Multilingual resources: discourse relations in English TED talks and their translation into Lithuanian, Portuguese and Turkish

Deniz Zeyrek a,*, Giedrė Valūnaitė Oleškevičienė b and Amália Mendes c

a Cognitive Science Dept., Middle East Technical University, Ankara, Turkey
E-mail: dezeyrek@metu.edu.tr
b Institute of Humanities, Mykolas Romeris University, Vilnius, Lithuania
E-mail: gvalunaite@mruni.eu
c Center of Linguistics, University of Lisbon, Portugal
E-mail: amaliamendes@letras.ulisboa.pt

Abstract. The creation of multilingual resources is of key importance for crosslinguistic research and making such resources accessible for the Linked Data paradigm is a pressing need. The current paper is exploratory, aiming to reveal the potential of an annotated corpus, TED-Multilingual Discourse Bank, for the Linked Data paradigm. The paper focuses on a sample of two TED talk transcripts selected from this corpus in English, the source language, and the translations into three less-studied languages, Lithuanian, Portuguese, Turkish. It examines how low-level coherence is established in English versus target languages, mainly focusing on the comparison of discourse connective classes, explicit versus implicit relations, and the matches between English and the target languages in conveying discourse relations.

Keywords: multilingual corpus, discourse relations, discourse connectives, explicitation, implicitation

1. Introduction

The researchers in the field of LLOD (linguistic linked open data) are working to develop standards and methods for linking linguistic data. Given the various linguistic levels that are annotated and numerous annotation schemes used, linguistically annotated data is a challenge for creating a common infrastructure for LLOD [1]. However, linguistic corpora have demonstrated their vast potential for the Linked Data paradigm. The current paper is exploratory, aiming to reveal the potential of an annotated corpus (TED-Multilingual Discourse Bank, or TED-MDB) for the linked data paradigm. It focuses on a sample of two TED talk transcripts selected from this corpus in English, the source language, and the translations into Lithuanian, Portuguese, Turkish.

Discourse is the level of language above the sentence. TED-MDB is a corpus of 6 TED talks in 7 languages (English, German, Polish, Russian, Turkish, Portuguese, Lithuanian) annotated at the discourse level following the rules and principles of the Penn Discourse TreeBank (PDTB) [2].

Two texts of TED-MDB were investigated with the aim to reveal the patterns of discourse revolving around connectives in the texts of English and three typologically distinct languages. The choice of these three languages was motivated by the fact that they are less studied compared to English, and would benefit from linguistic research as well as technologies applied to the development and sharing of resources. There is significant potential of Linked Data for such languages enabling crosslinguistic analysis to better understand linguistic diversity [3]. Thus, despite its
modest size, one of the values of the paper lies in its empirical value.

Discourse relations are one way of understanding the way clauses or sentences are structured to achieve coherence and are named after the senses they convey (contrast, condition, elaboration, etc.). They are expressed by a range of linguistic devices, including but not limited to syntactic classes such as conjunctions (and, but, so), adverbials (however, moreover, in addition) or prepositional phrases (in summary) [4–6]. (See example (1)). These words or groups of words are generically referred to as discourse operators, discourse markers, cue phrases, or discourse connectives [7], the term used in the present study. Clearly, readers do not need connectives to infer the discourse relation and these cases have been known as implicit relations (example (2)).

(1) Michael went to the store for Christmas shopping. However, he couldn’t find any gifts for his mum.

(2) Michael went to the store for Christmas shopping. He bought a pair of earrings for his wife and toys for his children.

Semantically, discourse connectives express a two-place relation, where the text spans they relate have an abstract object interpretation (eventualities, propositions, facts) as in [8] or are full clauses as in [9]. Elements that have a discourse connective role may also have interactional functions in discourse. (e.g. so is not only a causal discourse connective but also a marker with the function of organizing text segments). Thus, connectives are a heterogenous class of elements with multiple functions. Although this polyfunctionality is not a problem for the present paper, it is a fact on which the annotators were trained.

In the rest of the paper, it is assumed that "there is a definable set of coherence relations, and that the recovery of such coherence relations are essential for comprehension” [10]; also see [7, 11–13]. To examine how coherence is established in English and the translated texts under investigation and reveal the usefulness of the corpus for future research, three research questions were formulated:

(a) What are the lexical classes of discourse connectives in SL vs TLs?
(b) How are explicit and implicit discourse relations distributed to semantic categories in SL vs TLs?
(c) Do the translation of discourse relations match the SL relations? i.e., How often are explicit>implicit translation (implicaiton) and implicit>explicit translation (explicitation) observed in the data? What semantic categories tend to be implicitated or explicitated?

The rest of the paper is organized as follows. Section 2 forms the theoretical background of the paper and outlines the role of connectives in discourse research both from a monolingual and multilingual perspective. Section 3 describes the annotation framework of the corpus with examples, followed by an explanation of the annotation pipeline and the reliability of the corpus reported in a previous paper. Section 4 introduces the two transcripts that were analyzed and explains their characteristics through descriptive statistics. Section 5 answers research question (a), briefly comparing different classes of lexical elements that serve as discourse connectives in each language set. Section 6 and Section 7 address research questions (b) and (c), respectively, and finally, Section 8 concludes the paper with a brief summary.

2. Background: The role of connectives in discourse research

Ever since 1970s, connectives have been exploited as a way to understand the structure of texts [4, 14]. Researchers have addressed a range of issues including the connectives’ role in cognition [15], or the effect of translation on coherence [16]. Lately, with the advance of corpus linguistics and big data, researchers have been interested in investigating connectives in annotated corpora. In a study on the PDTB 2.0, linguistic features were derived from the clauses and local discourse dependencies and a machine learning model that exploits these linguistic structures was built [17]. The training data showed that synchronous, asynchronous, conjunction and contrast relations were explicitly realized in the majority of the time, while instantiation, reason, result and restatement were mostly left implicit. The best model successfully predicted the presence or omission of discourse connectives, demonstrating the empirical value of discourse-annotated corpora for discourse research.

In a recent line of research inspired by Continuity Hypothesis [15, 18], researchers have investigated the reasons that underlie the implicit/explicit realization of connectives by examining linguistic corpora, either monolingual or multilingual. The Continuity Hypothesis holds that “connectives impact on-line processing
to the extent that they signal a text event that represents a departure from the continuity of the events stated in the text” [19]. Thus, in a study on PDTB 2.0 that builds on the Continuity Hypothesis and the causality-as-default hypothesis of [20], it was revealed that discontinuous relations (concession, temporal precedence) tended to be more explicit than continuous relations (causality, temporal sequence, instantiation) [21].

In the domain of contrastive discourse analysis, scholars have long noticed that connectives are highly volatile elements at risk of omission in translation [5], and have looked into the question of why they are dropped. In a paper that made use of the Cognitive Approach to Coherence Relations [22] as the annotation framework, eight connectives were annotated to investigate the features of coherence relations that make them “more or less likely to be conveyed implicitly” [23] (p. 115). The data was a parallel multilingual corpus derived from Europarl translated into Dutch, German, French, Spanish. Based on the notion of complexity of coherence relations in [24] and the evidence from previous literature, it was hypothesized that cognitively simple relations would be prone to implicitation more often than more complex relations. The results showed that complexity affected the linguistic marking of coherence relations, for example, condition relations were less frequently implicit than causal or additive relations, and that some relations types, such as speech-act relations were more likely to be omitted. The data led the authors to conclude that frequently implicated coherence relations did not differ between the languages in the corpus. This finding was in line with the expectation that cognitively complex relations would hold cross-linguistically. Moreover, whether syntactic and relational dependency (one relation containing another relation or sharing a segment with another relation [17]) affected implicit/explicit marking of relations was examined. It was found that relational dependency affected implicit marking of relations.

In another paper, the notion of discontinuity was extended to include all relations involving a form of perspective shift between two related texts [25]. Three polysemous French connectives and their translations into German, English and Spanish were extracted from Europarl, where all the selected connectives conveyed distinct coherence relations. The analyses revealed that relations that can easily be conveyed implicitly were also those that were most often left implicit in translation. For all connectives and relations under investigation, the discontinuous discourse relations (condition, concession and confirmation) led to a higher number of explicit translations compared to continuous relations (causal and temporal).

On the basis of a sample of three TED talks (4720 words, 234 sentences), highly polyfunctional markers and, but, so in Czech, French, Hungarian and Lithuanian were examined by [26] through the lens of underspecification (a notion related to low information value). Frequent omission of the discourse marker in the translated texts was viewed as a sign of low information value, either due to high polyfunctionality (as in the case of and, so) or to a highly bleached meaning (as in the case of now). The researchers concluded that omission targets the same discourse markers in the same ranking in all the languages of the sample stressing the role of underspecification in omitted connectives across languages.

There are similarities and differences between previous research and the present paper. First and foremost, the present paper is not concerned with the reasons why discourse relations are realized implicitly or explicitly in texts. Rather, it examines a small but fully annotated data to understand the overall patterns of discourse relations to find preliminary answers to the research questions above to gain an insight into how low-level coherence is maintained in TED talks. Secondly, the overarching aim of TED-MDB has been to construct an annotated multilingual corpus from scratch as a resource for future investigations of discourse. It differs from [23] in the style of annotation, because TED-MDB not only annotates English but also the TLs in the corpus, and secondly, it includes explicit-to-implicit translation as well as explicit-to-implicit and implicit-to-implicit translation of relations. Different from [23, 25, 26], TED-MDB does not limit the discourse relations to a specific set of markers and their translation. Thus, all texts in TED-MDB are annotated from start to end, an annotation style that reflects the incremental comprehension of texts by readers. The current paper also differs from [27] extending it with a new language, Lithuanian and offering new analyses (the lexical class of connectives across languages, the distribution of relation subsenses across the data, the tendency of certain subsenses for implicature, among others).

3. Annotation framework: The PDTB annotation style

As in PDTB, in TED-MDB, annotation means specifying the discourse connective, if any, marking the
The PDTB sense hierarchy involves four major semantic categories as the top-level senses to tag explicit and implicit discourse relations and AltLexes: Expansion, Temporal, Contingency, and Comparison. Each of these categories is constructed out of subsenses at the level right below. Briefly, Expansion refers to the elaboration relations between two text spans. The category Temporal subsumes time-related eventualities. Contingency relations encompass Cause and Condition relations and their sublevels. Comparison refers to the relations between two eventualities where differences are highlighted. In addition to these categories, a fifth first-level relation referred to as Hypophora has been annotated in TED-MDB. Hypophora refers to questions and the meaningful answers given to them and are taken as alternative means of lexicalizing a relation. Questions and answers are annotated as arguments to Hypophora. EntRels and NoRels are specified and annotated together with their binary arguments. A sense tag is not assigned to them.

The annotation style is bottom-up in the sense that with a set of informative guidelines, the annotators’ inferences and intuitions regarding discourse connectives, relation types, and senses are gathered and reflected in the corpus. The annotation scheme is summarized below.

- Explicitly conveyed discourse relations

Once the annotator infers an explicit relation, his task is to annotate the discourse connective, its binary arguments and assign a sense label to the relation. In TED-MDB, explicit relations are annotated both at the intra-sentential (example (3)) and the inter-sentential level (example (4)).

(3) If your prosthetic socket is uncomfortable, you will not use your leg, and that is just simply unacceptable in our age. (TED Talk ID 1971) (Explicit; Expansion:Conjunction)

(4) I was born and raised in Sierra Leone, a small and very beautiful country in West Africa, a country rich both in physical resources and creative talent. However, Sierra Leone is infamous for a decade-long rebel war in the ’90s when entire villages were burnt down. (TED Talk ID 1971) (Explicit; Comparison:Concession)

- Alternative lexicalizations

Alternative lexicalizations, or AltLexes are connecting devices that fall outside the set of conjunctions, adverbs, or prepositional phrases. They vary from fixed expressions (for this reason), to free expressions (one of the most vivid examples is) and annotated both inter- and intra-sententially.

(5) Success motivates us, but a near win can propel us in an ongoing quest. One of the most vivid examples of this comes when we look at the difference between Olympic silver medalists and bronze medalists after a competition. (TED Talk ID 1978) (AltLex; Expansion:Instantiation)

- Implicit discourse relations

As in the PDTB scheme, implicit relations are annotated by providing an explicit signal that best expresses the sense of the relation. They are annotated at the inter-sentential level, leaving intra-sentential implicit relations for further research.

(6) I kept a packed bag with some clothes and cans of food tucked away in the back of a closet. There was a deep restlessness in me a primal fear that I would fall prey to a life of routine and boredom. (TED Talk ID 1971) (Implicit = because: Contingency:Cause)

- Entity relations (EntRels)

These involve cases “where no discourse relation can be inferred and where the second sentence only serves to provide some further description of an entity in the first sentence” [29]. Once an EntRel is inferred, it is simply marked together with its arguments. EntRels are sought and annotated inter-sententially.
Another example is Pentair. Pentair is a U.S. industrial conglomerate . . . (TED Talk ID 1971) (EntRel)

– No relation

The tag no relation (NoRel) is used for the cases where no semantic relation can be inferred between adjacent sentences though the sentences are interpretable within the broader discourse context (e.g. topic shifts and sentences at text boundaries). Similarly to EntRels and implicit relations, NoRels are sought and annotated inter-sententially.

The Paris Review got it out of James Baldwin when they asked him, “What do you think increases with knowledge?” and he said, “You learn how little you know.” Success motivates us, but a near win can propel us in an ongoing quest. (TED Talk ID 1978) (NoRel)

4. Data and methodology

While the PDTB includes annotations on written texts (selections from the Wall Street Journal) [2], TED-MDB encompasses TED talk transcripts which display the characteristics of both spoken and the written genre. The annotation pipeline (of the entire corpus) started by studying the PDTB annotation guidelines and adapting them to TED talks. The annotations on all six talks were performed by monolingual teams, which comprised minimally an expert on discourse and a trained annotator. Each monolingual annotator went over the data file by file, annotating relations in each adjacent clause and sentence. In order to avoid the effect of SL on TLs, annotations were performed independently by each monolingual expert, then discussed in multilingual group meetings for consistency.

To evaluate the reliability of the corpus, ~ 20% of the data was annotated by the second annotator of each monolingual team, and inter-annotator agreement was measured for each language set on discourse relation spotting, discourse relation type, and top-level sense separately. For a corpus of multiple languages, a perfect agreement is not easy to achieve as the relations and senses vary with respect to the language. Nevertheless, across languages, an F-score of 0.73-0.84 was achieved for DR spotting. Cohen’s Kappa (κ) values of 0.70-0.92 were obtained for DR types and 0.71-0.86 for top-level sense, respectively as explained in previous work [27].

The data for the present paper includes two talks randomly chosen from TED-MDB on different topics presented by different speakers. The first talk portrays nomads’ life through the eyes of a documentary.
photographer, while the second talk describes the social map of a city. After the annotation procedure has finished, these two talks were extracted, automatically aligned, then manually checked and revised where necessary.

The data consists of 5805 words, distributed over two talks and four languages (Table 2). For each TL, 121 relations were aligned with English, amounting to a total of 484 relations. Table 3 provides information on the frequency of annotated relation types per language. Due to translation differences, certain relations were not annotated, shown as no annotation, which range between 3-8% in the TLs.

Chi square analyses were run to understand whether two transcripts differed from each other in terms of explicit relations, implicit relations, AltLexes and non-discourse relations (EntRels, NoRels, no annotation cases). The results showed that in English and Turkish, the two texts did not significantly differ from each other (p > 0.50) and the effect size was weak or negligible (Cramer’s V test, < 0.2). In Portuguese and Lithuanian, the difference between the two texts was significant (p < 0.50), but the effect size was weak or negligible (Cramer’s V test, < 0.2). These results show that in all language sets, the difference between talks ID 2009 and 2150 is trivial.

5. The lexical class of connectives in the data

This section attempts to answer research question (a), examining the lexical class of discourse connectives in SL vs TLs.

The distribution of discourse connective types and tokens was examined (Appendix B). This analysis revealed *and* and the equivalent devices in TLs as the most frequent discourse connective token in each language. The frequency of *and* was followed by *but* and its equivalents in English, Lithuanian and Portuguese. In Turkish, *ama* ‘but/yet’ was not as common as *but* in other languages since two other contrastive/concessive connectives with similar meanings (*ancak* ‘however’, *fakat* ‘but’), typically found in formal registers were also present in the data. This may be a tendency of the translators to formally present scientific facts to the audience, or a general characteristic of the TED talk genre that is worth further scrutiny.

Table 4 displays the distribution of the grammatical class of discourse connectives in the data and shows that in SL and TLs, conjunctions were used more often than other discourse connective types. While Portuguese and Turkish made use of adverbials and other connective types, Lithuanian did not resort to such devices often, at least in the data.

The examination of the lexical class of discourse connectives is likely to indicate two things: the translations of explicit relations follow the SL closely, or they make use of different linguistic cues. Lithuanian has the highest number of conjunctions among TLs and it does not seem to choose connectives from other lexical classes often. The number of Lithuanian conjunctions comes close to the number of SL conjunctions. In other words, the English-Lithuanian pair seems to have the closest translation pairs, at least in what concerns explicit relations.

Turkish translations of explicit relations differ from Lithuanian translations and are illustrative of how translators make use of the range of linguistic cues available in the language rather than choosing direct translations of the discourse connectives in SL texts. Turkish is an agglutinating language where suffixation plays an important role in word derivation and inflection, which also affects inter-clausal cohesive links. The dispreference of conjunctions in conveying discourse relations indicates that, as opposed to other TLs, Turkish prefers devices belonging to different lexical classes, such as suffixal connectives subsumed under the other category in Table 4 (*-arak* ‘by’). These suffixal connectives, which are syntactically convertors, constitute the second most common discourse connective type after conjunctions in the Turkish subset and illustrates the way languages differ in conveying explicit relations. It is worth noting that this kind of inter-clausal discourse connectivity is functionally similar to that of other languages but due to their morphological characteristics, they resist string-to-string matching.
Table 3
Number of relations per type of relation and per language

<table>
<thead>
<tr>
<th>Type</th>
<th>ENG %</th>
<th>LI %</th>
<th>PT %</th>
<th>TR %</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explicit</td>
<td>62</td>
<td>51.24</td>
<td>50</td>
<td>41.32</td>
<td>44</td>
</tr>
<tr>
<td>Implicit</td>
<td>34</td>
<td>28.10</td>
<td>44</td>
<td>36.36</td>
<td>53</td>
</tr>
<tr>
<td>AltLex</td>
<td>8</td>
<td>6.61</td>
<td>6</td>
<td>4.96</td>
<td>4</td>
</tr>
<tr>
<td>EntRel</td>
<td>8</td>
<td>6.61</td>
<td>10</td>
<td>8.26</td>
<td>9</td>
</tr>
<tr>
<td>NoRel</td>
<td>8</td>
<td>6.61</td>
<td>3</td>
<td>2.48</td>
<td>7</td>
</tr>
<tr>
<td>No annot.</td>
<td>1</td>
<td>0.83</td>
<td>8</td>
<td>6.61</td>
<td>14</td>
</tr>
<tr>
<td>Total</td>
<td>121</td>
<td>100</td>
<td>121</td>
<td>100</td>
<td>121</td>
</tr>
</tbody>
</table>

Table 4
Frequency of connective classes in the data (Other: phrases, suffixal connectives)

<table>
<thead>
<tr>
<th>Type</th>
<th>ENG Type</th>
<th>ENG Token</th>
<th>LI Type</th>
<th>LI Token</th>
<th>PT Type</th>
<th>PT Token</th>
<th>TR Type</th>
<th>TR Token</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conjunction</td>
<td>10</td>
<td>54</td>
<td>12</td>
<td>48</td>
<td>7</td>
<td>33</td>
<td>8</td>
<td>32</td>
</tr>
<tr>
<td>Adverb</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Adposition</td>
<td>3</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>7</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Other</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>Total</td>
<td>17</td>
<td>62</td>
<td>14</td>
<td>50</td>
<td>15</td>
<td>44</td>
<td>14</td>
<td>49</td>
</tr>
</tbody>
</table>

6. Distribution of explicit and implicit relations into semantic categories

In an attempt to provide an answer to research question (b), this section examines the data globally. It first provides the distribution of discourse relation types across languages, then turns to the distribution of explicit and implicit relations to the top- and second-level semantic categories of the PDTB sense hierarchy in SL vs TLs. The third level of senses are left out of scope of the current paper as the data is too sparse.

Table 3 shows that the majority of the relations are explicitly or implicitly conveyed in the data, with explicit relations occurring more often in all languages except Portuguese. Given that EntRels and NoRels are non-discourse relations [30], they are expected to occur less often than explicit and implicit relations. Indeed, this is what we find across languages: EntRels and NoRels occur at 13.43% on average in the sample.

Table 5 displays the distribution of explicit and implicit relations to top-level semantic categories, and Figures 2 and 3 complement this table by visually representing the distribution of explicit relations and implicit relations to second-level semantic categories, respectively.

Figure 2 demonstrates that across languages, explicit Conjunction relations appear more often than the other explicit relations. Figure 3 shows that across languages, implicit Conjunction and implicit Level-of-detail tend to occur more often than the other implicitly conveyed senses, and Portuguese stands out as it conveys these senses by implicit means more often than the other languages. A comparison of Figures 2 and 3 reveals that in all languages, Level-of-detail is more often implicit than explicit.

In summary, this analysis shows that Conjunction can be realized with or without a connective; neither strategy would lead to loss of coherence. On the other hand, Level-of-detail tends to be expressed without a connective more often than with a connective.

7. Explicit-to-implicit and implicit-to-explicit translation (implicitation and explicitation)

This section zooms in the matches between English and TL relations in translations and aims to answer research question (c).

It has been shown by [31] that during translation, the realities of a situation, the realities of the context, as well as language- and culture-specific issues need to be taken into consideration. At the level of discourse relations, different translators of the same language or translators of different languages may tend to organize the “realities of the situation” and “the realities of the context” in different ways.

Table 6 involves a subset of the relations already provided in Table 3 and presents the matches be-
Table 5

Distribution of explicit relations (E) and implicit relations (I) to semantic categories across languages (T: The total number of sense-assigned relations.)

<table>
<thead>
<tr>
<th></th>
<th>ENG</th>
<th>LI</th>
<th>PT</th>
<th>TR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>E</td>
<td>I</td>
<td>T</td>
<td>E</td>
</tr>
<tr>
<td>COMPARISON</td>
<td>14</td>
<td>4</td>
<td>18</td>
<td>13</td>
</tr>
<tr>
<td>TEMPORAL</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>CONTINGENCY</td>
<td>15</td>
<td>6</td>
<td>22</td>
<td>14</td>
</tr>
<tr>
<td>EXPANSION</td>
<td>32</td>
<td>24</td>
<td>58</td>
<td>20</td>
</tr>
<tr>
<td>Total</td>
<td>62</td>
<td>34</td>
<td>100</td>
<td>50</td>
</tr>
</tbody>
</table>

Fig. 2. Distribution of explicit DRs to second-level semantic categories across languages

Fig. 3. Distribution of implicit DRs to second-level semantic categories across languages
The volatility of connectives such as *and* is well-attested in the literature. *And* typically conveys the additive sense but in different contexts, it may express temporal or causal senses. Due to this kind of underdetermination [32], or underspecified meaning [26], it is likely to be affected by contextual changes in translation and inclined to omission [5]. Note that the conjunction *and* was found as the most frequent discourse connective in the corpus. This result should not contradict with the *and*-omissions in the data: *and* tends to be omitted but still remains a frequent explicit device conveying Expansion sub-senses. To further investigate the implication phenomenon and *and*-omission, looking into co-occurring multiple connectives could give further insights.

In English, an adverb (*otherwise, instead*) frequently co-occurs with a conjunction (*because, if*). These are called concurrent relations and exist in many languages, as in English [33–35], Catalan and Spanish [36], as well as Turkish [37]. There are very few co-occurring multiple connectives, or CMC instances in the data; yet, how they are captured in translation could be interesting from a crosslinguistic perspective because this raises the question of whether CMCs are kept in translation or undergo implication partly or totally.

(9) **EN** I’m here on the green side, down on the far right where the geeks are and TEDx also is down on the far right (Explicit, Expansion:Conjunction)  
**LT** Aš esu žaliajo pusėje, toli dešinėje su kitaikais moksliaukais, [ Implicit= ir *and* ] TEDx irgi ten dešinėje (Explicit, Expansion:Conjunction)  
**PT** Eu estou aqui de o lado verde , no extremo direito , onde estão os " geeks " [ Implicit= e *and* ] O TEDx também está em baixo , no extremo direito (Explicit, Expansion:Conjunction)

The adverb *otherwise* was found as the most frequent discourse connective in the corpus. This result should not contradict with the *and*-omissions in the data: *and* tends to be omitted but still remains a frequent explicit device conveying Expansion sub-senses.

To further investigate the implication phenomenon and *and*-omission, looking into co-occurring multiple connectives could give further insights.
Table 6
Explicit relations in English that correspond to various discourse relations in the TLs

<table>
<thead>
<tr>
<th></th>
<th>ENG (Exp.)</th>
<th>LI %</th>
<th>PT %</th>
<th>TR %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Talk ID: 2009 (N =24)</td>
<td>Explicit 16</td>
<td>25.81</td>
<td>21</td>
<td>33.87</td>
</tr>
<tr>
<td></td>
<td>Implicit 5</td>
<td>8.06</td>
<td>2</td>
<td>3.23</td>
</tr>
<tr>
<td></td>
<td>AltLex 1</td>
<td>1.61</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>EntRel 0</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>NoRel 0</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>No annot. 2</td>
<td>3.23</td>
<td>1</td>
<td>1.61</td>
</tr>
<tr>
<td>Talk ID: 2150 (N =38)</td>
<td>Explicit 32</td>
<td>51.61</td>
<td>20</td>
<td>32.26</td>
</tr>
<tr>
<td></td>
<td>Implicit 5</td>
<td>8.06</td>
<td>16</td>
<td>25.81</td>
</tr>
<tr>
<td></td>
<td>AltLex 0</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>EntRel 0</td>
<td>0.00</td>
<td>2</td>
<td>3.23</td>
</tr>
<tr>
<td></td>
<td>NoRel 0</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>No annot. 1</td>
<td>1.61</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>Total</td>
<td>62</td>
<td>100.00</td>
<td>62</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Fig. 4. Explicit-to-explicit vs. explicit-to-implicit translation across TLs

Fig. 5. Implicit-to-implicit vs. implicit-to-explicit translation across TLs
The CMC and so is considered as a possible case of addition (less integration) or composition (more integration), depending on the context [36]. In cases of addition, and so has the meaning 'and as a consequence', while in the case of composition, it indicates discourse re-orientation. In example (10), the CMC is a case of addition: each connective has its scope, and the meaning is addition + consequence.

Cases of addition was expected to be annotated with two senses, and this is what happened in English. It was also expected that each connective could be either kept or omitted independently of the other in the TLs.

In translating and so, the Lithuanian translator used only one connective, the equivalent to and, and omitted so. In Portuguese, both connectives are missing entirely in the translation, although an implicit Contingency sense is inferred. In Turkish, the translator captured the relation in three sentences rather than two, and this relation did not get aligned with the original text. However, three sentences were linked to each other implicitly and the translation bears evidence for another omitted case of and.

These instances raise the possibility that the omission of and takes place regardless of where it appears; i.e., both when it is used as a single connective or as part of a CMC.

In example (11), the English version but then responds to the criteria of the composition cases: it expresses a single meaning rather than two adjoined meanings (concession + temporal/consequence). It was annotated as a single connective, where but was taken to convey a unique sense (concession). But then introduces a subtopic that the speaker wishes to introduce in comparison to the previous comment. The new subtopic was undoubtedly why the Lithuanian translation kept the connective equivalent to but (bet), omitted then, and the annotator annotated the example with Expansion:Arg2-as-detail. In Turkish, but then was translated by means of a CMC fakat ... ayrıca, where fakat functions to move the discourse forward, and ayrıca encodes the importance of the new subtopic. The sentence was annotated as conveying a dual meaning, namely Concession and Conjunction senses. In Portuguese, but then was translated through a single connective e ‘and’ (well in accordance with the type composition) and it was labelled with the sense Expansion:Conjunction. So, the translations and annotations of this relation tend to consider but then as expressing a move forward in the discourse.

These limited examples show that what is a CMC in English might well be translated to the TL as a single explicit connective or as an implicit relation. Thus, in the future, it would be worthwhile to consider CMCs as potential signals for implicitation.

Finally, most SL implicit relations were translated to SL texts implicitly; in other words, the data has very few implicit>explicit translations (explicitation) (See Figure 5). Although not decisive, this result echoes [38], where the author did not find support for the explicitation theory of [16] in Norwegian-English translations.

To sum up, the analysis in this section showed that TLs resorted to implicitation at varying degrees. Portuguese had the highest number of implicitations among TLs, but the number of implicitations in each transcript were not the same.

8. Conclusion

In order to understand the nature of low-level coherence in multilingual data, this paper investigated three TLs in comparison to English in a small dataset. The data comprised two TED talks from TED-MDB that were as comparable as could be, hence keeping the risks related to unbalanced distribution of relations low.

The major contribution of the paper has been to reveal the importance of the following aspects of the TED talks under investigation:

- There was a tendency for all TLs to translate explicit relations as explicit relations. Regarding the lexical class of discourse connectives, conjunctions were used often in all languages. While Lithuanian did not seem to prefer connecting devices from other classes, Turkish and Portuguese used conjunctions as well as connecting devices from other classes. This result indicates that Lithuanian translations are close to English in rendering explicit relations.
Across TLs, Conjunction relations tended to be rendered either explicitly or implicitly. In conveying Level-of-detail, all languages favored implicit means.

Portuguese exhibited the highest number of implicitations in the data, particularly in one of the transcripts. In Portuguese, Conjunction relations were more often implicitated than other languages (and other senses). This finding indicates that there is variability in the number of implicitation both across languages and within the translations of the same language. The number of explicitations was negligible in the data.

These results have limited generalizability due to the small sample used. Secondly, although professionals check the translations, semi-trained or untrained translators provide the translations, and high quality of translation is not ensured. Subtitling, which is a specific translation technique [39] used in translating the transcripts of TED talks, has limitations of its own. Such methodological issues are likely to affect the translation of different talks in different languages, and might lead to annotation differences. Nevertheless, by examining two TED talk transcripts from TED-MDB, it is hoped that the value of the richly annotated discourse-level multilingual corpus has been shown for future research. Linked data based on aligned multilingual corpora would ensure comparability between languages. Once the discourse relations and their sense tags in TED-MDB have been aligned, the corpus would be perfect to compare aspects of discourse across languages. The aim for the future is to complete the relation alignment procedure and refine the multilingual corpus to fit the Linked Data paradigm.

### Appendix A. PDTB3 Sense hierarchy

<table>
<thead>
<tr>
<th>Temporal</th>
<th>Comparison</th>
<th>Contrast</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asynchronous</td>
<td>Precedence</td>
<td>Simultaneity</td>
</tr>
<tr>
<td>Synchronous</td>
<td>Result</td>
<td>A permanent cause</td>
</tr>
<tr>
<td>Contingency</td>
<td>Result</td>
<td>A probable cause</td>
</tr>
<tr>
<td>Cause</td>
<td>Reason</td>
<td>An necessary condition</td>
</tr>
<tr>
<td>Purpose</td>
<td>Reason</td>
<td>An accidental or fortuitous cause</td>
</tr>
<tr>
<td>Condition</td>
<td>Result</td>
<td>An evident cause</td>
</tr>
<tr>
<td>Condition</td>
<td>Reason</td>
<td>An apparent cause</td>
</tr>
<tr>
<td>Condition</td>
<td>Result</td>
<td>An adequate reason</td>
</tr>
<tr>
<td>Negative Condition</td>
<td>Reason</td>
<td>An inadequate reason</td>
</tr>
<tr>
<td>Negative Condition</td>
<td>Result</td>
<td>An insufficient cause</td>
</tr>
<tr>
<td>Negative Condition</td>
<td>Result</td>
<td>An insufficient reason</td>
</tr>
</tbody>
</table>

### Appendix B. Connective types and tokens in the data

<table>
<thead>
<tr>
<th>Lang.</th>
<th># of Types</th>
<th># of tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENG</td>
<td>17</td>
<td>1</td>
</tr>
<tr>
<td>LI</td>
<td>14</td>
<td>1, 20</td>
</tr>
<tr>
<td>PT</td>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td>TR</td>
<td>14</td>
<td>4, 6, 20</td>
</tr>
</tbody>
</table>

### References


