Manuscript Details

Manuscript number	PRAGMA_2018_376
Title	Functions and translations of underspecified discourse markers in TED Talks: a parallel corpus study on five languages
Article type	Full length article

Abstract

Discourse markers are highly polyfunctional, particularly in spoken settings, and this variation in meaning can be challenging to structure. In addition, their syntactic optionality makes them optimal candidates for omission in translations, even more so in the restricted space of subtitles, as is the case with parallel transcripts of TED Talks. In this study, we combine the methods of discourse annotation and translation studies to investigate the functions, translation equivalents and omissions of the three most frequent English discourse markers and, but, so and their translation into Czech, French, Hungarian and Lithuanian. In particular, we study them through the lens of underspecification, of which we distinguish several types. We have observed that some processes of underspecification are common to the languages in our sample, that they proceed in parallel, based on the semantics of discourse markers. However, not all discourse marker types nor their functions are equally affected by underspecification. Besides, monolingual and multilingual underspecification do not always map for a particular marker.

Keywords	discourse markers; translation; underspecification; sense annotation; TED Talks
Corresponding Author	Ludivine Crible
Corresponding Author's Institution	Université catholique de Louvain
Order of Authors	Ludivine Crible, Agnes Abuczki, Nijole Burkšaitienė, Giedre V. Oleškevičienė, Šárka Zikánová
Suggested reviewers	Wilbert Spooren, Ekaterina Lapshinova-Koltunski

Submission Files Included in this PDF

File Name [File Type]

Cover letter.docx [Cover Letter]

Highlights.docx [Highlights]

Title page.docx [Title Page (with Author Details)]

JoP article_subm_anonym.docx [Manuscript (without Author Details)]

Vita.docx [Author Biography]

To view all the submission files, including those not included in the PDF, click on the manuscript title on your EVISE Homepage, then click 'Download zip file'.

Nothing to declare.

Highlights:

- Applying discourse annotation to translation corpora reveals three types of underspecification that affect discourse markers.
- The frequency of omissions and the variety of translation equivalents vary with the function of the original marker across languages.
- The pragmatic spectrum of *and*, *but*, *so* and their equivalents in different languages can be structured across a limited number of functions and domains in a comparable way.

Functions and translations of underspecified discourse markers in TED Talks: a parallel corpus study on five languages

Ludivine Crible

Université catholique de Louvain

1 place Blaise Pascal, 1348, Louvain-la-Neuve, Belgium

ludivine.crible@uclouvain.be

Ágnes Abuczki

MTA-DE Research Group for Theoretical Linguistics

H-4010 Debrecen Pf. 47, Hungary

abuczki.agnes@gmail.com

Nijolė Burkšaitienė

Mykolas Romeris University, Institute of Humanities

Ateities 20, LT-08303, Vilnius, Lithuania

burksa@gmail.com

Giedrė Valūnaitė Oleškevičienė

Mykolas Romeris University, Institute of Humanities

Ateities 20, LT-08303, Vilnius, Lithuania

gentrygiedre@gmail.com

Šárka Zikánová

Charles University, Institute of Formal and Applied Linguistics

Malostranské náměstí 25, 118 00 Prague, Czech Republic

zikanova@ufal.mff.cuni.cz

Functions and translations of underspecified discourse markers in TED Talks: a parallel corpus study on five languages

Abstract

Discourse markers are highly polyfunctional, particularly in spoken settings, and this variation in meaning can be challenging to structure. In addition, their syntactic optionality makes them optimal candidates for omission in translations, even more so in the restricted space of subtitles, as is the case with parallel transcripts of TED Talks. In this study, we combine the methods of discourse annotation and translation studies to investigate the functions, translation equivalents and omissions of the three most frequent English discourse markers *and*, *but*, *so* and their translation into Czech, French, Hungarian and Lithuanian. In particular, we study them through the lens of underspecification, of which we distinguish several types. We have observed that some processes of underspecification are common to the languages in our sample, that they proceed in parallel, based on the semantics of discourse markers. However, not all discourse marker types nor their functions are equally affected by underspecification. Besides, monolingual and multilingual underspecification do not always map for a particular marker.

Key-words: discourse markers; translation; underspecification; sense annotation; TED Talks

1. Introduction

Discourse markers have been a primary topic of interest in the field of pragmatics and discourse analysis for several decades (e.g. Erman 2001, Fraser 1990, Fraser 1999, Fuller 2003, Redeker 1990, Schiffrin 1987, Van Dijk 1979). These procedural expressions, which contribute to discourse structure and coherence, have never ceased to intrigue researchers investigating their many forms and functions in both speech and writing. The interest in discourse markers (and their challenge) stems from their great semantic-pragmatic variation: the same form, e.g. *and* or *so*, can be interpreted in several ways in different contexts of use, which depends not only on the semantics of the marker, but also on its structural configuration and the language user's world knowledge. Different authors tend to refer to this phenomenon as polyfunctionality, ambiguity, polysemy or, as we will be exploring in this study, underspecification. These notions can sometimes be vague and seem to cover multiple mechanisms of meaning variation. This study sets out to explore how underspecification applies to discourse markers through the methodological lens of translation and discourse annotation.

Parallel corpora are an invaluable resource for pragmatic approaches to discourse markers, since they allow detecting nuances in interpretation and cross-linguistic differences which are not only useful for the comparison of different language systems, but are also relevant to further understand a category or phenomenon from a monolingual point of view. Unfortunately, such multilingual aligned resources are very scarce when it comes to spoken or conversational data: most parallel corpora are written texts that usually belong to rather formal discourse genres (newspapers, academic articles, religious texts) and therefore do not allow accessing particular uses of discourse markers that are more specific to speech. While there is a common core of forms and functions that discourse markers display across the two modalities (speech and writing), it has been shown that their semantic-pragmatic variation is larger in the former, where they perform additional functions related to the management of the interaction (Biber 2006, Castellà 2006, Fox Tree 2014, Kunz & Lapshinova-Koltunski 2015, [Author]). For this reason, the recent database of TED Talks (broadcast presentations on specific topics usually given in English, with subtitles translated into a large number of languages) allows for

innovative analyses of discourse markers in spoken(-like) parallel data (Dupont & Zufferey 2017, Steele 2015).

In this paper, we report on the methods and findings of a multilingual corpus study focusing on the functions of English discourse markers and their translations into Czech, French, Hungarian and Lithuanian, in a selection of TED Talks. We address the following theoretical and empirical research questions: (1) What can parallel corpora and discourse annotation tell us about different types of underspecification, how to detect and measure them? (2) In what way(s) are discourse markers underspecified? How frequent is this phenomenon in spoken-like data? 3) Are there any typical processes of underspecification that occur in parallel across the analysed languages? Can any general cross-linguistic features be observed in the process of underspecification? We start with the analysis of omissions, or implicitations, in discourse marker translations, which leads us to a detailed analysis of the three most frequent devices in the sample, i.e. *and*, *but*, *so*, their functions in the English original and their translations into four target languages. In doing so, we not only provide a fine-grained semantic-pragmatic description of these high-frequency discourse markers, but also illustrate the potential of TED Talks as a resource for cross-linguistic and translation research in spoken-like data.

Firstly, we outline approaches to the polyfunctionality of discourse markers, including those used in previous translations studies. Secondly, we describe our data and the annotation scheme. Finally, we present the analysis of omissions and functions of discourse markers in the sample, linking up our corpus-based findings to a more theoretical discussion of the notion of monolingual and multilingual underspecification.

2. Theoretical background

2.1 Discourse markers: categorisation and polyfunctionality

Discourse markers are defined by Schiffrin (1987: 31) as "sequentially dependent elements which bracket units of talk". They consist in lexical elements used to join clauses, utterances, paragraphs or ideas through coherence relations such as contrast, consequence or exemplification, among many others. In speech, they also perform additional functions related to the management of turns, topics or the speaker-hearer relationship (Fischer 2006, [Author]). Discourse markers are grammatically heterogeneous and can originate from the syntactic classes of coordinating conjunctions (*and, but, or*), subordinating conjunctions (*because, although*), adverbs (*well, actually*), verbal phrases (*you know, I mean*) or prepositional phrases (*in fact*). Detailed criteria for discourse marker identification will be provided in Section 3.2.

This formal heterogeneity is only rivaled by the great polyfunctionality of discourse markers, both as a category and as individual members. For instance, according to Schiffrin (1987), the most frequent English DM *and* has little semantic meaning and two basic discourse uses, including coordination and continuation. Besides, *and* also has contrastive uses, e.g. *We tried to win. And we lost*. It can also preface the outcome of a reason, connect two pieces of support at a higher level of idea structure or a general conclusion drawn from a list of specific events. Working on native and learner speech, Buysse (2012) classified *so* across three domains (factual, argumentative, textual) and 10 functions (e.g. result, conclusion, summary, self-correction), concluding with a highly polysemous discourse marker, which the author relates to possible diachronic change. Aijmer (2002) discusses variable meanings of a selection of discourse markers, including *actually*, *now* or *sort of* and identifies a number of different functions for each of them (e.g. *actually* can express seven functions, including contrast, elaboration or topic change).

Some authors have proposed to structure this functional variation across three, sometimes four "domains" or dimensions of discourse structure. Beyond a basic distinction between "semantic" (or objective, external) and "pragmatic" (or subjective, internal) uses (cf. Halliday & Hasan 1976, Sanders et al. 1992, Van Dijk 1979), Redeker (1990) proposed a refined tripartite distinction between ideational structure (connecting real-world events), rhetorical structure (connecting assumptions or speech acts) and sequential structure (connecting larger units of discourse such as turns or topics). This model has inspired several proposals (Cuenca 2013, Degand 1996) and some of them identified a fourth domain dedicated to the interpersonal dimension of speech (González 2005, [Author]). What is common to all these approaches is that they distinguish between a particular type of meaning or discourse relation which is expressed by a marker on the one hand, and the layer of discourse structure or the speaker intention which is targeted by this marker, on the other hand. In other words, as Buysse (2012) and others have showed, a discourse marker (e.g. *so*) can express a single meaning (e.g. result) across more than one domain (e.g. factual result, argumentative result, textual result). The approach adopted in this study is in line with such proposals and will be detailed in Section 3.2.

Such polyfunctionality poses a challenge for cross-linguistic comparisons of discourse markers, especially for translators who have to adapt them to a new language and culture, in which textual strategies involving their use are often different from those of the source text (Zufferey & Degand 2013).

2.2 Discourse markers in translation studies

Freely available parallel data resources are limited, especially with regard to the variety of languages and text tpes. One example is Europarl (Koehn, 2005), which consists of the translations of the proceedings of the European Parliament into most European languages (at most 50 million words). However, it only covers the very specific domain of parliamentary proceedings. For other, more conversational types of settings, only comparable corpora are available, such as the GECCo corpus for English-German (Kunz & Lapshinova-Koltunski 2015), the [anonymized] dataset for English-French [Author] or those presented in the studies by Aijmer & Simon-Vandenbergen (2006).

There remains, however, a vast amount of literature on discourse markers and discourse relations in parallel translated data (e.g. Aijmer 2007; Altenberg 1999, 2007; Cuenca 2008; Degand 2004; Mortier & Degand 2009). We focus on Hoek et al. (2017) and Dupont & Zufferey (2017), two recent studies which are particularly relevant to the concept of underspecification. In the former, the authors investigate the types of discourse relations which tend to be implicitated in translation. Their hypothesis is that cognitively simple relations are expected and should therefore more often occur implicitly than more complex relations. Their parallel corpus study on English parliamentary debates translated into Dutch, German, French and Spanish showed that some relations types are more prone to implicitation, including speech-act relations and positive causal relations. This comprehensive study on the factors of implicitation did not however include the most frequent English discourse marker *and*, since "this connective is extremely general and used to signal many different kinds of relation, often as an underspecified marker" (2017: 121), a gap which the present study intends to fill.

Dupont & Zufferey (2017) resort to translation corpora to investigate the effect of register, translation direction and translator's expertise on meaning shifts and omissions of English and French markers of concession. Their study is particularly relevant to our analysis since one of their corpora under investigation is the TED Talks corpus (Cettolo et al. 2012), of which they

discuss the specific characteristics and methodological issues. The authors mention in particular: subtitling as a specific type of translation (cf. Lefer & Grabar 2015); the hybrid mix of spoken and written features in TED Talks; the variety of TED Talks speakers (non-native speakers or speakers of various regional varieties of English); the low level of expertise of amateur translators. They found that concessive markers usually have one or two most frequent translation equivalents in TED Talks, whereas in news articles they are more varied, often implicit or "under-specified" (e.g. *néanmoins* 'nevertheless" translated by the more generic *but*). Dupont & Zufferey suggest that translations might be more faithful to the original in TED Talks, but remain careful because of the "noise arising from the specific translation features of the TED corpus" (2017: 284). Still, omissions are the least frequent in this corpus compared to the other two genres investigated, which is a surprising result given the space restrictions of subtitles, which they explained by the higher necessity to "maintain highly explicit links [in] argumentative language" (2017: 286).

2.3 Discourse markers and underspecification

The notion of underspecification is used primarily in three disciplines. Firstly, in formal linguistics, it refers to the decision to assign an underspecified meaning to ambiguous words in order to avoid making unnecessary disambiguation steps. In this sense, as used by Egg (2010), Egg & Redeker (2007) or Irmer (2011), underspecification is a methodological bias which can be used in formalisms to enhance reliability. Secondly, in psycholinguistics, underspecification refers to a type of processing when mental representation is not as fully developed as what the linguistic material suggests. Frisson (2009) and Frisson & Pickering (2001) report on findings that suggest that comprehenders do not always associate a word with its specific and detailed meaning but rather "stop" at an underspecified sense, which Ferreira et al. (2002) termed "good-enough representations".

Thirdly, in discourse analysis, underspecification has mostly been applied to discourse markers by Spooren (1997), in a corpus study where he examines their acquisition and functional use by children first-language learners and adult second-language learners of spoken Dutch. He defines underspecification as a mismatch between the semantics of the marker and that of its interpretation (i.e. the discourse relation which it expresses) and relates it to Horn's (1984) R- and Q-principles, in other words to speaker and hearer economy, respectively. Spooren further explains that an underspecified interpretation of a discourse marker is due to conversational implicatures and gives the example of *and* expressing a causal or contrastive relation, in which case the added meaning is superimposed over the basic meaning (addition for *and*). His results show that underspecification is less frequently used by more proficient speakers, as if more seasoned speakers "take into consideration the needs of the conversational partner" (1997: 165).

The present study aims at exploring underspecified discourse markers in Spooren's (1997) sens, not only with monolingual but also with multilingual data. In doing so, we expect to identify more than one type of underspecification: besides the semantic mismatch discussed above, underspecification also has tight links to translation, as a strategy to translate a specific marker by a more polyfunctional one. This second type was already discussed in Dupont & Zufferey (2017: 281), for instance when the French discourse marker *néanmoins* 'nevertheless' is translated by *but*. As it has already been mentioned above, the authors found that such underspecification in the translation is particularly frequent in news articles, and not so much in parliamentary debates nor in TED Talks. Thirdly, based on Baker's (2011) observation that languages tremendously vary in terms of types of conjunctions as well as their frequency, which poses a real challenge for translators, at times leading to the choice of omission,

implicitation (omission) of discourse markers could be considered as a third type of underspecification.

In sum, underspecification appears to be a multi-faceted notion which always corresponds to a mismatch or imbalance between a particular form and its contextual function. Translation data and corpus annotation is expected to shed more light on this pervasive phenomenon of (spoken) language use.

3. Methodology

3.1 Data

For this study, we extracted three texts from the multilingual parallel collection of TED Talks in English (original language), Czech, French, Hungarian and Lithuanian (target languages). This choice was made on the grounds that parallel texts are considered to be ideal for optimal comparability between languages (Kenning 2010) as they provide more flexible and accurate ways to compare discourse markers (Cettolo et al. 2012, Samy & González-Ledesma 2008).

TED Talks videos are available through the TED website together with subtitles provided in many languages. The talks are translated by volunteers into multiple languages. The open nature of the parallel collection of the talks makes them attractive to research as it ensures such advantages as: the set of subtitles continuously increases, subtitles are available in a substantial number of languages and the topics cover a wide span of knowledge fields, which makes the data applicable in multiple domains (Cettolo at al. 2012).

However, there are also certain limitations concerning the use of parallel transcripts of TED Talks for discourse marker research. The collection of TED Talks is unidirectional, thus it cannot be used for exemplifying differences for different translation directions. Also, volunteer translators do not necessarily ensure a high-quality translation. Finally, it should be kept in mind that even the original (English) scripts of TED Talks are semi-prepared, thus they do not provide genuine spoken linguistic data (cf. Dupont and Zufferey 2017).

Our data consists of three TED Talks comprising 234 sentences and 4720 words in the original English. The translations were manually aligned at sentence level.

3.2 Annotation method

The investigation was conducted in two stages: identification of discourse marker tokens in the original transcriptions and their translation (or lack thereof) in the target languages, followed by functional disambiguation of the English tokens of *and*, *but* and *so*.

The first identification stage was carried out manually and all expressions that met the following criteria were extracted: syntactically optional (not integrated in any syntactic relation), formally fixed (grammaticalized), with a procedural meaning and a discourse-structuring function. This identification was therefore bottom-up as we did not start from a list of pre-selected markers. Some borderline expressions which we decided not to select include "Here is the example" (a whole phrase with a propositional meaning), "if" used as a complementizer (as in "imagine if…") or "maybe" (used as a modal epistemic marker, not working on discourse structure). The five most frequent discourse markers extracted from the sample in the five languages under scrutiny are reported with their frequency in Table 1. The full list of all 41 discourse marker types in English is the following (ranked by order of frequency):

and, but, so, now, because, when, if, then, actually, okay, whereas, well, once, instead, even, also, in fact, equally, as, while, whenever, until, the thing is, so that, or, not only ... but, not least, in particular, yet, i. e., I think, I mean, for one thing, for instance, for example, even though, even if, either, before, and then, after all.

Table 1: Top-five most frequent discourse markers in each language and their frequency

	1st	**2nd**	**3rd**	**4th**	**5th**
EN	and (48)	but (35)	so (28)	now (26)	because (18)
CZ	a *'and'* (31)	ale *'but'* (22)	když *'when, if'* (17)	protože *'because'* (9)	tedy *'thus'* (8)
FR	et *'and'* (25)	mais *'but'* (25)	si *'if'* (12)	parce que *'because'* (12)	donc *'so'* (12)
HU	de *'but'* (21)	és *'and'* (19)	ha *'if'* (12)	amikor *'when'* (11)	mert *'because'* (9)
LI	ir *'and'* (29)	bet *'but'* (24)	jeigu *'if'* (12)	taigi *'so'* (10)	kai *'when'* (10)

It appears that the most frequent discourse markers in all the investigated languages are *and* and *but* (and their cross-linguistic equivalents), followed by some subordinating conjunctions and the adverbial *so*. Although our corpus is small, these findings are consistent with larger corpus studies on spoken English [Author] and on written English (e.g. Prasad et al. 2008), except for *so* which is much less frequent in writing than in speech.

For the second step, we used [Author]'s taxonomy of domains and functions, which is specifically designed for annotating discourse markers used in spoken discourse. It consists of four domains:

- the ideational domain is linked to "states of affairs in the world, semantic relations between real events";
- the rhetorical domain is linked to "the speaker's meta-discursive work on the ongoing speech";
- the sequential domain is linked to "the structuring of discourse segments, both at macro- and micro-level";
- the interpersonal domain is linked to "the interactive management of the exchange, in other words, to the speaker-hearer relationship."

In addition to the domains, fifteen functions can be assigned to the discourse markers, including addition, contrast or specification, as can be seen in Table 2. Domains and functions are independent, i.e. any domain can apply to any function and any function can apply to any domain. According to [Author: 20], annotators "can choose to start at domain-level or function-level, to annotate both levels simultaneously or independently". The authors believe that this system vouches for a reliable annotation (high inter-annotator agreement), because of the reduced number of labels and the independence of the two levels (i.e. domains and functions).

Table 2. [Author]'s revised taxonomy with cross-domain functions

Ideational	**Rhetorical**	**Sequential**	**Interpersonal**
[addition] [alternative] [cause] [closing] [concession] [condition] [consequence] [contrast] [enumeration] [opening] [punctuation] [resuming] [specification] [temporal] [topic-shift]			

In this study, we only annotated the domains and functions for three English discourse marker types, i.e. *and*, *but*, *so* (see the motivation in the next section). The disambiguations were performed independently by two experts, and any disagreement was later resolved through discussion among all the authors, until a gold standard for each token was reached.

In a last step of the analysis, we started from the translation equivalents of English *and*, *but* and *so* in the four target languages (e.g. *et*, *mais* and *donc* in French) and examined the types of English original discourse markers of which they are the translations. This method, similar to translation spotting, allows us to account for a type of underspecification specifically related to translation, namely cases where a "strong" marker in the original is translated by a "weaker" marker.

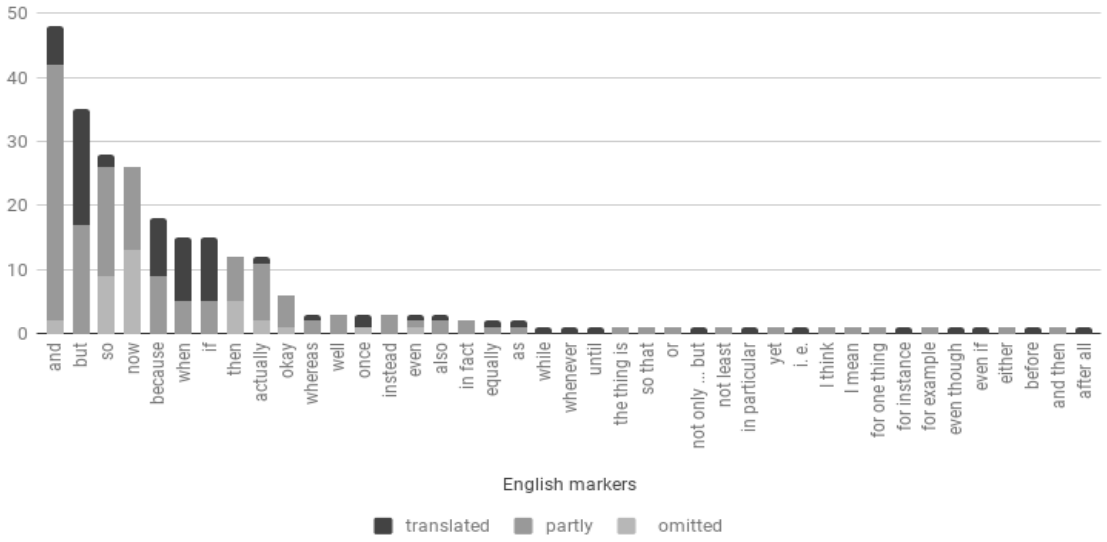
4. Results

4.1 Discourse marker omissions

In the sample, we extracted 261 tokens of discourse markers in English. However, not all of them were maintained in the translations. The number of such omissions varies across languages: there are 79 cases of omission in Lithuanian, 101 in Czech, 115 in French and even 133 in Hungarian, which is more than half of the original tokens.

Omissions do not target all discourse markers equally, thus some types seem to be much more affected by this translation strategy than others. Figure 1 shows each English type in the sample and the proportions of cases which were translated into all four languages ("translated", dark grey on the graph), omitted in at least one language ("partly", grey) or omitted in all four languages ("omitted", light grey). As we can see, there are two relatively frequent markers (more than 10 occurrences) which are never translated into all four languages, i.e. *now* and *then*. This first group is closely followed by *and* and *so*, which are very frequently omitted in the translations (only a few occurrences in dark grey are translated across all languages). By contrast, other types are never completely omitted in all four languages. This concerns *but* and *because*, which are translated in about half of their occurrences and only partly omitted in other cases. Similarly, the subordinating conjunctions *when* and *if* are translated in all four languages most of the time, which is also the case for much rarer discourse markers in the sample, including such as *once*, *while*, *whenever* etc., which only occur once.

Figure 1: Proportion of omissions for all English discourse marker types



Interestingly, whether or not a discourse marker is frequently omitted or maintained in the translations seems to clearly relate to its semantics and functional behaviour. Omissions mainly affect two types of markers: i) speech-specific discourse markers functioning as "punctuators" (*okay*, *now*, *then*), or ii) basic conjunctions expressing a wide range of discourse relations (*and*, *so*). Both types are presented in the following example:

(1) **Okay**, **so** let's imagine **then** that you picked your perfect partner **and** you're settling into a lifelong relationship with them. **Now**, I like to think that everybody would ideally like to avoid divorce, apart from, I don't know, Piers Morgan's wife, maybe?

This excerpt illustrates the discourse-structuring functions of *okay*, *so* or *now*, which are (almost) never translated. In particular, *now* is produced with a high frequency by one speaker only in the sample and functions as a filler, to open a lot of utterances. This marker is not found in the other two texts of the sample, which is why we will not be addressing it any further in this paper (see Aijmer 1988, Schourup 2011 for further analysis).

On the other hand, omission is not as frequent for markers with a more specific semantics (i.e. less polyfunctional items), such as *but*, *because*, *when* or *if*: they respectively express contrastive, causal, temporal and conditional relations, which do not seem to be implicitated often. This finding relates to Hoek et al.'s (2017) study, where they showed that conditional (*if*) and negative additive (*but*) relations are often explicit in translations. In sum, omission appears to be tightly related to the semantics and polyfunctionality of the markers to be translated.

Omissions are strikingly similar across all target languages regarding the types and ranking of omitted markers, as illustrated in Table 3. While it could be hypothesized that omissions are partly influenced by the presence or absence of a formal-functional equivalent in the target language, this table rather shows that the translation strategy of omission or implicitation is in fact primarily guided by the semantics and pragmatic function of the markers in the original text. This result is even more striking given that the languages under scrutiny belong to different typological families (Germanic, Baltic, Finno-Ugric, Slavic and Romance), so that the recurring patterns of omissions cannot be traced back to family resemblance in the category of discourse markers. In fact, all languages have a direct equivalent of *and*, *so* and *but*, while only some have an equivalent for *now* (Hungarian *nos*, Lithuanian *na*).

Table 3: Number of omissions for the most frequent English discourse markers (5+ cases) in each target language

Lithuanian	now (18), and (17), so (12), then (8), actually (5)
Hungarian	and (25), now (23), so (22), but (15), then (10), okay (6), actually (6)
Czech	now (24), and (18), so (17), then (9), but (6), actually (5)
French	now (23), and (20), so (19), but (11), actually (8), then (8)

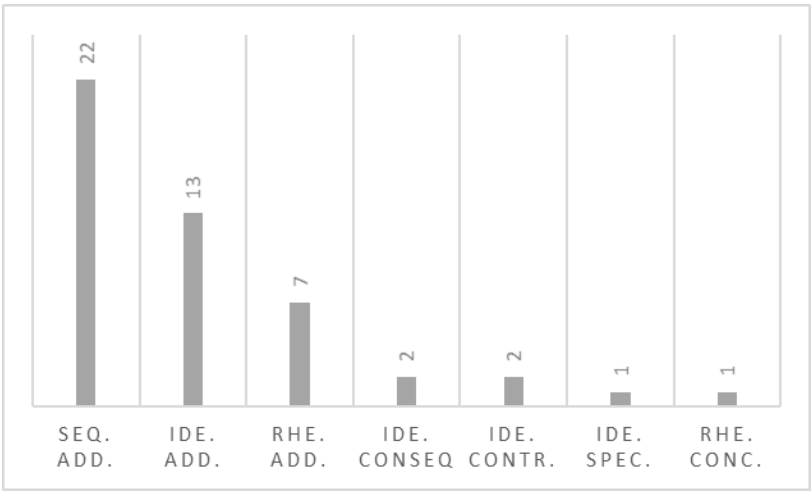
To come back to the notion of underspecification, we can say that frequent omission is a sign of low information value, which the translator did not find relevant or worthwhile to maintain in the target language, and as such, omission corresponds to a first type of multilingual underspecification. In the remainder of the study, we will focus on the three most frequent markers in the sample, i.e. *and, but, so*, which are omitted to different extents (cf. above). The large proportion of cases where *but* is maintained in all four languages suggests that this marker displays a narrower functional spectrum than the other two markers, which we expect to be more variable or, in other words, more underspecified. We will test these hypotheses through a functional analysis in the next sections.

4.2 Discourse marker functions

4.2.1 Functions and translations of "and"

The 48 tokens of the original *and* were assigned to three domains (ideational, rhetorical, sequential), five functions (addition, concession, consequence, contrast, specification) for a total of seven domain-function combinations, which are shown in Figure 2 along with their frequency.

Figure 2: Frequency of the different functions of *and* in the original English



First, we can see that the basic additive meaning of *and* takes up a vast majority of its uses, although they are spread across three functional domains: the addition primarily targets the sequential domain (i.e. *and* expresses continuation in the discourse flow), followed by the

ideational domain (*and* connects two objective contents) and, in only seven cases, the rhetorical domain (*and* introduces a new argument with some stylistic effect). The three "flavours" of addition are illustrated below.

(2) So these equations, they predict how the wife or husband is going to respond in their next turn of the conversation, how positive or negative they're going to be. **And** these equations, they depend on the mood of the person when they're on their own, the mood of the person when they're with their partner, but most importantly, they depend on how much the husband and wife influence one another. [sequential addition]

(3) So let's imagine then, that you start dating when you're 15 **and** ideally, you'd like to be married by the time that you're 35. [ideational addition]

(4) There'd be a huge spread in her scores. **And** actually it's this spread that counts. [rhetorical addition]

These examples all correspond to the basic meaning of *and*, the variation only comes from the particular speaker intention or aspect of discourse that is targeted by the marker. By contrast, *and* was also found in contexts where the relation is not that of addition but rather one of consequence (the second argument is the logical result of the first one), contrast (Example 6), concession (Example 7) or specification (the second argument adds a detail or an example to the first one).

(6) The people who fancy you are just going to fancy you anyway, **and** the unimportant losers who don't, well, they only play up to your advantage. [ideational contrast]

(7) How ironic that I work in human resources, a profession that works to welcome, connect and encourage the development of employees, a profession that advocates that the diversity of society should be reflected in the workplace, **and** yet I have done nothing to advocate for diversity. [rhetorical concession]

In Example 6, there is a clear contrast between "the people who fancy you" on the one hand, and "the unimportant losers" on the other. Example 7 further illustrates the role of co-occurring discourse markers in the interpretation of *and*, which clusters with *yet* here to express a concession or counter-expectation between the speaker's social profession and her own behaviour. These examples differ from Examples (2)-(4) above in that they no longer express the marker's basic meaning. Although much rarer than the basic sense of addition, such uses of *and* correspond to a different type of monolingual underspecification, where the pragmatic interpretation is richer than the encoded semantic instruction of the marker. This type of underspecification is not related to translation but rather to pragmatic inferences which comprehenders draw from the content of the arguments or from their world knowledge (Blakemore & Carston 1999, Luscher & Moeschler 1990).

We can now combine the two types of underspecification observed so far, namely (multilingual) omissions and (monolingual) functional shift, by relating each function of *and* to the proportion of omissions (see Table 4). First, only sequential addition has cases where *and* was omitted in all four languages. Moreover, this function only has one case (out of 22) which is translated in all languages, which suggests that this function is the most "expendable" use of *and*. Secondly, the vast majority of occurrences of *and* are partly omitted in at least one language, especially rhetorical addition with six partly omitted cases out of seven tokens. By contrast, two of the rare functions, i.e. ideational consequence and contrast, are translated in all languages in one case out of two. This latter finding is highly relevant to our discussion of underspecification and suggests that, when *and* expresses more than basic addition, it tends to

be maintained in the translations, which is even more striking given that these uses are very rare.

Table 4: Cross-tabulation of functions and omissions of the English *and*

Functions	**omitted (all)**	**partly omitted**	**translated (all)**	**Total**
sequential addition	3	18	1	22
ideational addition	0	9	4	13
rhetorical addition	0	6	1	7
ideational consequence	0	1	1	2
ideational contrast	0	1	1	2
ideational specification	0	1	0	1
rhetorical concession	0	1	0	1
Total	**3**	**37**	**8**	**48**

Overall, omission mainly affects the additive uses of *and*, and in particular sequential and rhetorical addition. Ideational addition, however, is preserved in about half of the occurrences of *and*, which suggests that *and* gives more information in those uses where the logical operation of addition is more prevailing than in sequential or rhetorical addition. When *and* is underspecified in the sense that its pragmatic interpretation is richer than its semantics, it is never completely omitted and it can even be translated into all languages when it expresses consequence or contrast. Therefore, multilingual and monolingual underspecification do not seem to match for *and*.

Turning to the translations, we have observed that, in all four target languages, *and* is translated into a number of different discourse markers: one in Czech, two in French, four in Hungarian and Lithuanian (see Table 5). The most frequent translation of *and* is its semantic equivalent, which is an additive coordinating conjunction in all languages.

Table 5: Translations of *and* into Czech, French, Hungarian and Lithuanian

Czech	**French**	**Hungarian**	**Lithuanian**
a *'and'* (29)	et *'and'* (23)	és *'and'* (19)	ir *'and'* (24)
	mais *'but'* (2)	s *'and'* (2)	o *'and/but'* (4)
		ehhez *'in addition'* or *'for this'* (1)	taip pat *'and also'* (1)
		egyébként *'otherwise'* (1)	ir todėl *'and so'* (1)

Other, much rarer translations of English *and* are not particularly associated with specific functions: for instance, French *mais* 'but' is used in a case of sequential addition and in one of rhetorical concession, where the translation subsumes the meaning of the original cluster "and yet". Lithuanian *o* has a double core meaning expressing both addition and contrast, which is reflected in the functions of the original *and* which is translated: two cases of sequential addition, one of rhetorical addition and one of ideational contrast. Lithuanian *ir todėl* specifies the additive *ir* with a resultative meaning and translates a case of ideational consequence.

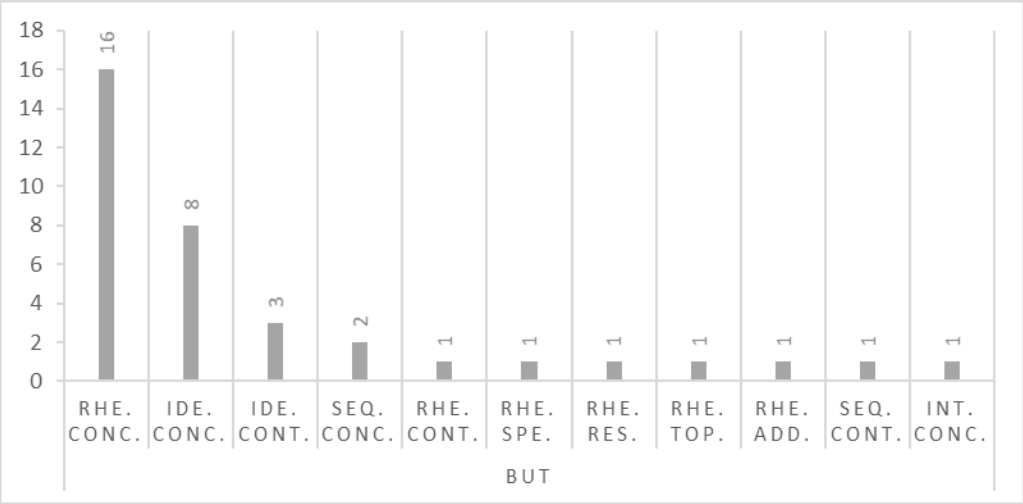
In sum, the functional spectrum of *and*, which is wider than mere addition, is only somewhat reflected in its translation, with only a few cases where the marker in the target language lexically encodes the enriched pragmatic interpretation of the underspecified original *and*. Similarly, patterns of omissions only partially correspond to functional variation, since this translation strategy mostly applies to *and* expressing sequential addition, which stands out as a low-information function, typical of speech where *and* simply connects discourse segments

without instructing any particular coherence relation. Underspecified uses of *and* such as consequence or contrast are not always omitted (although the data is very scarce) and can be translated by more specified markers, such as Lithuanian *ir todėl* 'and so' or French *mais* 'but'.

4.2.2 Functions and translations of "but"

The 36 tokens of the original *but* were assigned to all four domains and six functions (addition, concession, contrast, specification, topic-resuming, topic-shift) for a total of 11 domain-function combinations, which are shown in Figure 3 along with their frequency.

Figure 3: Frequency of different functions of *and* in the original (English)



We can see that a vast majority of cases are concessive uses of *but*, which is the only use in our corpus that can be expressed in all four domains, as illustrated in (8)-(11) below:

(8) Nobody thinks that she's ugly, **but** she's not a supermodel, either [ideational]

(9) Now of course, it takes bit more than just a low negativity threshold and not compromising to have a successful relationship. **But** I think that it's quite interesting to know that there is really mathematical evidence to say that you should never let the sun go down on your anger [rhetorical]

(10) Now just by using these very simple ideas, Gottman and his group were able to predict whether a given couple was going to get divorced with a 90 percent accuracy. **But** it wasn't until he teamed up with a mathematician, James Murray, that they really started to understand what causes these negativity spirals and how they occur [sequential]

(11) Every time I would think about coming out in the past, I would think to myself, **but** I just want to be known as Morgana, uniquely Morgana, but not "my lesbian friend Morgana," or "my gay coworker Morgana" [interpersonal]

Similarly to the additive function of *and* discussed in the previous section, here we can see that all these examples retain some trace of the basic concessive meaning of *but*, which targets different types of entities: facts (8), personal evaluations (9), steps in the narrative (10), opinions or pseudo-reported internal dialogue (11). This latter use is particularly interesting: the speaker is imitating an exchange with herself where, on the one hand, she would "think about coming out" yet, on the other, she disagrees with her own idea and starts her response by *but*. This interpersonal use has already been identified in French *mais* 'but' by [Author].

Another multi-domain function of *but* is contrast, which is less frequent than concession and corresponds to clear oppositions between two facts (ideational contrast), two subjective arguments (rhetorical contrast) or two larger units representing scenarios (sequential contrast), as in (12) below:

(12) Okay, another risk is, let's imagine, instead, that the first people that you dated in your first 37 percent are just incredibly dull, boring, terrible people. Now, that's okay, because you're in your rejection phase, so that's fine, you can reject them. **But** then imagine, the next person to come along is just marginally less boring, dull and terrible than everybody that you've seen before.

In this example, the audience is invited to "imagine" two scenarios (someone being boring and someone being marginally less dull). The distance between the two *imagine*-utterances, interrupted by an evaluation ("now that's okay"), and the size of the related segments create a higher-order structuring effect so that the contrast no longer targets facts (or facts alone) but (also) steps in the development of the speaker's presentation.

Figure 3 also shows rare cases of addition, specification, topic-resuming or topic-shift, all in the rhetorical domain. The latter two can be seen as desemanticized uses of *but* where no specific discourse relation is expressed beyond a discourse-structuring function. The former two are at first sight quite contradictory with the basic adversative meaning of *but*: the two related segments are no longer opposed but, on the contrary, added and seen as elaborations, which relates to the construction "not only, … but also", instantiated in (13) below:

(13) if we whitewash our stories for the sake of mass appeal, **not only** will we fail, **but** we will be trumped by those with more money and more resources to tell our stories [rhetorical addition]

We do not consider such uses as cases of monolingual underspecification since, contrary to *and*, the various interpretations of *but* are not in any way "richer" or more informative than its basic adversative meaning, they simply differ from it through specific structural configurations. It even seems as if *but* can "lose" its adversative nature in specific rhetorical contexts, where it simply adds a new topic, new information or an example, which relates more to semantic bleaching rather than to underspecification.

Turning to omissions, *but* radically differs from *and* as being translated into all four languages in more than half of its occurrences, and almost never omitted in all the languages (with only one exception), as can be seen in Table 6.

Table 6: Cross-tabulation of functions and omissions of the English *but*

Functions	omitted (all)	partly omitted	translated (all)	Total
rhetorical concession	0	7	9	16
ideational concession	0	4	4	8
ideational contrast	0	1	2	3
sequential concession	0	1	1	2
rhetorical contrast	0	1	0	1
rhetorical specification	0	1	0	1
rhetorical topic-resuming	0	0	1	1
rhetorical topic-shift	0	1	0	1
rhetorical addition	0	0	1	1
sequential contrast	0	0	1	1

interpersonal concession	1	0	0	1
Total	**1**	**16**	**19**	**36**

It appears that *but* is always translated into at least one target language, if not in all languages, with the notable exception of the case of interpersonal concession analysed above (Example 11 above). This particular interjective use of *but* is quite unusual and serves more to reflect the speaker's tone and state of mind rather than to actually structure discourse. No particular trend can be noted for other functions that are quite equally distributed across partly omitted and fully translated cases: some of the "bleached" uses discussed above are omitted (e.g. rhetorical specification) while others are translated (rhetorical addition); more frequent functions show similar proportions in each column. Overall, we can interpret this tendency towards translation (rather than omission) as the result of a strong adversative meaning of *but*, which has been shown elsewhere to favour explicitation rather than implicitation (Hoek et al. 2017). In other words, we can say that neither monolingual nor multilingual underspecification substantially apply to *but*.

Table 7 shows the translations of *but* and their frequency in the four languages of the sample. Again, the languages vary regarding the number of possible translation equivalents (six in Lithuanian, five in Czech, three in French and Hungarian). There is always one most frequent translation corresponding to a contrastive conjunction (*ale*, *mais*, *de*, *bet*), which can be used across all functions of *but*. The Czech *však* 'however' and the Lithuanian *tačiau* 'however' also show several occurrences: the latter is limited to the basic uses of ideational or rhetorical concession and contrast. It is interesting to note that rhetorical addition is translated into Lithuanian by a complex marker consisting of contrastive *bet* and additive *ir* (*bet ir*); similarly, the same use is translated by the Czech additive *taky* 'also', whereas French and Hungarian resort to typically adversative markers (*mais* 'but' and *hanem* '*but', 'instead'*).

Table 7: Translations of *and* into Czech, French, Hungarian and Lithuanian

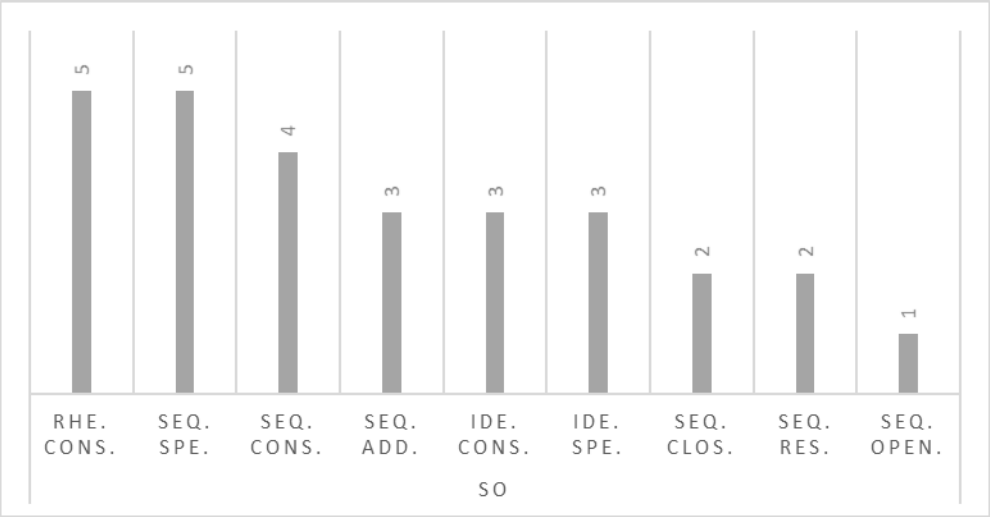
Czech	**French**	**Hungarian**	**Lithuanian**
ale *'but'* (20)	mais *'but'* (23)	de *'but'* (19)	bet *'but'* (23)
však *'however'* (5)	et *'and'* (1)	hanem '*but', 'instead'* (1)	tačiau *'however'* (5)
ovšem *'however* (3)	même si *'even if'* (1)	pedig *'in turn'*, *'now', 'although'* (1)	ir *'and'* (2)
a *'and'* (1)			bet ir *'but and'* (1)
taky *'also'* (1)			ir priedo *'in addition'* (1)
			o *'but & and'* (1)

There is one case of ideational concession where *but* is translated by a more specific discourse marker, i.e. French *même si* 'even if" (Example 8 above), which might indicate the translator's will to over-specify the meaning of *but* even in its standard basic function. Overall, it appears that most of the functional spectrum of *but* can be translated by its direct translation equivalents in the four languages, with no particular associations between functions and translations, except for the case of rhetorical addition in Czech and Lithuanian. We take this finding as another evidence of the strong polyfunctionality of *but*: rarely omitted, rarely translated by more specific markers, it can express concession in all four domains of our taxonomy as well as additional "bleached" meanings where adversativity is no longer perceivable.

4.2.3 Functions and translations of "so"

The 28 tokens of the original *so* were assigned to three domains (all but interpersonal) and six functions (addition, closing, consequence, opening, specification, topic-resuming) for a total of nine domain-function combinations, which are shown in Figure 4 along with their frequency. Contrary to both *and* and *but*, where one domain-function combination clearly dominates the sample, the functional spectrum of *so* is much more spread over cross-domain variants of consequence and specification mostly, in addition to other uses exclusively in the sequential domain (addition, closing, opening, topic-resuming).

Figure 4: Frequency of dfferen functions of *so* in the original (English)



Consequence displays uses in the ideational, rhetorical and sequential domains. As noted earlier, this variation targets the nature of related units and the speaker's intention or attitude towards their discourse. While the opposition between ideational (14) and rhetorical (15) consequence is fairly straightforward, as it corresponds to the well-known divide between objective and subjective (or semantic and pragmatic) relations (cf. Section 2.1), sequential uses once more depend more on the higher-order structure of the speech, the size and distance of the related units. As a marker of sequential consequence, *so* usually introduces a conclusion to a larger speech segment (Sweetser 1990), as in (16) below:

(14) Now, if you're following the maths, I'm afraid no one else comes along that's better than anyone you've seen before, **so** you have to go on rejecting everyone and die alone.

(15) Because I believe that mathematics is so powerful that it has the potential to offer us a new way of looking at almost anything. Even something as mysterious as love. And **so,** to try to persuade you of how totally amazing, excellent and relevant mathematics is, I want to give you my top three mathematically verifiable tips for love.

(16) [*long passage on a mathematical approach to choosing one's partner*] A Valentine's Day card like this. "My darling husband, you are marginally less terrible than the first 37 percent of people I dated." It's actually more romantic than I normally manage. Okay, **so** this method doesn't give you a 100 percent success rate, but there's no other possible strategy that can do any better.

In Example (16), *so* is equivalent to "to conclude" or "all in all" and indicates both a general conclusion to a lengthy development (consequence function) and the end of such development (sequential domain).

Considering that *so* is typically associated with the core meaning of consequence in the literature, it is striking to see that, in the sample, specification is almost as frequent as consequence (8 vs. 12 tokens), which points to the polysemy of the marker (cf. e.g. Buysse 2012). Specification varies across the ideational and sequential domains and is most frequent in the latter, as in (17) below:

(17) Thankfully, there's a rather delicious bit of mathematics that we can use to help us out here, called optimal stopping theory. **So** let's imagine then, that you start dating when you're 15 and ideally, you'd like to be married by the time that you're 35 [*continued*]

In this example, as well as in other cases of sequential specification, the segment introduced by *so* is an elaboration (example or more specific explanation) of the previous segment. Since this elaboration is lengthy and opens a new (yet related and hierarchically dependent) passage in the discourse, the specification operates at a higher level of discourse structuring, hence in the sequential domain and could be paraphrased by "let's talk about this in more details". This contrasts with the ideational basis of specification, which merely corresponds to a colon in punctuation, as is exemplified below:

(18) the way that people would vote would be very different. **So** Portia's scores would all be clustered around the 4 because everybody agrees that she's very beautiful, whereas Sarah Jessica Parker completely divides opinion.

Like *but*, *so* displays additional functions which are all in the sequential domain and which have little to do with its basic meaning(s). As a marker of addition, opening, closing or topic-resuming, it signals the inevitability of progress in the discourse. It can signal the opening, closing or resumption of a discourse segment or topic. In these uses, *so* frequently co-occurs with other discourse markers, especially with *then* (19) and *okay* (20):

(19) But, equally, you don't really want to leave it too long if you want to maximize your chance of long-term happiness. As my favorite author, Jane Austen, puts it, "An unmarried woman of seven and twenty can never hope to feel or inspire affection again." Thanks a lot, Jane. What do you know about love? **So** the question is **then**, how do you know when is the right time to settle down given all the people that you can date in your lifetime? [resuming]

(20) I think this is conclusive proof, if ever it were needed, that everybody's brains are prewired to be just a little bit mathematical. **Okay**, **so** that was Top Tip #2. [closing]

Our analysis of these cases does not resort to the notion of underspecification but rather to bleaching, where *so* seems to have lost its causal or inferential component (much like *but* when it no longer showed adversativity). The sequential domain is once more the most productive one in the polyfunctionality of *so*. This is partly reflected in the omissions: no sequential function of *so* is fully translated into all four languages, while most cases of complete omission mainly concern sequential uses, with two exceptions (ideational specification), as it can be seen in Table 8.

Table 8: Cross-tabulation of functions and omissions of the English *so*

Functions	**omitted (all)**	**partly omitted**	**translated (all)**	**Total**
rhetorical consequence	0	4	1	5

sequential consequence	1	3	0	4
sequential specification	3	2	0	4
sequential addition	3	0	0	3
ideational consequence	0	2	1	3
ideational specification	2	1	0	3
sequential closing	0	2	0	2
sequential resuming	0	2	0	2
sequential opening	0	1	0	1
Total	**9**	**17**	**2**	**28**

Very few (two) occurrences of *so* are maintained in all four languages, and they both correspond to the basic meaning of consequence in their ideational and rhetorical variants: these two uses are never omitted, which points to their centrality in the functional spectrum of *so* as well as to their "necessity" as markers of discourse coherence. The three cases of sequential addition are fully omitted, while the other bleached sequential functions of *so* are only partly omitted. In terms of underspecification, it is once more the sequential domain that appears most expendable, since it only targets the flow of discourse or the structuring process: these functions are not part of the core meaning of *so* and can therefore be rendered by other cues such as prosody or lexis. This result also relates to Hoek et al.'s (2017) finding on speech-act relations, which are often implicit in translations because they are easy to recognize and because they are "close to the I-here-now of the situation" (2017: 127).

Lastly, we finish the analysis of *so* by looking at how it is translated in the data (Table 9). We can see that Lithuanian again has the greatest number of translation equivalents (six), which is related to a lower number of omissions of *so* in this language; Hungarian has four translation equivalents, Czech three, and there is only one in French (*donc* 'so'). All languages, except for Hungarian, seem to favour one particular marker (Cz. *tedy*, Fr. *donc*, Lit. *taigi*), which are direct equivalents of the adverbial, while Hungarian translations are spread over markers which all share a resultative component.

Table 9: Translations of *so* into Czech, French, Hungarian and Lithuanian

Czech	**French**	**Hungarian**	**Lithuanian**
tedy *'so'* (6)	donc *'so'* (8)	így '*this way', 'thus'* (2)	taigi *'so'* (9)
takže *'so'* (3)		tehát *'so', 'therefore'* (2)	tai *'that'* (2)
aby *'in order to'* (1)		ehhez *''in addition to this', 'for this (purpose)'* (1)	tarkim *'let's say'* (2)
		szóval *'so'* (1)	dabar *'now'* (1)
			na *'well'* (1)
			tam kad *'in order to'* (1)

The Czech *takže* and *aby* are restricted to consequence while *tedy* is also used for text-structuring (sequential) functions. Similarly, Hungarian *ehhez* (*'in addition to this'*, *'for this purpose'*) and *így* ('*this way'* or '*thus'*) are restricted to the ideational and rhetorical consequence, while *szóval* (*'so'*) and *tehát* (*so'* or *'therefore'*) are used for sequential functions. In Lithuanian, some markers are only used to express consequence, even though they originate from different classes with different semantics (preposition of consequence *tam kad*, temporal

adverbial *dabar* 'now', speech particle *na* 'well'); *taigi* expresses a large range of functions across domains, while *tarkim* is restricted to the function of sequential specification, of which it is a literal translation (cf. our paraphrase above "let's talk about this in more details").

In sum, the functional spectrum of *so* seems to be divided into three unequal parts: consequence (and its cross-domain variants), which is translated by both generic and dedicated markers and is always at least partly translated; sequential functions (e.g. functions of opening, topic-resuming), which are the least frequent and tend to be translated by specific markers (in addition to the direct equivalents of *so*); specification, which is somewhere in-between the previous two uses in terms of frequency and meaning, with both ideational and sequential variants and a dedicated marker in Lithuanian (i.e. *tarkim*). We could see that, from a monolingual perspective, *so* is not concerned with underspecification, but multilingually, when looking at omissions and translations, patterns seem to emerge regarding whether and how to translate various functions of this polysemous discourse marker.

4.3 From translations to the original: a further type of underspecification

So far, the parallel data has led us to analyse two types of underspecification: a monolingual, functional one, where the function in context is "richer" than the encoded semantics; and a multilingual, translation-related one, where a discourse marker is omitted in the translation. We now turn to a further, somewhat intermediate type of underspecification, which is also specific to parallel data, namely the translation of a "strong" discourse marker by a "weaker", semantically less informative or more generic one in the target language.

To examine this phenomenon, we started from the translations, extracted all cases of the three basic discourse markers of addition, consequence and contrast in each language and looked at the English original discourse marker that they translate. The three markers in each language were identified on the basis of the analyses reported in the previous sections, as the most frequent translation equivalents of *and*, *but* and *so*: *a*, *ale*, *tedy* in Czech; *et*, *mais*, *donc* in French; *és*, *így*, *tehát, de* in Hungarian; *ir*, *o*, *taigi* in Lithuanian.

We found only six relevant cases in the sample:

- original *but* translated by an additive marker (one case of French *et*; one case of Czech *a*; two cases of Lithuanian *ir*);
- original *whereas* translated by the additive-contrastive Lithuanian marker *o*;
- original *in fact* translated by additive *et* in French.

In the first case (see Example 21), the original and the translation come from different semantic classes (contrastive and additive), which reflects the ability of *and* and its cross-linguistic equivalents to be used in underspecified contexts to express different meanings.

(21) Now, I think that we can all agree that mathematicians are famously excellent at finding love. **But** it's not just because of our dashing personalities, superior conversational skills and excellent pencil cases. [Lithuanian *ir*; Czech *a*]

This case of *but* was annotated as rhetorical concession, so that this type of underspecification can affect uses of *but* which are very close to its core meaning. In the case of *whereas* and *in fact* (Examples 22 and 23), the translations come from roughly the same semantic class (contrastive and additive-elaborative, respectively), but they differ in their degree of specificity or informativeness (stronger in the original).

(22) So let's say that you think somebody's attractive, but you suspect that other people won't necessarily be that interested. That means there's less competition for you and it's an extra incentive for you to get in touch. **Whereas** compare that to if you think somebody is attractive but you suspect that everybody is going to think they're attractive. [Lithuanian *o*]

(23) Now if you do this, it can be mathematically proven **in fact** that this is the best possible way of maximizing your chances of finding the perfect partner.

(23') *Si vous faites cela,* ***et*** *c'est mathématiquement démontrable, c'est la meilleure façon possible de maximiser vos chances de trouver le partenaire idéal*.

We can see in Example 23' that, in the French translation, the discourse marker changed its position from the final *in fact* to the initial *et* with respect to the parenthetical aside "it can be mathematically proven", while retaining the same meaning.

The low frequency of this phenomenon can be explained by the small size of the sample and also relates to Dupont & Zufferey's (2017) finding that this type of underspecification is rare in TED Talks. Nevertheless, we can see once again that underspecification is multi-faceted, that high-frequency generic discourse markers are involved in more than one type of underspecification, and that parallel corpus data is highly valuable to detect such phenomena in translations.

5. Conclusion

Starting from the observation that discourse markers are highly polyfunctional, and that the notion of underspecification covered multiple phenomena, this parallel corpus study investigated the functions, translation equivalents and omissions of three high-frequency discourse markers in English, i.e. *and*, *but*, *so*, in a sample of TED Talks and their translation into Czech, French, Hungarian and Lithuanian. The combined use of parallel data and discourse annotation in functions and domains allowed us to identify three types of underspecification: functional shift (monolingual), omission and weaker translation (multilingual).

Besides this theoretical contribution, our study refined the semantic-pragmatic description of the three most frequent English discourse markers (*and*, *but*, *so*) by not only identifying their various functions but also several variants of these functions across [Author]'s "domains". The independence between functions and domains allowed us to draw a detailed yet structured portrait of the polyfunctionality of these markers by distinguishing, in particular, various subtypes of addition, concession and consequence, among others.

Our findings reveal that not all discourse marker types nor their functions are equally affected by underspecification, and that monolingual and multilingual underspecification do not always map for a particular marker. However, we have observed that the processes of underspecification are common to the five languages under scrutiny, that they proceed in parallel, depending on the semantics of discourse markers.

While discourse markers have been repeatedly studied in translation corpora in the past, the present annotation methodology is considerably innovative and bears theoretical and empirical implications. Nevertheless, this study also has limitations, the first of which being the small size of the sample. Although this restriction was imposed by the time-consuming methodology of functional annotation, it does prevent us from generalising our quantitative findings beyond the particular texts and register under scrutiny. Our findings remain in line with previous pragmatic descriptions of *and*, *but*, *so* and with previous translation studies, so that we believe

quality was not hurt by quantity. Besides corpus size, the data was also restricted to one very specific type of setting (between speech and writing) and of translation (subtitling done by amateurs). These two limitations call for further research on more data from different registers in order to check whether underspecification (and the markers it affects) varies with register, translator's expertise and possibly other factors.

References

[Author]

[Author]

[Author]

Aijmer, Karin. 1988. Now may we have a word on this: the use of *now* as a discourse particle. Corpus linguistics: hard and soft, ed. by M. Kyto et. al. Rodopi.

Aijmer, Karin & Simon-Vandenbergen, A. M. 2006. *Pragmatic Markers in Contrast*. Amsterdam: Elsevier.

Aijmer, Karin. 2007. The interface between discourse and grammar: The fact is that. 31-46. 10.1075/pbns.161.05aij. 2007

Altenberg, Bengt. 1999. Adverbial connectors in English and Swedish: Semantic and lexical correspondences. In Hasselgård & Oksefjell 249-68.

Altenberg, Bengt. 2007. The Correspondence of Resultive Connectors in English. Nordic Journal of English Studies. Vol. 6, 1. Available from http://ojs.ub.gu.se/ojs/index.php/njes/article/view/11/14

Baker, Mona. 2011. *In Other Words – A Coursebook on Translation*. London: Routledge.

Biber, Douglas. 2006. *University Language: A Corpus-Based Study of Spoken and Written Registers*. Philadelphia: John Benjamins.

Blakemore, Diane & Carston, Robyn. 1999. The interpretation of and-conjunctions. In Iten, C. & A Neelman (eds.), University College Working Papers in Linguistics 11:1-20. London: UCL.

Buysse, Lieven. 2012. *So* as a multifunctional discourse marker in native and learner speech. *Journal of Pragmatics* 44 (13), 1764-1782.

Castellà, Josep M. *Oralitat I Escriptura. Dues Cares de la Complexitat del Llenguatge*. Barcelona: Publicacions de l'Abadia de Montserrat.

Cettolo, M., Girardi, C., and Federico, M. 2012. WIT: Web Inventory of Transcribed and Translated Talks. *Proceedings of the 16th EAMT Conference* (pp. 261-268). Trento, Italy.

Cuenca, Maria-Josep. 2008. Pragmatic markers in contrast: The case of well. Journal of Pragmatics. 40. 1373-1391.

Cuenca, Maria-Josep. 2013. The fuzzy boundaries between modal and discourse marking. In L. Degand, B. Cornillie and P. Pietrandrea (eds), Discourse Markers and Modal Particles. Description and Categorization, pages 191-216, John Benjamins, Amsterdam.

Degand, Liesbeth. 1996. *Causation in Dutch and French. Interpersonal Aspects*. In: Ruqaiya Hasan, Carmel Cloran, David Butt, *Functional Descriptions: Theory in Practice* (Current Issues in Linguistic Theory; 121), John Benjamins: Amsterdam, p. 207-237.

Degand, Liesbeth. 2004. *Contrastive analyses, translation and speaker involvement : the case of "puisque" and "aangezien"*. In: Achard Michel, Kemmer Suzanne (eds.), *Language, culture and mind*, CSLI Publications: Stanford (Canada), p. 251-270.

Dupont, Maïté; Zufferey, Sandrine. 2017. *Methodological issues in the use of directional parallel corpora*. In: *International Journal of Corpus Linguistics*, Vol. 22, no.2

Egg, M; Redeker, G. 2007. Underspecified discourse representation. In A. Benz and P. Kühnlein, editors, Constraints in Discourse, Amsterdam. Benjamins.

Egg, Markus. 2010. Semantic underspecification. *Language and Linguistics Compass* 4(3): 166-181.

Erman, Britt. 2001. Pragmatic markers revisited with a focus on *you know* in adult and adolescent talk. *Journal of Pragmatics* 33: 1337-1359.

Ferreira, F; V. Ferraro; K.G.D. Bailey. 2002. Good-enough representations in language comprehension. *Current Directions in Psychological Science*, 11, pp. 11-15

Fischer, K. 2006. Towards an Understanding of the Spectrum of Approaches to Discourse Particles: Introduction to the Volume. In: K. Fischer (Ed.) *Approaches to Discourse Particles* (pp. 1–20). Oxford/Amsterdam: Elsevier.

Fox Tree, Jean E. 2014. Discourse markers in writing. *Discourse Studies*, Volume: 17 issue: 1, pages: 64-82.

Fraser, Bruce. 1990. An approach to discourse markers. Journal of Pragmatics 14. 383-398.

Fraser, Bruce. 1999. What are discourse markers?. Journal of Pragmatics. 31. 931-952.

Frisson, Steven & Pickering, Martin J. 2011. Obtaining a Figurative Interpretation of a Word: Support for Underspecification, Metaphor and Symbol, 16:3-4, 149-171

Frisson, Steven. 2009. Semantic Underspecification in Language Processing. Language and Linguistics Compass. 3. 111-127.

Fuller, Janet. 2003. The influence of speaker roles on discourse marker use. Journal of Pragmatics. 35. 23-45.

González, Montserrat. 2005. Pragmatic markers and discourse coherence relations in English and Catalan oral narrative. In: Discourse Studies. Volume: 7 issue: 1, pages: 53-86.

Halliday, M. A. K., and Ruqaiya Hasan. 1976.*Cohesion in English.* London: Longman.

Halliday, M.A.K.; Hasan, R. 1985. Language, Context and Text: Aspects of language in a social semiotic perspective. Geelong, Victoria: Deakin University Press.

Hoek, Jet & Zufferey, Sandrine & Evers-Vermeul, Jacqueline & Sanders, Ted. (2017). Cognitive complexity and the linguistic marking of coherence relations: A parallel corpus study. Journal of Pragmatics. 121.

Horn, Laurence. 1984. Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature.

Irmer, Matthias. 2011. Bridging Inferences: Constraining and Resolving Underspecification in Discourse Interpretation. Berlin/Boston: De Gruyter.

Kenning, Marie-Madeleine. 2010. What are parallel and comparable corpora and how can we use them?. The Routledge Handbook of Corpus Linguistics.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In Proceedings for the Tenth Machine Translation Summit, pp. 139-142, Phuket, Thailand

Kunz, Kerstin & Lapshinova-Koltunski, Ekaterina. 2015. Cross-linguistic analysis of discourse variation across registers. Nordic Journal of English Studies 14(1): 258-288.

Lefer, Marie-Aude & Grabar, Natalia. 2015. Super-creative and over-bureaucratic: A cross-genre corpus-based study on the use and translation of evaluative prefixation in TED talks and EU parliamentary debates. Across Languages and Cultures: A Multidisciplinary Journal for Translation and Interpreting Studies, 16(2), pp. 187–208.

Luscher, Jean-Marc & Moeschler, Jacques. 1990. Approches dérivationnelles et procédurales des opérateurs et connecteurs temporels: les exemples de et et de enfin. In Cahiers de linguistique française 11, pp. 77-104.

Mortier, Liesbeth & Degand, Liesbeth. 2009. Adversative discourse markers in contrast. International journal of corpus linguistics, 14(3), pp. 338-366.

Redeker, Gisela. 1990. Ideational and pragmatic markers of discourse structure. Journal of Pragmatics 14(3):367-81.

Samy, Doaa & González-Ledesma, Ana. 2008. Pragmatic Annotation of Discourse Markers in a Multilingual Parallel Corpus (Arabic- Spanish-English).

Sanders, T. J.M. 1992. Discourse Structure and Coherence: Aspects of a Cognitive Theory of Discourse Representation. Doctoral dissertation, Tilburg University.

Schiffrin, Deborah. 1987. Discourse Markers. Cambridge University Press, Cambridge.

Spooren, Wilbert. 1997. The processing of underspecified coherence relations. Discourse processes 24(1), pp. 149-168.

Steele, David. 2015. Improving the translation of discourse markers for Chinese into English. In *Proceedings of NAACL-HLT 2015 Student Research Workshop*, Denver, June 1st: 110-117.

Sweetser, E. 1990. From Etymology to Pragmatics. Cambridge: Cambridge University Press.

Van Dijk, T. 1979 “Pragmatic connectives”. Journal of Pragmatics 3:447-456.

Zufferey, Sandrine; Degand, Liesbeth. *Explicit and implicit discourse relations across languages*. 13th International Pragmatics Conference (New Delhi).

Ludivine Crible obtained her PhD in Linguistics at the Université catholique de Louvain in 2017, on the topic of discourse markers and disfluency in spoken English and French. She published several papers on crosslinguistic and multimodal (writing, speech, gesture) studies of a range of phenomena such as discourse markers, but also fillers or repetitions. She uses corpus-based methods (especially corpus annotation) in order to provide systematic accounts of complex linguistic categories from a contrastive perspective.

Nijolė Burkšaitienė, PhD, is a professor of Education Sciences at the Institute of Humanities at Mykolas Romeris University, Vilnius (Lithuania). The researcher conducts interdisciplinary studies into teaching and learning English for Specific Purposes, creativity development in university studies and translation. She is the author of a number of research articles, a co-author of three books as well as a co-translator of two monographs from English into Lithuanian and a monograph from Lithuanian into English. She also acts as a member of editorial boards of two research journals.

Assoc. Prof. Dr. Giedrė Valūnaitė Oleškevičienė, affiliated to Mykolas Romeris University, Institute of Humanities, defended her doctoral dissertation "Making Sense of Social Media Use in University Studies" in 2016. Her scientific interests include areas such as methodology of social research, problems of contemporary education philosophy, development of creativity in modern education system, etc. The researcher is also actively involved in research on teaching foreign languages, as well as pursuing scientific interests in linguistics and translation.

Ágnes Abuczki is currently working as a post-doc researcher at the MTA-DE (Hungarian Academy of Sciences – University of Debrecen) Research Group for Theoretical Linguistics, and she will be joining the Károli Gáspár University (Hungary) as a senior researcher in September 2018. She defended her PhD dissertation at the University of Debrecen in 2015 entitled *A Core/Periphery Approach to the Functional Spectrum of Discourse Markers in Multimodal Context*. Her academic interests include Conversation Analysis, Computational Pragmatics and Corpus Linguistics.

PhDr. Šárka Zikánová deals with research of text coherence and the grammatical structure of a language. In her doctoral thesis *The placement of verbal predicates in Older Czech (1500-1620),* she analyzed the relation between word order and information structure. In her current work, she is interested in the interplays of discourse relations, coreference and bridging anaphora, information structure and syntax. She is a senior research associate at the Institute of Formal and Applied Linguistics (Charles University, Prague).