BRILL

CONTRASTIVE PRAGMATICS 3 (2022) 452–479

CONTRASTIVE PRAGMATICS
A Cross-Disciplinary Journal
brill.com/jocp

An Exploratory Analysis of TED Talks in English and Lithuanian, Portuguese and Turkish Translations

Results from the Analysis of an Annotated Multilingual Corpus

Deniz Zeyrek | ORCID: 0000-0001-9248-0141
Professor, Graduate School of Informatics, Middle East Technical University, Department of Cognitive Science, Ankara, Turkey
Corresponding author
dezeyrek@metu.edu.tr

Amália Mendes | ORCID: 0000-0001-6815-2674
Associate Professor, Center of Linguistics, School of Arts and Humanities, University of Lisbon, Lisbon, Portugal
amaliamendes@letras.ulisboa.pt

Giedrė Valūnaitė Oleškevičienė | ORCID: 0000-0001-5688-2469
Associate Professor, Institute of Humanities, Mykolas Romeris University, Vilnius, Lithuania
gvalunaite@mruni.eu

Sibel Özer | ORCID: 0000-0002-2202-3399
PhD student, Graduate School of Informatics, Middle East Technical University, Department of Cognitive Science, Ankara, Turkey
sibel.ozer@metu.edu.tr

Abstract

This paper contributes to the question of how discourse relations are realised in TED talks. Drawing on an annotated, multilingual discourse corpus of TED talk transcripts, we examine discourse relations in English and Lithuanian, Portuguese and Turkish translations by concentrating on three aspects: the degree of explicitness in discourse relations, the extent to which explicit and implicit relations are encoded inter- or intra-sententially, and whether top-level discourse relation senses employed in English differ

in the target languages. The study shows that while the target languages differ from English in the first two dimensions, they do not display considerable differences in the third dimension. The paper thus reveals variations in the realisation of discourse relations in translated transcripts of a spoken genre in three languages and offers some methodological insights for dealing with the issues surrounding discourse relations.

Keywords

discourse relations – TED talks – annotation – multilingual corpus – translation

1 Introduction

Discourse is the language level that goes beyond sentences, though it can also be found within a sentence.[1] Discourse conveys more than the sum of its parts, which is partly due to the semantic or pragmatic relations that hold between (or within) the individual sentences that comprise the discourse. These relations are known as discourse relations (DRs) and are an essential aspect of discourse structure (Asher and Lascarides, 2003; Fraser, 1999; Kehler, 2002; Mann and Thompson, 1988). The semantic content of discourse relations (Expansion, Cause, Condition and the like) may be signalled by various linguistic cues, such as discourse connectives, or expressed implicitly, where features such as the adjacency of discourse units, lexical relations and anaphoric links signal the discourse relation. To interpret texts, readers need to uncover discourse relations by recognising these clues and to infer the sense of the discourse relation. This paper contributes to the growing literature on discourse relations, and, by drawing on an annotated, multilingual discourse corpus of TED talk transcripts, it aims to throw light on the manifestations of discourse relations in English and Lithuanian, European Portuguese and Turkish translations in an oral genre. It compares and contrasts these languages on the basis of three dimensions: the degree of explicitness in discourse relations, whether discourse relations are encoded inter- or intra-sententially, and whether discourse relation senses expressed in English differ in the target languages. It is assumed that these dimensions interact with each other in structuring the discourse of TED talks at a low level, and comparing them in a multilingual corpus can give hints about the means by which local coherence is established in the languages under examination.

1 We use the terms *discourse* and *text* synonymously though they can be differentiated by defining *discourse* as the process of purposeful, communicative activity and *text* its product.

Parallel or translated corpora[2] are much-needed resources for all levels of linguistic analysis as they enable researchers to understand cross-linguistic differences and to pinpoint linguistic phenomena that are not directly observable in monolingual corpora. As argued by Mauranen (1999: 161), "a parallel corpus can capture relations of sense as well as form, which would be very hard to capture without such data". But parallel corpora in discourse and pragmatics are still scarce, particularly in regard to datasets involving multiple languages. Moreover, most existing parallel corpora not only involve resource-rich languages but are also drawn upon formal, written language, which could be a limited representation of the spectrum of variation in discourse and pragmatics (but see the small TED talk corpus in Crible et al., 2019).

This paper exploits an existing resource, the TED-Multilingual Discourse Bank or TED-MDB, a freely available annotated discourse corpus of TED talks in English and their translations into five languages (German, Polish, Russian, European Portuguese and Turkish) (Zeyrek et al., 2020). It was developed as a joint effort by an international team of researchers following the rules and principles of the Penn Discourse Treebank (PDTB) (Prasad et al., 2008), a 2-million-word resource that has the goal of capturing aspects of discourse structure by annotating discourse relations in Wall Street Journal texts. The TED-MDB expands the PDTB framework to TED talks by systematically annotating multiple languages with a common descriptive paradigm. In this way, a resource is created for discourse analysis in English and translations into various languages in a spoken genre. The current corpus can be used by linguists, computational linguists, discourse analysts, translation experts or teachers for pedagogical purposes.

We focus on two relatively less-resourced languages in the TED-MDB that have been described by Zeyrek et al. (2020) (Turkish and Portuguese) and a third language recently added to the corpus, Lithuanian. Two factors guided our choice of these languages besides English: Firstly, in our earlier work (on all languages except Lithuanian) over *two texts* of the TED-MDB, we found that the percentage of implicit discourse relations among the language sets placed English and Turkish at one end of the spectrum, and Portuguese at the other end. But the tendency shown by Portuguese or Turkish for implicitation, a chief parameter that could inform the strength of explicitness in translated corpora, was not examined in detail. Secondly, since Lithuanian was not part of the corpus at that time, it was not analysed. The current work intends to extend the initial findings by re-examining the ways in which discourse relations are realised not only in Turkish and Portuguese but also in Lithuanian

2 These terms are used interchangeably in the paper to mean corpora which consist of texts in one language and translations into other languages.

from a broader perspective based on translated and annotated data from all *six* TED transcripts in the TED-MDB.

A complete delineation of discourse structure in the genre of TED talks is not the aim; instead, the current research intends to describe the behaviour of English and three translated languages in conveying discourse relations in a spoken genre in general, and more particularly, the extent to which discourse relations differ in terms of the three dimensions introduced above in a multilingual corpus of TED talks. Thus, with a frame that sees TED talks as a particular type of oral speech, the following research questions are asked:

– RQ1: Do target texts differ from English and each other in conveying discourse relations explicitly?
– RQ2: Do target texts differ from English and each other in encoding discourse relations inter-sententially versus intra-sententially? (i.e., is information distributed across sentences or gathered in a single sentence across languages in the corpus)?
– RQ3: Do target languages retain the annotated top-level sense of corresponding discourse relations?

The layout of the rest of the paper is as follows. We start with a brief overview of work related to the discourse phenomena under investigation (Section 2). Then, in Section 3, we present the discourse relation realisation types that are recognised by the PDTB framework. Section 4 provides a synopsis of our major annotation categories and the annotation procedures, and offers an evaluation of the annotated corpus. The section continues with the technique of assigning senses to discourse relations and ends with data alignment and linking methods. Section 5 presents the results, and Section 6 summarises and concludes the paper.

2 A Brief Overview of Related Work

The aim of this section is to provide an overview of how the explicitness of discourse relations is tackled in the literature and to summarise a few representative works dealing with the omission of discourse connectives in translation. Rather than presenting a comprehensive review, the aim is to provide background for the discourse phenomena explored in the current work.

2.1 *Explicitness*

Explicitness is a phenomenon referring to the use of overt linguistic material in structuring information in clauses. In the current paper, it is used in a strict sense referring to the realisation of discourse relations through discourse connectives.

In one of the earliest works, Asr and Demberg (2012) investigated the reasons that underlie the implicit versus explicit realisation of discourse relations by examining the PDTB. Their research was inspired by the Continuity Hypothesis, which holds that "connectives impact online processing to the extent that they signal a text event that represents a departure from the continuity of the events stated in the text" (Murray, 1997: 227). Building on the Continuity Hypothesis and the causality-as-default hypothesis of Sanders (2005), the researchers revealed that discontinuous relations (e.g., Adversatives) tended to be conveyed explicitly, as opposed to continuous relations (e.g., Cause), which tended to be expressed implicitly. This work showed that the degree of explicitness conveyed by discourse relations could vary even within a single language, depending partly on continuity. A study by Maier et al. (2016) dealt with the argumentative discourse genre of commentary. With a focus on cues such as referential continuity, lexical chains and the use of discourse connectives, the authors characterised a group of relations (e.g., Contrast and Corrective Elaboration) as having a high degree of 'glueyness'. Other relations, such as Elaboration and Result, differed from the first group as they tended to be conveyed implicitly. This study suggested that discourse genres can affect the overtness of discourse relations. Hofmockel et al. (2017) continued to investigate the effect of genre on the overtness of discourse relations and examined argumentative and narrative discourse. They observed that in the argumentative genre, Explanation, Continuation and Comment tended to be realised with a lower degree of overtness and Elaboration with a higher degree of overtness. Beyond pointing out the effect of discourse genre on the explicitness of discourse relations, these works share a common denominator in showing that relation semantics has a role in the explicitness of discourse relations. Another study in this line of research is the work conducted by Das and Taboada (2018), which also emphasises the role of relation semantics in the explicitness of discourse relations. By annotating a corpus that followed the principles of the Rhetorical Structure Theory (Mann and Thompson, 1998), the authors found that most relations in the corpus were signalled by discourse markers (*and*, *but*, *since*) and that most of those signalled relations had additional signals such as referential links and semantic, syntactic and graphical features.

Recent years have seen an upsurge in the analysis of discourse phenomena through multilingual corpora. In a study that draws on translated corpora, Steiner (2015) compared and contrasted English and German in terms of several dimensions, such as the different orientations of discourses towards explicitness and 'information density', a term used by Fabricius-Hansen (1996) (also

see Section 3.3). The study found that German texts tended to encode not only very explicitly but also with high density conjunctive relations (e.g., Temporal, Adversative) and used adverbials rather than paratactic constructions. Like Steiner, House (2015) adopted a corpus-based approach and compared translations from popular English science texts to German in terms of 'linking constructions', which are multi-word lexico-grammatical patterns such as *in addition* and *after all*. She compared linking constructions from a diachronic perspective, and revealed the range of translation equivalents of the linking constructions used in English and German translations in different periods. Her research showed that while linking constructions were typical of English texts, they were either translated by a zero form or by syntactic integration into German. House attributed such differences to deep-seated syntactic differences between the two languages.

2.2 *Omission of Connectives*

Since the 1970s, discourse connectives have been exploited to understand the structure of texts (Halliday and Hasan, 2014; Schiffrin, 1986). In coherence-based theories, they are known as cues that make the discourse relation salient, facilitating the inference of the discourse relation (Fraser, 1999; Kehler, 2002; Knott and Dale, 1994; amongst others). Translation scholars, on the other hand, have long noticed that connectives are highly volatile elements which risk omission in translated texts (Halverson, 2004). This has led to the question of why connectives are dropped in translation. To name a few works on this topic, in a paper that draws on the notion of (dis)continuity (Murray, 1997), it was observed that discontinuous discourse relations such as Condition and Concession led to a higher number of explicit translations compared to continuous relations (Zufferey, 2016). A comprehensive study (Hoek et al., 2017: 114) argued that cognitive complexity affected the linguistic marking of coherence relations. That is, "if an unexpected relation [such as Condition] is not marked, its inference requires too much effort for the resulting cognitive effect". The authors argued that for such relations, there would be much less room for variation in translation. Examining highly polyfunctional markers *and, but* and *so* in English and translations into multiple languages in TED talks, Crible et al. (2019) argued that the omission of connectives could be explained by the concept of underspecification (a notion related to low information value) and the function of the original marker across languages.

The omission of discourse connectives in translation will be considered again, although only very briefly, in Section 5 (Results and Analysis), when the explicitness dimension of discourse relations is analysed. The potential effect

of the Continuity Hypothesis or relation semantics on the explicitness of discourse relations in TED talks, on the other hand, is beyond the scope of our analysis.

3 The PDTB Framework

This section first introduces the different realisations of discourse relations that are recognised by the PDTB framework (Section 3.1), describes how we analyse discourse relation senses (Section 3.2) and explains how we record the inter- and intra-sentential encoding of discourse relations in the multilingual corpus (Section 3.3).

3.1 *Realisation of Discourse Relations*

The PDTB considers discourse connectives as lexico-syntactic cues that signal the presence of a discourse relation. Semantically, discourse connectives express a two-place relation, where the text parts they relate to have an abstract object interpretation (eventualities, propositions, facts), as explained by Asher (1993). These text parts, i.e., the constitutive units of discourse relations, are referred to as arguments, and coherence results from inferring a semantic relation between the arguments. The PDTB recognises six ways in which discourse relations are realised.

Explicit Discourse Relations: An explicitly conveyed discourse relation is signalled by discourse connectives such as subordinating or coordinating conjunctions or adverbs, as in example (1). The discourse connective (shown in small capitals) fully specifies the sense of the discourse relation.

(1) The child is crying BECAUSE he is hungry.

Implicit Discourse Relations: The implicitness of discourse relations is a notion based on adjacency. An implicit discourse relation can be inferred by the mere adjacency of discourse units, where linguistic material in the clauses/sentences offers cues about the discourse relation sense (Fabricius-Hansen, 1996; Kehler, 2002). Moreover, an implicit discourse relation can easily be rephrased by a discourse connective to confirm the conveyed sense; for instance, example (2) could be paraphrased by *instead*. These are referred to as 'implicit connectives' in the PDTB framework.

(2) We haven't seen any of the planets. We've only detected them indirectly.

Alternative Lexicalisation, Entity Relation, No Relation: The PDTB distinguishes those cases where an implicit discourse relation could be rephrased by a connective from other cases where it could not due to the presence of some other expression that conveys the relation. These expressions are called 'alternative lexicalisations' (AltLex) (Prasad et al., 2010), which involve lexically frozen expressions (*quite the contrary, what's more*) as well as syntactically and lexically free ones (*this is why*), as shown in examples (3) and (4). The recognition and analysis of such terms enable us to see that discourse relations can be anchored by devices beyond syntactic classes.

(3) We are the best; THIS IS WHY we won the game!
(4) Most people resort to art AS A WAY to find meaning and purpose in life.

An indispensable aspect of coherence is the mechanism that relates NP referents to subsequent clauses (Karamanis, 2007; Knott et al., 2001). In the PDTB framework, this is called an *entity relation* (EntRel). They are taken as implicit relations at the inter-sentential level, where an entity is introduced in the first sentence and discussed in the following sentence (example (5)). One distinction between implicit discourse relations and EntRels is that while implicit discourse relations can be rephrased by using a connecting device, EntRels would often sound unnatural with one.

(5) Art is a diverse range of human activity. It involves creative talent, technical skills or emotional power.

There may be cases where neither an implicit discourse relation nor an entity-based relation can be inferred between a pair of adjacent sentences. These cases are known as 'no relation' (NoRel), and indicate that the establishment of coherence has failed. They should be distinguished from other cases where coherence is maintained. In oral language, speakers can switch to a new topic by either signalling the shift with an overt marker such as *now*, *so*, *well*, or implicitly without a particular marker. In an approach like our own, where differently realised discourse relations are scrutinised, they can be considered to be NoRels (see example (6)). In a framework that aims to account for global discourse features, it is perfectly plausible to analyse them as functional elements of discourse.

(6) That's four billion middle-class people demanding food, energy, and water. Now, you may be asking yourself, are these just isolated cases?

Additionally, in speech presented in the presence of a live audience, the speakers can move between an impromptu mode and an expository mode to make the speech livelier (as in example (7)). These shifts can also be considered as NoRels. Still, in a different framework that aims to capture the interactional features of speech, they could well be characterised differently, highlighting their specific function.

(7) You gotta love the Aussies, right? CalPERS is another example.

3.2 *Discourse Relation Senses*

Whether or not there exists a finite list of the possible semantic relations that link discourse segments is still an unresolved issue (Asher, 1993). Nevertheless, researchers have agreed on identifying a specific group of useful senses to study the common properties of discourse relations at various levels of discourse structure. The PDTB 3.0 introduces a sense tagset which is organised hierarchically (Webber et al., 2019). It entails four major semantic categories as the top-level, alternatively referred to as Level-1 senses: Expansion, Temporal, Contingency and Comparison. Briefly, Expansion refers to the elaboration relations between two text spans. The Temporal category subsumes time-related eventualities. Contingency encompasses relations linked by cause, condition and purpose, and Comparison refers to the relations between two eventualities where differences and similarities are highlighted. The semantics of each of these categories is further refined at the second level. A third level specifies the semantic contribution of each argument (Prasad et al., 2008). For instance, example (8) below presents an implicit discourse relation, where the second clause provides more detail about the situation described in the first clause. The clauses are linked with the Level-1 sense of *Expansion*, the Level-2 sense of *Level-of-detail* and the Level-3 sense of *Arg2-as-detail*.

(8) We have a population that's both growing and aging, we have seven billion souls today, heading to 10 billion at the end of the century. (TED ID 1927) [Expansion: Level-of-detail: Arg2-as-detail]

To deal with the interactive characteristics of TED talks, a new top-level sense called Hypophora is brought to bear. Hypophora refers to questions and the answers given to them by the speaker him/herself. In the TED-MDB, it is defined as an alternative means of lexicalising a relation with questions and answers as the constitutive units of the discourse relation (Zeyrek et al., 2020). Hypophora exists in most forms of verbal communication, with the rhetorical

function of enhancing interest in the speech. This is illustrated by example (9) from our corpus.

(9) Are investors, particularly institutional investors, engaged? Well, some are, and a few are really at the vanguard.

While discourse relations are assigned senses from all three levels of the PDTB 3.0 sense hierarchy in the TED-MDB, in the current work, the analysis of senses is limited to Level-1 senses.

3.3 *Inter- versus Intra-sentential Encoding of Discourse Relations*

Researchers have acknowledged that an understanding of discourse relations not only necessitates an understanding of how semantic information is expressed by differently realised discourse relations but also whether the discourse relation is stated within the boundaries of a sentence or over more than one sentence in discourse. Fabricius-Hansen (1996) was one of the first authors to notice the importance of this aspect of discourse. Following the Discourse Representation Theory (Asher and Lascarides, 2003), she analysed German and Norwegian in terms of 'informational density', highlighting it as a problem for translation. By bearing in mind this notion, she compared two texts conveying the same meaning, one with less linguistic material than the other, as demonstrated by version (a) versus version (b) of example (10). Both texts answer the question 'what is going on?', but (a) expresses the same content by conveying the information in a more compact and informationally dense form than (b). In addition, while (a) is more complex syntactically, the interpretation of (b) invites accommodation and inference.

(10) a. France mourns the death of a very famous French actor.
b. A French actor died. He was very famous. France mourns his death.

Despite being a very useful approach to translation, the notion of informational density is difficult to apply in our case because sentences such as (10)a would frequently be disregarded in our work as the information is not expressed by a discourse relation. However, cases like (10)b are relevant to our analysis and correspond to implicitly conveyed discourse relations. Thus, we will only be concerned with a rough distinction of inter-sentential versus intra-sentential encoding of discourse relations, as in (10)b versus example (11) below (also see Section 4.1).

(11) France mourns SINCE a very famous French actor died.

4 Methodology

This section first summarises the major annotation categories that capture the different realisations of discourse relations (Section 4.1). Then, it overviews the data and the annotation tool with which the TED-MDB was created and provides an evaluation of the data's stability (Section 4.2). It explains the technique of sense assignment (Section 3.2) and ends with an account of the data alignment and linking methods employed (Section 4.3).

4.1 *Major Annotation Categories*

In Section 3, a number of the discourse relation realisation types were introduced; these also constitute the major annotation categories of the TED-MDB, which are annotated in the following manner:

- Explicit and implicit discourse relations and AltLexes are searched and annotated together with their arguments, both at the inter-sentential and intra-sentential levels, and assigned sense labels.
- EntRels and NoRels are annotated only at the inter-sentential level together with their arguments; a sense tag is not assigned to them. Topic shifts and mode shifts are considered NoRels.

Although TED talks are monologues, the transcripts contain punctuation marks, and we consider periods, exclamation marks and question marks to be the delimiters of a sentence. We specify discourse relations as inter-sententially encoded if their arguments are separated by one of the delimiters of a sentence. Discourse relations are characterised as intra-sententially encoded if the arguments are separated by a comma, semi-colon or colon. In the absence of one of these marks, discourse relations are also identified as intra-sentential. Once the data alignment and linking procedures have been completed (see Section 4.3), the discourse relations are automatically tagged for their intra-/inter-sentential encoding type with the punctuation rules described.

4.2 *Data and the Annotation Tool*

TED talks are highly structured, prepared, short speeches that are published online. They have become a very popular form of public speech and an 'emergent genre' (Ludewig, 2017). Their availability through TED's own website or web inventories have made them an excellent source of data for researchers needing parallel data. Raw data for the TED-MDB was obtained from the Web Inventory of Transcribed and Translated Talks (WIT3) (Cettolo et al., 2012). Table 1 lists the number of sentences in each transcript per language being considered in the current paper.

TABLE 1 Sentence counts in each TED-MDB talk

Talk ID	EN	LT	PT	TR
Talk 1927	114	122	128	117
Talk 1971	27	31	28	28
Talk 1976	88	96	85	100
Talk 1978	82	88	83	83
Talk 2009	30	32	31	31
Talk 2150	44	45	57	62
Total	385	414	412	421

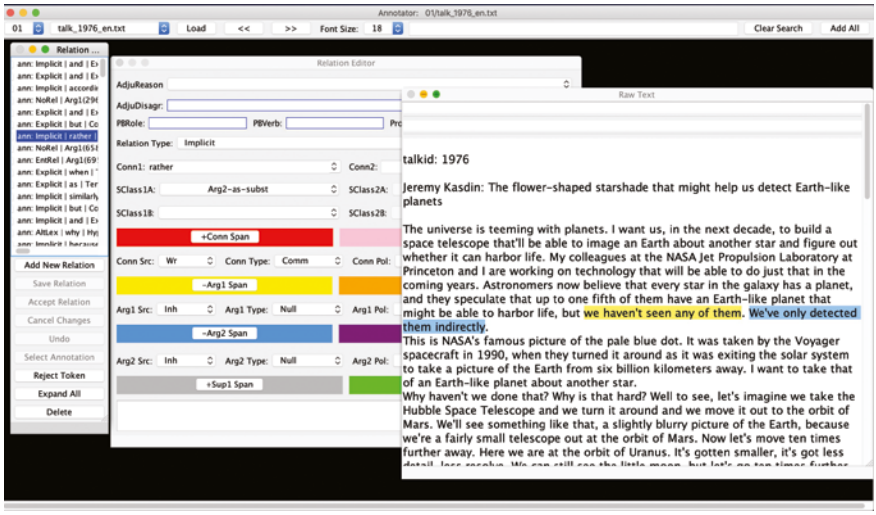


FIGURE 1 The annotation tool

Following a set of guidelines, the transcribed texts were annotated from start to end by paying attention to clause-to-clause and sentence-to-sentence transitions. This annotation style reflects the incremental comprehension of texts by readers and it is a bottom-up technique because the annotators' inferences and intuitions are reflected in the corpus. The PDTB annotation tool (Lee et al., 2016) (Figure 1) is used to create the corpus as it provides a user-friendly environment, enabling users to read the text and annotate the components of the relations. The tool presents the list of annotated relations, their realisation type and sense in the left panel and highlights one discourse relation at a time in the right panel. The annotation tool has a pipe-delimited data

format, which can be conveniently converted to other data formats such as spreadsheets.

The stability of the annotations was evaluated by F-score and Cohen's Kappa (Cohen, 1960), as explained by Zeyrek et al. (2020) and Oleškevičienė et al. (2018). In all languages, discourse relations were spotted with an average F-score of 0.83, and a good level of interrater reliability was reached in discourse relation realisation type and Level-1 sense between two annotators with an average κ of 0.83 and 0.84, respectively.[3] The data has undergone several updates since it was first released. The most recent update involves an automatic sentence-to-sentence alignment performed over the source text and target text pairs and a linking procedure where the labels over the annotated relations of English are linked to those over the target texts, as summarised in the following section (also see Özer et al., 2022 for a detailed explanation of these procedures).

4.3 *Linked and Unlinked Data*

As a result of sentence-to-sentence alignment and relation linking procedures, we currently have a set of linked and unlinked data (see Table 2 below), which can be exploited to investigate a range of discourse phenomena in TED talks (Özer et al. 2022).

TABLE 2 Total and linked DR counts in the TED-MDB

	EN	LT	PT	TR
Total	716	821	680	760
Linked	–	591	597	608

The linked data involves discourse relations in English that have corresponding discourse relations in the target languages. They can be used to examine, for example, conversions from one realisation type to another in the corpus. Example (12) presents an instance where an English implicit discourse relation is changed to an explicit one in Lithuanian and Turkish:[4]

3 κ values of 0.61–0.80 are regarded as substantial agreement, and values of 0.81–1.00 are considered almost perfect agreement.

4 In the following examples, the first constitutive unit, i.e., the first argument of the discourse relation (Arg1), is rendered in italics, and the second argument (Arg2) in bold font. The

(12) EN *The petals unfurl,* (Implicit = IN OTHER WORDS) **they open up.** (TED ID 1976).
LT *Pamatysite,* KAIP **skleidžiasi žiedlapiai.** 'You will see as petals open up.'
TR **Yapraklar açıl**IP *genişliyor*. 'The petals open up and expand.'[5]

Example (13) illustrates an AltLex-to-explicit conversion showing the translator's choice of not applying word-for-word translation for the English AltLex *this is why*, but rather turning to the commonly used conjunction *todėl* 'so' in Lithuanian:

(13) EN *Long-term value creation requires the effective management of three forms of capital: financial, human, and physical.* THIS IS WHY **we are concerned with ESG**. (TED ID 1927)
LT *Ilgalaikės vertės kūrimui reikia efektyviai valdyti tris kapitalo formas: finansinį, žmogiškąjį ir fizinį.* TODĖL **mums ir rūpi** *ASV.*

The unlinked data, on the other hand, includes relations that are annotated in one language but are not annotated in the other language for one reason or another. For instance, example (14) involves two annotated explicit relations signalled by *until* and *so that* on the English side, but only one of these discourse relations, namely the discourse relation anchored by *kadar* 'until', exists on the Turkish side. An equivalent of the *so that*-relation is not annotated on the Turkish side. However, Turkish annotates another relation anchored by the causal postposition *için* 'since/in order to' that is not annotated in English.

(14) EN UNTIL we live in a society where every human is assured dignity in their labor SO THAT they can work to live well, not only work to survive, there will always be an element of those who seek the open road as a means of escape, of liberation and, of course, of rebellion. (TED ID 2009)
TR Her insanın kendi işinin onurundan emin olduğu bir dünyada yaşayana KADAR, [*onlar iyi yaşamak için çalışacaklar*, **sadece hayatta kalmak için değil, daima bir anlamda kaçmak** IÇIN], özgürlük için ve tabii ki başkaldırı için.

relation sense is indicated once, where relevant. Languages are abbreviated as EN (English), LT (Lithuanian), PT (Portuguese) and TR (Turkish).

5 In addition to the lexical connectives of other languages, Turkish has suffixal connectives such as -IP in example (12). These are essentially converbial suffixes that function as subordinators.

> 'UNTIL they live in a world where each person is sure of the dignity of her job, they will work for the good, [in order not to survive but always in some sense to escape] for freedom and of course for rebellion.'
> (Back translation)

The quantitative analyses in the rest of the paper will make use of both linked and unlinked data.

5 Results and Analysis

As previously explained, the main goal of this paper is to compare and contrast how discourse relations are manifested in English and three target languages on the basis of three discourse parameters in an annotated multilingual corpus. This goal was broken down into three sub-goals as follows:

- RQ1: Do target texts differ from English and each other in conveying discourse relations explicitly?
  a. Across texts in the corpus, are discourse relations frequently explicit or implicit?
  b. How often are English discourse relations explicitated (i.e., how often is a discourse connective that is not present in the source discourse relation added to the target discourse relation) or implicitated (how often is a discourse connective that is present in a source discourse relation dropped in the target discourse relation)?
- RQ2: Do target texts differ from English and each other in encoding discourse relations inter-sententially versus intra-sententially? (i.e., is information distributed across sentences or gathered in a single sentence across languages in the corpus)?
  a. How are inter-sentential and intra-sentential discourse relations distributed across English and translated texts?
  b. Does the explicit/implicit realisation of discourse relations have an impact on their inter-/intra-sentential encoding?
- RQ3: Do target texts retain the annotated top-level sense of corresponding discourse relations?
  a. How are top-level senses distributed in the corpus?
  b. Is there a tendency for sense shifts, and what causes them?

The rest of this section introduces the results, both quantitative and qualitative, organised under three subsections devoted to providing answers to each respective research question. In both the tables and the analyses, *English* refers

to the source texts in the corpus, and *Turkish*, *Lithuanian* and *Portuguese* refer to the respective translations.

5.1 *Explicitness across Languages*

To answer RQ1a, we start with the overall distribution of discourse relation realisation types in the data (both linked and unlinked). Table 3 lists the percentage of each discourse relation realisation type in the corpus and indicates that English tends to express discourse relations explicitly. This pattern is quite closely followed by Turkish and Lithuanian. However, Portuguese tends to encode relations implicitly.

Let us now turn to the linked data to answer RQ1b and to examine the tendencies for implicitation and explicitation. The heatmaps (a–c) in Figure 2 depict how often target texts implicitate and explicitate source text discourse relations, demonstrating that 78–81% of explicit discourse relations and 79–85% of implicit discourse relations are preserved in target texts. Notable results from the analysis of frequency counts in the heatmaps are provided below, focusing on changes that go beyond 10%:

- In Turkish, implicitation occurs relatively infrequently (10% of source text explicits are implicated). This process occurs more often in Lithuanian and Portuguese, where 16% and 19% of explicits are implicitated, respectively.
- Explicitation occurs less often than implicitation in our data; 11% of implicits are explicitated in Lithuanian and 12% in Portuguese. Turkish explicitation is slightly higher than other languages, where 15% of implicits are converted to explicits.
- In addition to these, AltLexes display a variation in how they are retained or changed in target texts, as depicted in Figure 2: In Lithuanian, 37% of the original AltLexes are converted to explicits. This ratio is 20% and 21% in Portuguese and Turkish, respectively. However, Portuguese converts 29% of AltLexes to implicits, Lithuanian 21%, while Turkish changes 12%

TABLE 3 DR realisation types across languages

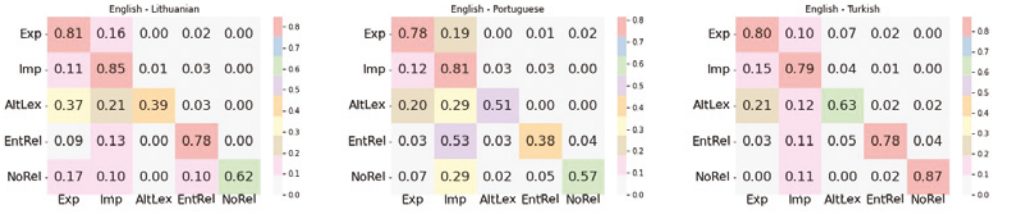
	AltLex (%)	**EntRel (%)**	**Explicit (%)**	**Implicit (%)**	**NoRel (%)**	**Total**
ENG	46 (6.42)	78 (10.89)	289 (40.36)	254 (35.47)	49 (6.84)	716
LT	18 (2.19)	79 (9.62)	377 (45.92)	315 (38.37)	32 (3.90)	821
PT	29 (4.26)	38 (5.59)	269 (39.56)	311 (45.74)	33 (4.85)	680
TR	60 (7.89)	70 (9.21)	315 (41.45)	264 (34.74)	51 (6.71)	760
Total	153	265	1250	1144	165	2977

of AltLexes to implicits. These percentages imply that an examination of implicitation and explicitation should also involve the analysis of AltLexes (see the discussion of Table 4 below).

Based on the answers to RQ1a and b, we can conclude that Turkish and Lithuanian tend to preserve the degree of explicitness in source text discourse relations at more or less equal levels. We continue to scrutinise the linked data through Table 4, including AltLexes in the analysis. By examining AltLex+explicit-to-implicit translation, we find that Portuguese and Turkish behave quite differently. As opposed to Turkish, which only implicitates 10.56% of AltLexes and explicits, Portuguese implicitates 20.29% of these connecting devices. Lithuanian lies between the two languages, implicitating 17.12% of AltLexes and explicits.

As for explicitation, Lithuanian and Turkish explicitate implicits and AltLexes at a slightly higher percentage than Portuguese. However, the difference between the three target languages is not significant, which prevents us from using explicitation to measure the strength of explicitness. Turning to implicitation as a more relevant signal of the strength of explicitness, and evaluating Figure 2 together with Table 4, we are led to a cline of explicitness from Turkish to Portuguese as follows: TR>LT>PT. We believe this provides a better answer to RQ1b, showing that despite an overall tendency to keep the explicitness level of the original discourse relations in translations, Turkish favours a higher level of explicitness than the other languages in the corpus, particularly in regard to Portuguese, which stands out as preferring implicitness to convey discourse relations.

In examining implicitation, the question which immediately arises is whether the process is sensitive to discourse relation sense. Although this



(a) English-Lithuanian (b) English-Portuguese (c) English-Turkish

FIGURE 2 Heatmap visualisations for the distribution of discourse relation realisation types in source text-target text pairs. Rows correspond to English discourse relations and columns denote target language discourse relations. The matrices are normalised row-wise, where each cell represents the percentage of English discourse relations that are converted to the respective label in the target language. For example, Figure 2a shows that in Lithuanian, 16% of English explicit relations are rendered implicitly (Özer et al., 2022).

TABLE 4 Implicitation/explicitation counts across languages

Language	Total Linked AltLex+Explicit DR Count	Implicitation Count	Total Linked AltLex+Implicit DR Count	Explicitation Count
LT	257	44 (17.12%)	248	38 (15.32%)
PT	276	56 (20.29%)	240	32 (13.33%)
TR	284	30 (10.56%)	244	40 (16.39%)

TABLE 5 Counts of *and* omission and other connectives in Expansion relations in target texts

Language	Total no. of Explicit Expansions	Omission of 'and'	Omission of other connectives
LT	23	15 (65.22%)	8 (34.78%)
PT	35	31 (88.57%)	4 (11.43%)
TR	17	14 (82.35%)	3 (17.65%)

was not among our research questions, even a quick inspection of the data shows that among the implicitated Expansion relations, the discourse connective *and* (one of the most common markers of Expansion:Conjunction in our framework) is omitted more often than other discourse connectives signalling Expansion. Table 5 provides the number of explicit Expansion relations in the linked data and shows how often *and* is omitted compared to other discourse connectives. All languages implicitate *and*, and Portuguese ranks highest for the number of implicitated cases (see example (15) below).

5.2 *Inter- vs. Intra-sentential Encoding of Discourse Relations across Languages*

This section tackles RQ2, i.e., whether the inter- vs. intra-sentential encoding of discourse relations differs in English and the target texts. To answer the first sub-part of the question (RQ2a), which enquires about the distribution of inter- and intra-sentential discourse relations across languages, we examine the overall data (both the linked and unlinked parts of the corpus) (Table 6). In English, inter- and intra-sentential discourse relations are distributed almost equally, with a 50.14/49.86% split, but target languages seem to differ from the original language. In particular, a Chi-square goodness of fit test reveals that

TABLE 6 Distribution of inter-sentential and intra-sentential relations across languages

	Inter-sentential (All relation types)		Intra-sentential (All relation types)						
	count	%	count	%	Total	%	Chi-Square	Asymp. Sig	Result
ENG	359	50.14	357	49.86	716	100	0.0056	0.94042	>0.05
LT	385	46.89	436	53.11	821	100	3.1681	0.07509	>0.05
PT	369	54.26	311	45.74	680	100	4.9471	0.02614	<0.05*
TR	396	52.11	364	47.89	760	100	1.3474	0.24574	>0.05

Note: *significant difference

the difference between inter-and intra-sentential discourse relations is significant in Portuguese.

Table 7 lists the distribution of inter- and intra-sentential discourse relations in terms of their realisation types to help answer RQ2b (the impact of implicit or explicit realisation on inter-/intra-sentential encoding). We infer from the table that in English, inter-sentential implicit discourse relations (implicit+NoRel+EntRel) constitute the majority of all inter-sentential discourse relations (76.05%). Lithuanian and Turkish have similar values to English, implicitly encoding 74.80% and 72.98% of inter-sentential discourse relations. Portuguese behaves differently by encoding 82.38% of inter-sentential discourse relations implicitly. In addition, in all languages, the number of explicit intra-sentential discourse relations is approximately 2 to 3 times the number of implicit intra-sentential discourse relations.

All in all, two facts are revealed by the analysis carried out to answer RQ2b: Regarding inter-sentential discourse relations, Portuguese is distinguished from other languages in the corpus by its tendency to encode implicit discourse relations inter-sententially more often than other languages. Portuguese then prefers to establish a low-level discourse structure by relating *sentences* to each other *implicitly*. Secondly, it was observed that all languages tend to encode *intra-sentential* discourse relations *explicitly*, although the extent to which they do so varies. That is, all languages tend to convey intra-sentential discourse relations with a discourse connective. Further research into the types of discourse connectives that are used in intra-sentential discourse relations could better reveal each language's patterns.

TABLE 7 Distribution of inter- and intra-sentential DRs in terms of realisation types

	Inter-sentential						Intra-sentential					
Lang.	**AltLex**	**EntRel**	**Explicit**	**Implicit**	**NoRel**	**Total**	**AltLex**	**EntRel**	**Explicit**	**Implicit**	**NoRel**	**Total**
EN	30 (8.36%)	78 (21.73%)	56 (15.60%)	146 (40.67%)	49 (13.65%)	359	16 (4.48%)	0	233 (65.27%)	108 (30.25%)	0	357
LT	16 (4.16%)	79 (20.52%)	81 (21.04%)	177 (45.97%)	32 (8.31%)	385	2 (0.46%)	0	296 (67.89%)	138 (31.65%)	0	436
PT	27 (7.32%)	38 (10.30%)	38 (10.30%)	233 (63.14%)	33 (8.94%)	369	2 (0.64%)	0	231 (74.28%)	78 (25.08%)	0	311
TR	48 (12.12%)	70 (17.68%)	59 (14.90%)	168 (42.42%)	51 (12.88%)	396	12 (3.30%)	0	256 (70.33%)	96 (26.37%)	0	364

5.3 *Relation Senses across Languages*

This section has two aims: to report on the analysis of Level-1 senses to answer RQ3 and to discuss the sense shifts observed in translating Expansion relations (Section 5.4), in an attempt to understand whether translation (particularly implicitation) affects the interpretation of the semantic content of discourse relations (Section 5.5).

Figure 3 shows the quantitative results of the analysis of Level-1 senses in the linked and unlinked parts of the corpus. It shows that target texts mimic the distribution of the Level-1 senses of the source texts. In other words, in English and the target languages, Expansion constitutes more than half the total number of sense-annotated relations, followed by Contingency (almost 50% of Expansion) and Comparison (almost 50% of Contingency). The number of temporal relations is almost half that of Comparison, with the lowest count of all the sense-annotated relations.

This is called a Zipfian distribution. Zipf's Law (Zipf, 1945) is a power law that reveals a puzzling fact about human language, such that linguistic elements occur according to a systematic frequency distribution, where their frequency is inversely proportional to their rank in the corpus.

5.4 *Sense Shifts*

Despite the overall similarity in the distribution of Level-1 senses across languages, there is not always a perfect match in the Level-1 senses assigned to parallel discourse relations. By manually examining the Expansion relation in English and the translations, we have noticed that these sense shifts may be triggered by translation, crucially, the implicitation of a discourse connective or annotation methodology. In Portuguese, for example, implicitation of *and* seems to trigger Expansion-to-Contingency shifts; in the absence

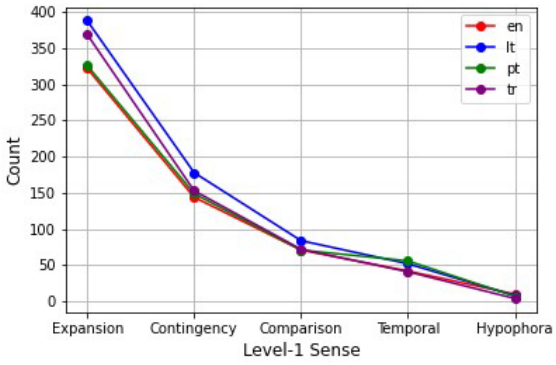


FIGURE 3 Distribution of Level-1 senses in the corpus

of a discourse connective in the translated discourse relation, the annotator appears to infer a sense which differs from the sense annotated for the corresponding English relation. Example (15) illustrates how the omission of *and* leads the annotator to assume a Contingency:Cause:Result+Belief relation.

(15) EN *I believe that a city is the relationships of the people that live there,* AND **I believe that if we can start to document those relationships in a real way then maybe we have a real shot at creating those kinds of cities that we'd like to have.** (TED ID 2150) [Expansion:Conjunction]

PT *Penso que uma cidade é a soma das relações das pessoas que lá vivem.* (Implicit = PORTANTO 'so') **Acho que, se começarmos a documentar essas relações de forma real talvez tenhamos uma hipótese para criar o tipo de cidades que gostaríamos de ter.** [Contingency:Cause:Result+Belief]
'*I think if we start to document these relationships in a real way,* **maybe we have a chance to create the kind of cities that we would like to have**.' (Back translation)

In Lithuanian, on the other hand, we find a different consequence of translation on Expansion-to-Contingency shifts, where several Expansion relations are translated with an incongruent discourse connective, leading the annotator to choose whatever sense the connective has in Lithuanian. This can be shown in example (16), where *and* is translated into Lithuanian by the causal *tad* 'so' and annotated accordingly in Lithuanian.

(16) EN *I think it's reckless to ignore these things because doing so can jeopardize future long-term returns.* AND **here's something that may surprise you: the balance of power to really influence sustainability rests with institutional investors, the large investors like pension funds, foundations, and endowments.** (TED ID 1927) [Expansion:Conjunction]

LT *Manau, yra neatsakinga ignoruoti šiuos dalykus, nes taip darydami keliame pavoju̧ ilgalaikei grąžai ateityje.* TAD **pasakysiu kai ką, kas gali jus nustebinti: galios balansas, galintis išties paveikti tvarumą, yra instituciniu̧ investuotoju̧ rankose.** [Contingency:Cause:Result] (TED ID 1927)
'*I think it is irresponsible to ignore these things because in doing so we are jeopardizing long-term returns in the future.* **So I will say something that may surprise you: the balance of power that can really affect sustainability is in the hands of institutional investors**'. (Back translation)

A third type of sense shift concerns multiple senses of a relation.

(17) EN It was first suggested by Lyman Spitzer, the father of the space telescope, in 1962, and he took his inspiration from an eclipse. You've all seen that. *That's a solar eclipse.* (Implicit = BECAUSE) **The moon has moved in front of the sun.** [Contingency:Cause:Reason] (TED ID 1976)

PT *É um eclipse solar.* (Implicit = ISTO É 'that is') **A Lua colocou-se à frente do Sol**. [Expansion:Level-of-Detail:Arg2-as-detail] '*It's a solar eclipse.* **The Moon placed itself ahead of the Sun**'. (Back translation)

In example (17), the English and Portuguese discourse relations can be understood as conveying Cause as well as Expansion senses; i.e., it could be inferred that the second sentence not only provides more detail about the solar eclipse mentioned in the first sentence (as reflected in the Portuguese annotation) but also gives the reason why this is a solar eclipse (as reflected in the English annotation). Both discourse relations could be assigned two senses. The TED-MDB allows the annotators to give more than one sense to a discourse relation (implicit or explicit). Still, these cases have not been annotated systematically, resulting in differences between source and target texts. We conclude that cases like example (17) should be addressed in a different study since they are due to reasons which fall outside the relation sense, i.e., the annotation methodology. Examples (15) and (16) prove that the implicitation of a discourse connective and the incongruent translation of a discourse connective can trigger sense shifts in the data. The former example can be taken as evidence for the negative effect that implicitation has on the interpretation of discourse relation sense, and future studies could illuminate the extent to which such instances appear in the corpus. The examples further highlight some limitations of working with translated data.

6 Summary and Conclusion

This research set out to investigate one of the essential aspects of discourse structure, namely discourse relations in TED talk transcripts in the source language, English, and their translations into three languages of the TED-MDB. The paper focused on the degree of explicitness that is encoded in discourse relations, whether the latter are encoded inter- or intra-sententially, and whether the top-level senses annotated in the source texts are retained by the

target languages. Comparison of discourse relations in terms of these dimensions showed similarities across the languages, but there were also some differences. Regarding the level of explicitness in discourse relations, the analysis that concentrated on the implicitation of explicit relations and AltLexes placed Turkish (which tended to display a more substantial level of explicitness than the other languages) and Portuguese at opposite poles of a cline. Lithuanian was placed closer to Portuguese. So, this approach confirmed our earlier findings regarding the level of explicitness displayed by Turkish and Portuguese discourse relations. Secondly, the overall distribution of the number of inter-/intra-sententially encoded discourse relations in the data showed that Portuguese tended to express implicit discourse relations inter-sententially more frequently than the other languages. On the other hand, in all the target languages, intra-sentential discourse relations were more often conveyed with a discourse connective than without one.

Finally, although we found some evidence exemplifying the adverse effect of translation on the interpretation of senses, our results showed that discourse relations were annotated more or less similarly for Level-1 senses across the languages. These results are similar to those obtained from our earlier work, implying that the Level-1 discourse senses conveyed by the original TED speeches have been transmitted to target texts without any substantial changes. The inclusion of Lithuanian in the analysis did not seem to change the general picture. Admittedly, this is a somewhat broad picture of the semantic dimension of our corpus, because Level-1 senses provide rather coarse-grained information about the semantic content of discourse relations. A fine-grained semantic analysis could be conducted by examining the Level-2 and Level-3 sense tags over the discourse relations of the languages included in the corpus.

Given the scarcity of annotated multilingual discourse-level corpora, the rich discourse-level annotations contained in the TED-MDB are unprecedented. The current work is novel as it illuminates the explicitness and inter- vs. intra-sentential encoding of discourse relations in TED talks as vital aspects of discourse structure. Based on the TED-MDB, we described how three target languages rendered discourse relations and presented some methodological challenges when dealing with the multiple facets of discourse relations. It is hoped that the results will pave the way for new hypotheses on discourse structure in English and the target languages to be developed, and will facilitate further research using a similar methodology. However, the work is not without its limitations. Firstly, the workflow of TED talk translations consists of translating the text into a different language and then adjusting the translation to the subtitle lines. Although a reviewer always checks the translated text, the result may not be ideal, which could

be an area of potential confusion for our analysis. It should be noted that the results of our small-scale study on the omission of *and* appear to be in line with previous research describing the behaviour of this conjunction, so the translation process of TED talks may not have a significant impact, at least in regard to the omission of (certain) discourse connectives in translation. Secondly, the TED-MDB is a small resource; hence, the current work is restricted by the size of the corpus. These limitations affect the generalisability of the results. This issue can be addressed in future investigations by drawing upon larger amounts of data, and the results reported here can be compared with those from multilingual or monolingual corpora involving the languages under scrutiny.

References

Asher, Nicholas M. 1993. *Reference to abstract objects in discourse. Studies in linguistics and philosophy*; v. 50. Dordrecht: Kluwer Academic Publishers.

Asher, Nicholas M. and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge: Cambridge University Press.

Asr, Fatemeh T. and Vera Demberg. 2012. Implicitness of discourse relations. In: *Proceedings of COLING 2012*, 2669–2684.

Cettolo, Mauro, Christian Girardi, and Marcello Federico. 2012. WIT3: Web inventory of transcribed and translated talks. In: *Conference of European Association for Machine Translation*, 261–268.

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20(1): 37–46.

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

Das, Debopam and Maite Taboada. 2018. Signalling of coherence relations in discourse, beyond discourse markers. *Discourse Processes* 55(8): 743–770.

Fabricius-Hansen, Cathrine. 1996. Informational density: A problem for translation and translation theory. *Linguistics* 34(01): 521–566.

Fraser, Bruce. 1999. What are discourse markers? *Journal of Pragmatics* 31(7): 931–952.

Halliday, Michael Alexander K. and Ruqaiya Hasan. 2014. *Cohesion in English*. London and New York: Routledge.

Halverson, Sandra. 2004. Connectives as a translation problem. *An International Encyclopedia of Translation Studies* 1: 562–572.

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

Hofmockel, Carolin, Anita Fetzer, and Robert M. Maier. 2017. Discourse relations: Genre-specific degrees of overtness in argumentative and narrative discourse. *Argument & Computation* 8(2): 131–151.

House, Juliane. 2015. Global English, discourse and translation. Linking constructions in English and German popular science texts. *Target. International Journal of Translation Studies* 27(3): 370–386.

Karamanis, Nikiforos. 2007. Supplementing entity coherence with local rhetorical relations for information ordering. *Journal of Logic, Language and Information* 16(4): 445–464.

Kehler, Andrew. 2002. *CSLI Lecture Notes Series.* Vol. 104: *Coherence, reference, and the theory of grammar*. CSLI Publications.

Knott, Alistair and Robert Dale. 1994. Using linguistic phenomena to motivate a set of coherence relations. *Discourse Processes* 18(1): 35–62.

Knott, Alistair, Jon Oberlander, Michael O'Donnell, and Chris Mellish. 2001. Beyond elaboration: the inter-action of relations and focus in coherent text. In T. Sanders, J. Schilperoord, and W. Spooren, editors, *Text Representation: Linguistic and Psycholinguistic Aspects*, pages 181–196. Amsterdam: Benjamins.

Lee, Alan, Rashmi Prasad, Bonnie Webber, and Aravind Joshi. 2016. Annotating discourse relations with the PDTB annotator. In: *Proceedings of COLING 2016, The 26th International Conference on Computational Linguistics: System Demonstrations*, 121–125.

Ludewig, Julia. 2017. TED Talks as an Emergent Genre. *CLCWeb: Comparative Literature and Culture* 19(1): <https://doi.org/10.7771/1481-4374.2946>.

Maier, Robert M., Carolin Hofmockel, and Anita Fetzer. 2016. The negotiation of discourse relations in context: Co-constructing degrees of overtness. *Intercultural Pragmatics* 13(1): 71–105.

Mann, William C. and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text* 8(3): 243–281.

Mauranen, Anna. 1999. Will 'translationese' ruin a contrastive study? *Languages in Contrast* 2(2): 161–185.

Murray, John D. 1997. Connectives and narrative text: The role of continuity. *Memory & Cognition* 25(2): 227–236.

Oleškevičienė, Giedre V., Deniz Zeyrek, Viktorija Mažeikienė, and Murathan Kurfalı. 2018. Observations on the annotation of discourse relational devices in TED talk transcripts in Lithuanian. In: *Proceedings of the Workshop on Annotation in Digital Humanities Co-located with ESSLLI* Vol. 2155: 53–58.

Özer, Sibel, Murathan Kurfalı, Deniz Zeyrek, Amália Mendes, and Giedrė V. Oleškevičienė. 2022. Linking discourse-level information and induction of bilingual discourse connective lexicons. *Semantic Web*, Vol. Pre-press, pp. 1–22.

Prasad, Rashmi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation* (*LREC'08*), Marrakech, Morocco. European Language Resources Association (ELRA).

Prasad, Rashmi, Aravind Joshi, and Bonnie Webber. 2010. Realization of Discourse Relations by Other Means: Alternative Lexicalizations. In: *Coling 2010: Posters*. Beijing, China: Coling 2010 Organizing Committee, 1023–1031.

Sanders, Ted. 2005. Coherence, causality and cognitive complexity in discourse. In: *Proceedings/Actes SEM-05, First International Symposium on the Exploration and Modelling of Meaning*. University of Toulouse-le-Mirail Toulouse (Veranst.), 105–114.

Schiffrin, Deborah. 1986. Functions of *and* in discourse. *Journal of Pragmatics* 10(1): 41–66.

Steiner, Erich. 2015. Contrastive studies of cohesion and their impact on our knowledge of translation (English-German). *Target. International Journal of Translation Studies* 27(3): 351–369.

Webber, Bonnie, Markus Egg, and Valia Kordoni. 2012. Discourse structure and language technology. *Natural Language Engineering* 18(4): 437–490.

Webber, Bonnie, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. *The Penn Discourse TreeBank 3.0 Annotation Manual*. Philadelphia: University of Pennsylvania.

Zeyrek, Deniz, Amália Mendes, Yulia Grishina, Murathan Kurfalı, Samuel Gibbon, and Maciej Ogrodniczuk. 2020. TED Multilingual Discourse Bank (TED-MDB): A parallel corpus annotated in the PDTB style. *Language Resources and Evaluation* 54(2): 587–613.

Zipf, George K. 1945. The meaning-frequency relationship of words. *The Journal of General Psychology* 33(2): 251–256.

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

Biographical Notes

Deniz Zeyrek is Professor of Linguistics at Middle East Technical University and carries out interdisciplinary research by compiling language resources. Her research specialties are discourse and pragmatics and their role in understanding human cognition. Recently, she has concentrated on discourse mechanisms in an attempt to understand the role of discourse relations in human languages. She has been the principal developer of the Turkish Discourse Bank

(an electronic resource of Turkish annotated for discourse relations in the style of the Penn Discourse Treebank). She co-edited the book *Discourse Meaning, The View from Turkish* (2020) and has published internationally in edited volumes and peer-reviewed journals.

Amália Mendes is Associate Professor at the School of Arts and Humanities of the University of Lisbon. Her main research interests are corpus linguistics as well as discourse and lexical studies. She has been a coordinator and team member of several national and European projects. She is one of the executive directors of PORTULAN CLARIN – Research Infrastructure for the Science and Technology of Language, and a member of the Organizing Committee of the Comprehensive Grammar of the Portuguese Language, published by Fundação Calouste Gulbenkian (2013, 2020). She has developed the CRPC-Discourse Bank, a corpus of Portuguese annotated for discourse relations in the Penn Discourse Treebank style.

Giedrė Valūnaitė Oleškevičienė is Associate Professor at Mykolas Romeris University, Institute of Humanities. She defended her doctoral dissertation *Making Sense of Social Media Use in University Studies* in 2016. Her scientific interests include areas such as the methodology of social research, problems of contemporary education philosophy, development of creativity in the modern education system, etc. She is also actively involved in research on teaching foreign languages, as well as pursuing scientific interests in linguistics and translation.

Sibel Özer is a senior software engineer and a PhD student at Middle East Technical University, Cognitive Science Department. Her scientific research interests focus on discourse mechanisms in human languages, linguistic data mining and modelling. Her thesis work specifically focuses on linking discourse relation annotations over parallel corpora to obtain a better insight into cross-linguistic coherence devices in texts.