}@coli.uni-saarland.de
{lapshinovakoltun, christina.pollklaesene}@uni-hildesheim.de

Abstract

Discourse relations have different patterns of marking across different languages. As a result, 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 discourse relations are indeed more explicit in translation than texts written originally in the same language. Automatic annotations allow us to carry out translation studies of discourse relations on a large scale. Our methodology using relative entropy to study the specificity of connectives also provides more fine-grained insights into translation patterns.

1 Introduction

Discourse connectives such as *because* and *however* are considered volatile items in translation: translators often add, rephrase or remove them (e.g. Zufferey and Cartoni, 2014). Prior studies have often 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 an understudied 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 and 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 the cases of (under)specifications instead of a subjectively defined subset.

Our contributions are: 1) We demonstrate that automatic word alignments and discourse parsers facilitate a comprehensive study of discourse connectives and relations in translation. 2) We show evidence for explicitation in translation, in terms of both insertion and specification of DCs; 3) We compare the cross-lingual specificity of English and German DCs; 4) The automatically aligned and annotated data are publicly available[1].

2 Background

2.1 Explicitation Hypothesis

Previous studies show that the translation of discourse connectives depends on various factors. One of the most well-known accounts, the Explicitation Hypothesis, suggests that translations tend to be

[1]https://osf.io/ybfxp/?view_only=8ef5f7a591064b7ea3334f706e544118

more explicit than the source texts (Blum-Kulka, 1986). However, this does not mean that discourse relations are always explicitated in translation, or that explicitation of the relations is always due to the translation effect. 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. This type of explicitation is separate from structural, formal or stylistic differences between the two languages, and with culture-specific textual elements. Klaudy (2009) argues that, in order to identify any translation-inherent explicitations, corresponding *implicitation in the opposite translation direction* should be taken into account. That is to say, 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 and Cartoni, 2014; Crible et al., 2019), the identity of the source and target languages (Zufferey, 2016), register and translator expertise (Dupont and 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 propose a score to quantify the specificity gap between a connective and its translation, such as cases where a stronger connective is used in translation (e.g. "*and*" translated as "*außerdem*" in German). While previous works only study a limited subset of subjectively defined specification, our empirical approach allows us to 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. This means that the entropy of English *and* might differ from the precise value of the entropy of German *und*. 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. 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 and 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 (see Section 3.3).

2.3 Identification and alignment of discourse connectives

Prior work is often based on a restricted selection of connectives. This can be attributed to the fact that connective identification on a large scale can be difficult, because many discourse connectives can also be used in non-connective contexts (e.g., *indeed* is not always used as a DC). Consequently, prior corpus studies have mostly focused on a handful of connectives and senses. For example, Zufferey and Cartoni (2014) analyzed 200 occurrences each of the English causal connectives *since*, *because* and *given that* in Europarl. The frequent causal connective *as* was excluded because it is often used in a non-connective usage. A more comprehensive analysis that takes into account a larger range of

connectives and coherence relation senses in the same text is critical to be able to get more insight into the general translation patterns of connectives. The current study explores the feasibility of using automatic methods to identify and align discourse connectives.

Automatic word alignment was an essential step in statistical machine translation (Och and Ney, 2000). In the era of neural machine translation, word alignment is often used for annotation projection, including the projection of English discourse annotations (Versley, 2010; Laali, 2017; Sluyter-Gäthje et al., 2020). The focus of these works is to associate discourse sense labels annotated for the DCs in English with the DCs in the human or machine-translated texts, in order to create discourse-annotated resources in the other languages. In contrast, we use word alignments to examine where the DC marking differs between source and target languages, when DCs are inserted, omitted or their specificity is changed.

Another line of work uses automatic word alignments to generate cross-lingual lexica of connectives. For example, Bourgonje et al. (2017) extract alignments between German and Italian adversative connectives that are identified based on connective lexicons of both languages. Özer et al. (2022) link the multilingual annotation of the TED-MDB corpus (Zeyrek et al., 2019) to induce multilingual connective lexicons. Robledo and Nazar (2023) examine the mapping of English and Spanish connectives in order to identify possible new categories of relation senses. In this work, we use a similar technique to investigate whether connectives are explicitated by insertion or specification. In contrast to existing work, we also use language-specific discourse parsers to identify connectives and exclude tokens of non-discourse usage in English and German texts. We then use a neural word aligner which has reported lower error rates compared with statistical aligners.

3 Methodology

3.1 Data

We analyze the parallel texts taken from the Europarl Direct Corpus (Cartoni and Meyer, 2012), which are proceedings from the European Parliament. A total of 33 proceedings are used in the analyses.[2] 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.

3.2 Identification and alignment of DCs in English and German texts

We use two language-specific parsers to identify and annotate the discourse relations in the English and German texts. We use the Discopy parser (Knaebel, 2021) to identify and classify DCs in the English original and translated texts. This parser considers the semantic representation of a connective token and its contexts. The classifier distinguishes discourse and non-discourse usage of the connective and labels each with a sense label based on the PDTB 2.0 framework (Prasad et al., 2008). The reported accuracies are 97.20% for connective identification, and 92.12% / 86.26% respectively for 4-way coarse-grained / 14-way fine-grained classification of the relation sense.

For the German texts, we use the German Shallow Discourse Parser (Bourgonje and Stede, 2018; Bourgonje, 2021) to identify and classify DCs in the German original and translated texts. The parser is based on a BERT architecture with additional syntactic features and ambiguity knowledge from the DimLex lexicon (Stede, 2002). It has been trained on the Potsdam Commentary Corpus (PCC) 2.2 (Bourgonje and Stede, 2020) to predict a sense labels defined in the PDTB 3.0 hierarchy (Webber et al., 2019). The reported results on the accuracy of this German parser regarding discourse-usage identification is 87.57% and 85.63% / 80.57% respectively for 4-way coarse-grained / 16-way fine-grained classification of the relation sense.

We align the identified connectives cross-lingually using the Awesome Align word alignment model (Dou and Neubig, 2021), which extracts corresponding tokens (including m:n mappings and "null" alignments) in a pair of bilingual sentences based on multilingual embeddings of the tokens and fine-tuned on parallel texts. An error rate of 15.1% is reported evaluating against human annotation of English-German word alignments (of all words, not just DCs), which out-performs statistical alignment models such as GIZA++ (Och and Ney, 2000) and eflomal (Östling and Tiedemann, 2016).

To ensure that the annotation tools produce reliable output for our data, we manually analyzed the

[2]These 33 proceedings are selected because they overlap with instances included in the discourse-annotated DiscoGeM corpus (Scholman et al., 2022), which could be used in future contrastive studies.

automatic annotations of 200 randomly extracted connective pairs each from the English-German and German-English translation data. The accuracy (precision) of connective identification and 4-way sense classification are 85% and 92% for English and 83% and 90% for German. The alignment accuracy is 90%. Taking into account error-propagation, in our analysis, we annotate DCs only on one side and analyze their alignment to the other side without considering whether the aligned words are also identified as DCs. In addition, we improve the automatic annotations by syntactic rules that remove unlikely DC candidates (e.g. *damit...,..zu..* is not a DC) and "unalign" tokens that cannot mark connectives, such as *'power'* or *'reading'*). We analyze the alignments of the source/target English and German texts respectively, in order to identify explicitation and implicitation in both translation directions.

3.3 Quantifying specificity of connectives

We determine the specificity level of each English and German connective based on their manual annotation in existing discourse-annotated resources. For English connectives, we extract the distribution of sense labels (after removing the *speechact* and *belief* tags) assigned to the *explicit* connectives in PDTB3.0. We extract the sense distribution of each German connective similarly based on their sense annotation in the PCC2.0 corpus (Bourgonje and Stede, 2020).

It is possible that the corpora from which we extract the specificity information differ in domain or aspects of how the annotation schemes were applied, such that in one language, a wider variety of relations was annotated than in the other. In order to remove such effects, 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. We further 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.

Connectives that are aligned to "null" in the target text are considered *omissions*, and connectives that are aligned to "null" in the source text are considered *insertions*. Similarly, connectives in the source and target texts that are aligned to a *less specific* connective are identified as *under-specification* and *specification* respectively.

4 Results

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

4.1 Implicitation & explicitation of DCs

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. Alignments of connectives in the **source** texts to *"null"* or a higher entropy DC means **omission** and **under-specification**, while the corresponding alignments of connectives in the **target** texts would mean **insertion** and **specification** in translation.[3]

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 the previous qualitative conclusion that German is more explicit in terms of discourse relation marking (Becher, 2011b,a).

Zufferey and 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 than in German.

Moreover, our analysis of connectives typically expressing all types of relation senses provides a

[3]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	align to 'null' (omission)	align to a DC of higher rel. ent. (under-specif.)	impl. total	ttl. DC count	align to 'null' (insertion)	align to a DC of higher rel. ent. (specification)	expl. total
EXPANSION	2329	13.1%	9.2%	22.4%	2821	**20.6%**	**3.1%**	**23.7%**
CONTINGENCY	906	16.8%	6.8%	23.6%	1383	**33.0%**	**18.7%**	**51.8%**
COMPARISON	978	7.5%	13.3%	20.8%	979	**24.9%**	**35.4%**	**60.4%**
TEMPORAL	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	align to 'null' (omission)	align to a DC of higher rel. ent. (under-specif.)	impl. total	ttl. DC count	align to 'null' (insertion)	align to a DC of higher rel. ent. (specification)	expl. total
EXPANSION	1876	17.6%	3.0%	20.7%	1605	**13.8%**	**20.1%**	**33.9%**
CONTINGENCY	1146	24.5%	16.8%	41.3%	831	10.5%	**7.8%**	18.3%
COMPARISON	638	21.2%	32.1%	53.3%	673	**9.5%**	**15.9%**	**25.4%**
TEMPORAL	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*). *Impl.* and *expl.* totals are based on the sum of *omission/insertion* and *under-specification/specification* respectively. 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).

Implicitation	**Explicitation**
EN→DE omission	EN→DE insertion
and (177), also (69), when (62), if (49), but (43), so (41)	und (287), dann (121), wenn (88), also (61) damit (57), aber (52)
DE→EN omission	DE→EN insertion
und (105), dann (105), aber (78), sondern (68), wenn (52), deshalb (49)	and (158), also (26), but (26), if (25) when (25), so (13)
EN→DE under-specif.	EN→DE specification
also → auch (173), but → sondern (113), then → dann (54), because → da (22), so that → damit (16)	but→ jedoch (89), however→ jedoch (82), but→ doch (70), when→ wenn (67), although→ obwohl (26)
DE→EN under-specif.	DE→EN specification
aber → however (80), wenn → when (67), jedoch → however (32), denn → for (30), allerdings → however (12)	auch → also (281), dann→ then (126), sondern - but (86), damit → so that (25), sondern → rather (13)

Table 2: The most freqent connective omissions, insertions, under-specifications and specifications (counts in brackets) in both translation directions.

more comprehensive picture. The results show that the explicitation strategy also differs across different relation senses and translation directions. For example, relations are explicitated more by insertion, while more relations in German translation, in particular temporal relations, are explicitated by specification in English. Within German translation, many CONTINGENCY (33.0%) and TEMPORAL connectives (40.2%) are inserted, while comparatively, COMPARISON relations are explicitated more by specification (35.4%).

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, see Table 2. It can be seen 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 expected (see bolded figures 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. Table 2 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).

In addition, many of the explicitated COMPARISON relations come from the translation of *but* and *however*, which are ambiguous because they can signal both CONTRAST and CONCESSION relations. The German translation often specifically signals CONCESSION, such as *jedoch* and *allerdings*. We will analyze some of these cases in Section 5 to see if such explicitation is obligatory or translation-inherent.

4.2 Cross-lingual correspondence of DCs

Next, we look into the mutual correspondence between English and German connectives. Figure 1 shows the normalized distribution of the alignment between each source connective (x-axis) and their translation (y-axis; at least the top two most common translations are displayed). Higher numbers / darker colors represent more frequent translation alignments.

It can be observed that some connectives have one or two dominating translations (e.g. English: *also, and, if, then*; German: *auch, und, weil*), while others can have an even distribution of various translations (e.g. English: *so, but*; German: *deshalb*). While many of the correspondences in the two translation directions are asymmetrical (e.g. 82% of *auch* is translated to *also*, but only 45% of *also* is translated to *auch*), some correspondences are symmetrical, indicating that the pair of connectives are of mutual correspondence (e.g. *and* is frequently translated as *und* and vice versa; the same goes for *then* and *dann*).

Figure 1 also suggests a general trend that English connectives are translated to a wider range of German connectives, while German connectives more often have one dominating English translation (more darker color cells in the bottom figure). It is to be expected that English connectives are more ambiguous than German, as English is less explicit in terms of discourse markedness (House, 1997; Becher, 2011a). We quantify this observation by considering the *cross-lingual specificity* of English and German connectives based on the diversity of their translations. This is calculated as the entropy of the distribution of alignments of each unique connective in the source texts (i.e. the entropy of the distribution per column in Figure 1). Figure 2 shows the distribution of connectives grouped by the entropy of their translation alignments. Connectives with less than 20 occurrences

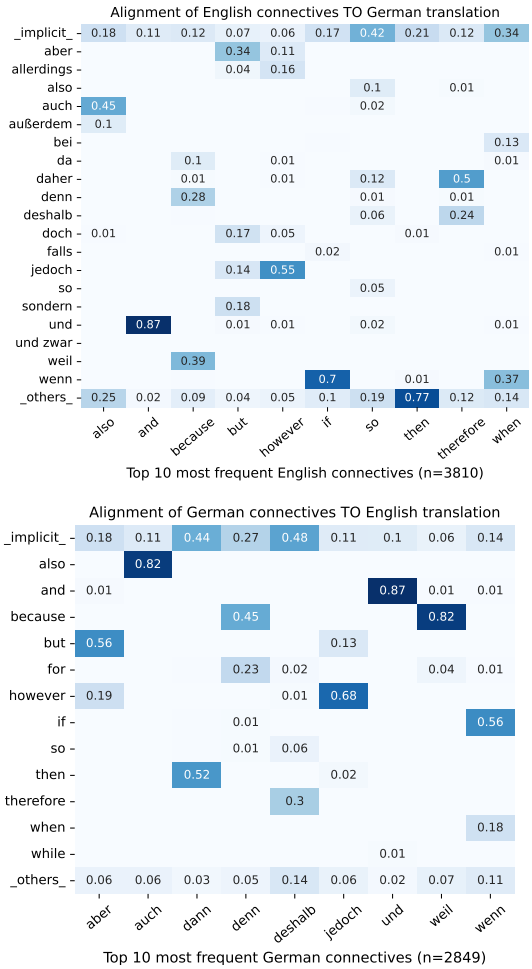

Figure 1: Alignment between connectives in the source texts (x-axis) with their corresponding tokens in translation (y-axis); the first row *_implicit_* means the connective is not aligned to any words in the target sentence, and the last row _others_ refers to the proportions of alignments to tokens that are not displayed on the y-axis.

are not included since the alignment distributions may divert from the actual distribution due to their sample size. It can be seen that most English DCs are more versatile and correspond to a wide range of German DCs, while a normal distribution is observed for the German DCs: some DCs have more correspondences and some have less.

To summarize, the automatic connective annotation and alignment procedure allows us to extract the complex mapping between connectives empirically and instantly. This enables us to identify systematic patterns such as the overall specificity of English connectives in terms of English-German translation. We found empirical evidence that explicitations counter-balance and exceed opposite implicitation.

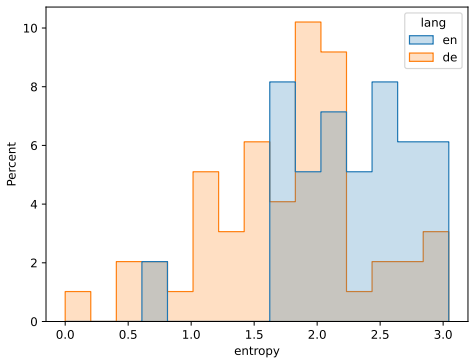

Figure 2: Distribution of connectives grouped by the entropy of their translation alignment.

We however also note that one needs to consider the effect of possible annotation errors using such an automatic approach. Based on our manual inspection of the 400 alignments, most of the error comes from the over-identification[4] of English *and* and German *und*: these often did not function as connectives, but were identified as such by the parsers. In most of these cases, *and* and *und* were aligned, which means that they were not counted as explicitation nor implicitation. Consequently, our reported explicitation / implicitation rate of EXPANSION could actually be higher, because the sample size should be smaller. Regarding errors specific to the alignment of connectives, we found that most alignment errors were false positives (i.e. a connective was aligned to a non-connective word, when in fact it was supposed to align to *null*), meaning the insertion / omission rates could actually be higher.[5] Therefore, manual qualitative analysis is still necessary to confirm the findings. This will also be demonstrated in the next section.

5 Qualitative analysis

The qualitative results show that there are more explicitations in translation after counter-balancing implicitation in the other translation direction. Now the question is, are these explicitations actually coming from the nature of the translation process, or are they due to the contrast between the two languages or other reasons? We try to gain some insights through a qualitative analysis.

[4]Note that the manual inspection did not include cases where a connective was missed by the parsers.

[5]The relative entropy of the falsely aligned words would most likely be "unknown", so they are not counted as (under-)specification.

We analyze the alignment instances to see if the explicitated translations are **obligatory** or **translation-inherent** (see Sec. 2.1). This analysis revealed various cases of obligatory explicitation. First, Table 1 shows that TEMPORAL relations are often specified in German to English translation. Table 2 suggests that the high explicitation rate of German TEMPORALs can be attributed to the frequent specification of *dann* (which can signal both TEMPORAL and CONDITIONAL according to PCC2.0) to *then* (which dominantly signals TEMPORAL in PDTB3.0). These explicitations are likely to belong to obligatory explicitations, because *then* is the only English DC that signals a PRECEDENCE relation like *dann* does, and has a similar level of markedness.

Second, for German translation, Table 1 also reveals that COMPARISON relations are often specified. The high specification rate of English COMPARISONs comes from the frequent translation of *but* to *jedoch* or *doch*, and *however* to *jedoch*, as seen in Table 2. The translation of *however* to *jedoch* might also be categorized as obligatory explicitation. The two connectives are very similar in their meaning and usage (both are predominantly be used to mark CONTRAST and CONCESSION), but English *however* is also occasionally used to mark SYNCHRONOUS relations among its many annotations in PDTB3.0 – this sense did not occur for *jedoch* in the PCC2.0. Similarly, the frequent specification of *wenn* to *when* belongs to this case. *Wenn*, which can ambiguously signal a CONDITION or SYNCHRONOUS relation, often has to be translated to the less specific *when* to mark a SYNCHRONOUS relation naturally because of a lack of other suitable DCs in English.

The translation of *but* to *doch/jedoch* differs from the previously discussed obligatory explicitations and might actually be translation-inherent: translators could have translated *but* to *aber*, which matches *but* semantically and also in terms of strength and specificity, instead of specifying the relation with *jedoch* or *doch*. To gain further insight into the reason for these explicitations, a trained translator manually analyzed these cases using a "substitution test": we produced an alternative translation using *aber*, making necessary grammatical changes. If the resulting translation is equally acceptable, then it could be a case of translation-inherent explicitation.

We found that in 35% of the *but*-instances that

were translated into *doch/jedoch*, these more specific could have been chosen because the resulting syntactic or stylistic structure is preferred; that is, they do actually appear to be cases of obligatory explicitation. For example:

> It is important to have EU and national targets, **but** it is also important to have a European directive...
> Es ist zwar wichtig, Ziele auf EU- und einzelstaatlicher Ebene zu setzen, **doch** ist es ebenso wichtig, eine europäische Richtlinie zu schaffen...

In this case, having chosen *zwar* in the previous clause, the translator likely used *doch*, because they often occur together. But in 65% of the cases, the use of *aber* is equally acceptable, and thus these cases appear to represent translation-inherent explicitation. For example:

> Its starting point is the European Year Against Racism 1997 **but** the context has moved on significantly.
> Ausgangspunkt war das Europäische Jahr gegen Rassismus 1997, **doch/aber** der Kontext wurde seither beträchtlich weiterentwickelt.

Among these acceptable cases, in 38% of the total cases, *doch* or *jedoch* sometimes fit better to the formality of a parliament discussion, while *but* is a lighter DC typical in spoken English, for example:

> **But** as has been pointed out, the adoption of a rigorous definition of the precautionary principle is crucial.
> ***Doch**, wie bereits festgestellt wurde, ist dabei die Verabschiedung einer strikten Definition des Vorbeugeprinzips von entscheidender Bedeutung.*

One possible explanation is the *domain gap* between the source and target texts. The source texts of the Europarl corpus are prepared speeches of the parliament, while the target texts are the published translation of these scripts. In other words, the source texts are prepared to be spoken while the target texts are for reading. This could be a reason that the discourse relations in the translated texts of the Europarl corpus are more specified than the original texts, corresponding to the *situational* and *translation-task* variables as discussed in House (2004). Analysis on data from another genre could confirm this domain and genre effect.

6 Discussion

The current study investigated explicitation and implicitation 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 evaluated the feasibility of this approach by first studying whether we could replicate the established effect of explicitation in translation between English and German texts. We furthermore extended existing findings by defining explicitation in a more fine-grained sense as specification of the relation sense, and investigating whether we can see a similar pattern of explicitation of connectives for those connectives that were already explicit in the source text.

Our quantitative results provide strong evidence for the *Explicitation Hypothesis*: taking into account the counter-balance of implicitation in the opposite translation direction, there is still considerable more explicitation in translation. Manual qualitative analysis suggests that a domain effect may have played a role. These findings are in line with already established effects in prior work, and thus support the reliability of the insights that the automatic approach can provide.

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. Furthermore, we found that English connectives are generally less specific than German ones, considering all types of connectives and their translation in our data. 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 discourse relations indeed tend to be explicitated in translation. Our proposed automatic approach is feasible for studying translation of connectives in parallel text. We were able to replicate known effects for German-English translations and extend these findings to specification of connectives using relative entropy. The cross-lingual analysis in large scale allows us to identify

language-specific patterns in discourse production, which is useful for the generation of multi-lingual discourses. Future work will focus on applying a similar methodolgy to less studied language pairings to gain further insight into the generalizability of DC translation and production patterns.

Acknowledgements

This project is supported by the German Research Foundation (DFG) under Grant SFB 1102 ("Information Density and Linguistic Encoding", Project-ID 232722074).

References

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

Viktor Becher. 2011a. *Explicitation and implicitation in translation. A corpus-based study of English-German and German-English translations of business texts*. Ph.D. thesis, Staats-und Universitätsbibliothek Hamburg Carl von Ossietzky.

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

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

Peter Bourgonje. 2021. *Shallow discourse parsing for German*, volume 351. IOS Press.

Peter Bourgonje, Yulia Grishina, and Manfred Stede. 2017. Toward a bilingual lexical database on connectives: Exploiting a german/italian parallel corpus.

Peter Bourgonje and Manfred Stede. 2018. Identifying explicit discourse connectives in german. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 327–331.

Peter Bourgonje and 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*, pages 1061–1066.

Bruno Cartoni and 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, pages 2132–2137.

Ludivine Crible, Ágnes Abuczki, Nijolė Burkšaitienė, Péter Furkó, Anna Nedoluzhko, Sigita Rackevičienė, Giedrė Valūnaitė Oleškevičienė, and Šárka Zikánová. 2019. Functions and translations of discourse markers in ted talks: A parallel corpus study of underspecification in five languages. *Journal of Pragmatics*, 142:139–155.

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

Maïté Dupont and 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.

John A Hawkins. 1986. *A comparative typology of English and German: Unifying the contrasts*. London Sydney: Croom Helm.

Jet Hoek, Jacqueline Evers-Vermeul, and 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*, pages 41–46.

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

Juliane House. 1997. *Translation quality assessment: A model revisited*. Gunter Narr Verlag.

Juliane House. 2004. *Explicitness in discourse across languages*. Bochum: AKS.

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

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

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

René Knaebel. 2021. discopy: A neural system for shallow discourse parsing. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 128–133.

Majid Laali. 2017. *Inducing discourse resources using annotation projection*. Ph.D. thesis, Concordia University.

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

Josep Marco. 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.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th annual meeting of the association for computational linguistics*, pages 440–447.

Robert Östling and Jörg Tiedemann. 2016. Efficient word alignment with Markov Chain Monte Carlo. *Prague Bulletin of Mathematical Linguistics*, 106:125–146.

Sibel Özer, Murathan Kurfalı, Deniz Zeyrek, Amália Mendes, and Giedrė Valūnaitė Oleškevičienė. 2022. Linking discourse-level information and the induction of bilingual discourse connective lexicons. *Semantic Web*, (Preprint):1–22.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. 2008. The penn discourse treebank 2.0. In *LREC*.

Hernán Robledo and Rogelio Nazar. 2023. A proposal for the inductive categorisation of parenthetical discourse markers in spanish using parallel corpora. *International Journal of Corpus Linguistics*.

Merel Scholman, Dong Tianai, Frances Yung, and Vera Demberg. 2022. Discogem: A crowdsourced corpus of genre-mixed implicit discourse relations. In *the 13th Language Resources and Evaluation Conference (LREC 2022)*, pages 3281–3290. European Language Resources Association.

Henny Sluyter-Gäthje, Peter Bourgonje, and Manfred Stede. 2020. Shallow discourse parsing for under-resourced languages: Combining machine translation and annotation projection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1044–1050.

Manfred Stede. 2002. Dimlex: A lexical approach to discourse markers. *A. Lenci and*, 501:1–15.

Yannick Versley. 2010. Discovery of ambiguous and unambiguous discourse connectives via annotation projection. In *Proceedings of Workshop on Annotation and Exploitation of Parallel Corpora (AEPC)*, pages 83–82.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The penn discourse treebank 3.0 annotation manual. *Philadelphia, University of Pennsylvania*, 35:108.

Deniz Zeyrek, Amalia Mendes, Yulia Grishina, Murathan Kurfali, Samuel Gibbon, and Maciej Ogrodniczuk. 2019. Ted multilingual discourse bank (ted-mdb): a parallel corpus annotated in the pdtb style. *Language Resources and Evaluation*, pages 1–38.

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

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