LINKING DISCOURSE-LEVEL INFORMATION: A STUDY ON DISCOURSE RELATION ALIGNMENT WITHIN MULTIPLE TEXTS AND LANGUAGES

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS OF
THE MIDDLE EAST TECHNICAL UNIVERSITY
BY

SIBEL ÖZER

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN
THE DEPARTMENT OF COGNITIVE SCIENCE

SEPTEMBER 2024

LINKING DISCOURSE-LEVEL INFORMATION: A STUDY ON DISCOURSE RELATION ALIGNMENT WITHIN MULTIPLE TEXTS AND LANGUAGES

submitted by **SIBEL ÖZER** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Cognitive Science Department, Middle East Technical University** by,

Prof. Dr. Banu GÜNEL KILIÇ
Dean, **Graduate School of Informatics** _____

Assoc. Prof. Dr. Barbaros YET
Head of Department, **Cognitive Science** _____

Prof. Dr. Deniz ZEYREK BOZŞAHİN
Supervisor, **Cognitive Science, METU** _____

Examining Committee Members:

Assoc. Prof. Dr. Barbaros YET
Cognitive Sience, METU _____

Prof. Dr. Deniz ZEYREK BOZŞAHİN
Cognitive Sience, METU _____

Prof. Dr. Cem BOZŞAHİN
Cognitive Sience, METU _____

Prof. Dr. Burcu CAN BUĞLALILAR
Computing Science, University of Stirling _____

Assoc. Prof. Dr. Aytaç ÇELTEK
Department of Western Languages and Literatures, Kırıkkale University _____

Date: 04.09.2024

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Sibel ÖZER

Signature :

ABSTRACT

LINKING DISCOURSE-LEVEL INFORMATION: A STUDY ON DISCOURSE RELATION ALIGNMENT WITHIN MULTIPLE TEXTS AND LANGUAGES

ÖZER, Sibel

Ph.D., Department of Cognitive Science

Supervisor: Prof. Dr. Deniz ZEYREK BOZŞAHİN

September 2024, 116 pages

This thesis examines the complex nature of cross-linguistic discourse structures and the expression of discourse relations within multilingual contexts, focusing specifically on the TED-MDB corpus. By aligning discourse relations in parallel corpora, the study explores variations in how discourse is realized, semantic shifts, and patterns of inter-sentential encoding across different languages. The analysis emphasizes differences in expression, implicitation and explicitation of discourse connectives, and the distribution of discourse senses, highlighting the nuances in discourse translation. In addition, the study develops methods for bilingual lexicon induction from naturally occurring data, creating valuable resources on multiple languages for discourse and pragmatic studies and the enhancement of natural language processing (NLP) systems. Future research directions include investigating alternative discourse annotation schemes, exploring domain-specific impacts on discourse translation, examining syntactic interactions, and expanding the analysis to other language pairs while aligning data to Linked Language Open Data (LLOD) standards. This research significantly contributes to the understanding of linguistic differences in conveying discourse relations and semantic adaptations in the translation of discourse relations across diverse languages.

Keywords: Discourse Relation Alignment, Lexicon Induction, Multilingual Corpora, Cross-linguistic Corpus Analysis

ÖZ

DERLEM SEVIYESI BILGININ BAĞLANMASI: BİRDEN FAZLA METİN VE DİLDE DERLEM İLİŞKİSİ HİZALAMASI ÜZERİNE BİR ÇALIŞMA

ÖZER, Sibel

Doktora, Bilişsel Bilimler Bölümü

Tez Yöneticisi: Prof. Dr. Deniz ZEYREK BOZŞAHİN

Eylül 2024, 116 sayfa

Bu tez, çok dilli bağlamlarda, çapraz-dil söylem yapılarının karmaşık doğasını ve söylem ilişkilerinin ifade edilişini, özellikle TED-MDB derlemine odaklanarak incelemektedir. Paralel derlemlerde söylem ilişkilerini hizalayarak, çalışmada söylemlerin nasıl gerçekleştirildiği, anlamsal kaymalar ve farklı dillerdeki cümleler arası kodlama desenlerindeki varyasyonlar araştırılmaktadır. Analiz, ifade farklılıkları, örtük bilgi, açıklık ve söylem anlamlarının dağılımı üzerindeki ayrıntıları vurgulamakta, söylem çevirisindeki incelikleri öne çıkarmaktadır. Ayrıca, çalışmada doğal olarak oluşan verilerden ikidilli sözlük türetme yöntemleri geliştirerek, pragmatik çalışmalar ve doğal dil işleme (NLP) sistemlerinin geliştirilmesi için değerli kaynaklar yaratılmaktadır. Gelecek araştırma yönleri, alternatif söylem işaretleme düzenlerinin araştırılması, söylem çevirisi üzerindeki alanlara özgü etkilerin incelenmesi, sözdizimsel etkileşimlerin gözden geçirilmesi ve çeşitli dil çiftlerine genişletilmiş analiz yapılması, veri bağlantılarının Linked Language Open Data (LLOD) standartlarına göre hizalanması gibi konuları içermektedir. Bu araştırma, farklı dillerde söylem çevirisinde dilsel incelikler ve anlamsal uyumlar hakkında daha derin bir anlayışa önemli ölçüde katkıda bulunmaktadır.

Anahtar Kelimeler: Derlem İlişkilerinin Hizalanması, Sözlük Türetimi, Çokdilli Derlem, Çapraz-Dil Derlem Analizi

To my little daughter, Asya

ACKNOWLEDGMENTS

My journey towards completing my PhD began in 2010. It has been a significant and transformative experience for me. My passion for studying linguistics predates even my Master's degree, and has grown stronger during my time as a doctoral student. While my PhD studies have posed their challenges and have not always been a straightforward path to success, they have allowed me to investigate the intricate workings of language in the human mind from various perspectives. From exploring neurolinguistics and analysing EEG signals to studying eye-tracking methodology and finally discourse mechanisms across languages, I have gained invaluable insights and knowledge.

Throughout this journey, I have come across personal milestones such as marriage, motherhood, and divorce. Despite these challenges, I am grateful for the continuous support and guidance of my colleagues and friends, without whom this thesis would not have been possible.

I am especially thankful to my supervisor, Deniz Zeyrek Bozşahin, whose assistance and inspiration have been indispensable throughout my research. It is hard to find the appropriate acknowledgement words that can reflect my gratitude. I could not ever make this thesis possible without her support. Her guidance, feedback, and encouragement have been instrumental in shaping my academic and personal growth as well as my work. Studying linguistics with her is a delight, honour and chance for me.

I also extend my gratitude to my TIK committee members, Cem Bozşahin and Burcu Can Buğlalılar, for their valuable insights and suggestions, which have significantly contributed to the clarity of my work. I also thank my PhD Thesis Defense Jury members, Barbaros Yet and Aytaç Çeltek, for their valuable and constructive comments on the first draft of my thesis.

I would like to express my appreciation to Murathan Kurfalı for his collaboration and insightful discussions, which have enriched my research.

I am fortunate to have a close friend like Burcu Ayşen Ürgen. Throughout my years in the PhD program, we have had the opportunity to discuss various topics ranging from life to academia. She has provided me with continuous encouragement and support, both as a friend and by sharing her experiences as a mentor to numerous students.

A special thank you goes out to my beloved family members: my mother Nazmiye Özer, my father İbrahim Özer, and my brother Volkan Özer. Their unconditional love and constant support has enabled me to focus on my thesis surpassing the challenges of being a single parent.

Lastly, a heartfelt thank you is for my little daughter Asya, whose patience, love, and understanding have been my source of strength throughout this journey of growth and transformation. This is for you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

DR	Discourse Relations
TED-MDB	TED Multilingual Discourse Bank
SMT	Statical Machine Translation
DIMLex	Dialogue Markup Language Extensions
SL	Source Language
TL	Target Language
DR	Discourse Relation
DC	Discourse Connective
LLOD	Linguistic Linked Open Data
POS	Part of Speech
NLP	Natural Language Processing
NLP	Natural Language Processing
NLU	Natural Language Understanding
LOD	Linked Open Data
URI	Unique Resource Identifier
RDF	Resource Description Framework
SPARQL	SPARQL Protocol and RDF Query Language
OLiA	The Ontologies of Linguistic Annotation
PDTB	Penn Discourse Treebank
RST	Rhetorical Structure Theory
SDRT	Segmented Discourse Representation Theory
EDU	Elementary Discourse Unit

BERT	Bidirectional Encoder Representations from Transformers
EN	English
TR	Turkish
LT	Lithuanian
PT	Portuguese
DE	German
PL	Polish
RU	Russian
AltLex	Alternative Lexicalization
EntRel	Entity Relation
NoRel	No Relation
WIT3	Web Inventory of Transcribed and Translated Talks
UPlug	Universal Plug and Play
LASER	Language-Agnostic SEntence Representations
BLEU	Bilingual Evaluation Understudy
IAA	Inter-Annotator Agreement
TP	True Positives
FP	False Positives
TN	True Negatives
FN	False Negatives
XML	eXtensible Markup Language
HTML	HyperText Markup Language
URL	Uniform Resource Locator
CSS	Cascading Style Sheets
TMX	Translation Memory Exchange

CHAPTER 1

INTRODUCTION

Discourse deals with the semantic or pragmatic relations within and across sentences. One of the essential aspects of discourse coherence is discourse relations such as contrast, temporal, contingency. The effective use of discourse connectives plays a crucial role in maintaining coherence across texts. These linguistic tools help connect ideas and structure discourse. Despite their importance, discourse relations have often been examined within a single language in monolingual corpora. This excludes some recent attempts that deal with discourse-level information in multilingual corpora ([2],[3],[4] and [5]).

Given the scarcity of discourse-level investigation in multiple languages, the objectives of this study are as follows:

- Given the availability of TED-Multilingual Discouse Bank which involves discourse-level annotations independently created in English (source texts) and multiple translated (target) texts, it aims to introduce a framework and establish a methodology for aligning discourse relation annotations.
- It creates bilingual discourse connective lexicons, and generates a resource that enhances understanding of the complex dynamics surrounding bilingual discourse connectives in translated texts.
- It contributes to discourse-level investigations of relatively low-resourced languages like Turkish and Lithuanian.
- Thus, despite the source language being English and the target languages being translation languages, the methodology of the current thesis can be adapted with minor modifications for application to other parallel corpora.

1.1 Problem Statement and Motivation

One of the recently introduced multilingual discourse corpora is TED-Multilingual Discourse Bank, or TED-MDB [6], created by a team of researchers including English, Turkish, Polish, Russian, Portuguese, German and Lihtuanian. TED-MDB is a corpus annotated for discourse relations in English TED talks and their translations into several languages. It follows the rules and principles of the Penn Discourse Treebank [7] for English, and with a lexicalized approach to discourse, it annotates both

“explicit" connectives (i.e. overtly marked connectives such as *since*, *because*, *however*) as well as implicit relations, considering they are also anchored by “implicit" connectives (i.e., connectives that can be inserted in a discourse relation to make their semantics salient).

In starting the TED-MDB project, most languages did not have a lexicon of discourse connectives, so the project team used syntactic cues (subordinating, coordinating conjunctions, and adverbials) to identify explicit discourse connectives.

TED-MDB provides an ideal platform for conducting cross-lingual discourse analysis and developing mono- and bilingual discourse connective lexicons for a range of languages. Thus, it forms the data source of the current thesis.

However, the TED-MDB project is built on the idea that each research team should annotate the texts in that language independently of English, by native speakers of each language. This was considered necessary so that English annotations did not influence the annotations in the other languages because one of the aims of the project was to contribute to an understanding of the range of discourse relations in target languages and to reveal how discourse relations unfold in transcribed texts.

Although TED-MDB is one of the earliest multilingual corpora at the discourse level, it significantly lacks the alignment of discourse relations. That is, given its design features, relation annotations in English are not serialized with those in target texts and not linked to them.

The current work is motivated by the need to conduct cross-lingual discourse analysis on TED-MDB, which requires aligning the corpus at the level of discourse relations, which in turn needs alignment at sentence or word level. It should be noted that aligning TED-MDB at the sentence or word level would be insufficient for cross-linguistic comparisons unless discourse relations are also aligned. So, our ultimate motivation in this thesis is to align discourse relation annotations of English with those of target languages. We devised a framework and methodology which first aligns the texts at sentence level, then link discourse relations annotated in English with those annotated in target texts.

1.2 Contributions of the Study

In a nutshell, the contributions of this thesis are as follows:

Development of a Discourse Relation Linking Framework: This work develops of a framework and methodology for linking discourse-level information. This framework provides a systematic approach for analyzing and identifying discourse connectives in TED-MDB and is hoped to prove useful for corpora that involve low-resource target languages.

Bilingual Discourse Connective Lexicon Induction: We induced a discourse connective lexicon from the linked discourse connective annotations. In this way, we contributed to the discourse connective inventory of multiple languages. This attempt is hoped to contribute to the creation of more effective and comprehensive natural language understanding (NLU) tools for diverse linguistic contexts.

Annotation Validation: The output of our discourse relation annotation linking methodology makes it easy to focus on missing annotations as it captures the non-linked annotations in the source language. In this way, target language annotations can be checked against source language annotations, which

would be quite costly if performed manually. Thus, annotation validation spots any relations that have not been annotated in the target languages of TED-MDB, showing the gaps in the annotated data.

Improved Cross-Linguistic Analysis: The induced bilingual (English-target language) discourse connective lexicon not only contributes the connective lexicons of each language, but it is hoped to enhance cross-linguistic analysis and the understanding of how discourse is structured and translated. This can lead to improved language processing and translation systems that take into account the nuances of discourse markers in multiple languages.

Extending TED-MDB Annotations: The annotations in Turkish, Lithuanian, European Portuguese, and partially in English were refined and extended using discourse relation alignments and by identifying missing links. Also, new properties such as inter- and intra-sentential distinctions, POS tags for discourse connectives, and subtitle location were added.

Overall, the study's contributions could lead to advancements in the field of linguistics, natural language processing, and cross-linguistic communication systems.

1.3 Organization of the Thesis

This thesis is structured into six chapters. The content of some sections of the thesis have already been published or presented. References are provided to each of these works within the text.

Chapter 2 introduces the three-partite framework of the thesis involving alignment, projection and linking precedures. It also serves as an introduction to parallel corpora and covers the related literature.

Chapter 3 outlines the primary research conducted in this thesis, focusing on discourse relation alignment methodology. The chapter addresses the method's limitations, performance evaluation, error analysis, and concludes with descriptive statistics.

Chapter 4 deals with the creation of bilingual discourse connective lexicons, emphasizing the induction process and structural characteristics of these lexicons.

Chapter 5 details the utilization of unaligned data for the enhancement of annotations in Turkish, Lithuanian, and European Portuguese. This chapter also discusses the integration of inter- and intra-sentential characteristic of discourse relations, POS Tags and subtitle location information into the alignment dataset, which contributes to further cross-linguistic analysis.

Chapter 6 offers a detailed analysis of aligned discourse relations, focusing on five key topics: discourse relation types, higher-level senses, intra- and inter-sentential realization, the impact of subtitling, and contextual effects on discourse connective translation. This chapter discusses each finding with respect to the maintenance of local coherence in multilingual translational corpora.

Chapter 7 concludes the thesis by summarizing the findings and main contributions and outlines possibilities for future research.

CHAPTER 2

FRAMEWORK

In this section, the three-partite framework will be introduced together with the key terminology relevant to this framework.

2.1 Three-partite framework

This thesis is based on a framework that uses three procedures in computational linguistics: (a) alignment, (b) projection, and (c) linking.

In this chapter first, each of these procedures will be explained, and then how we integrated these aspects in our framework will be introduced.

2.1.1 Alignment

The alignment procedure involves the automatic serialization of texts of different languages. Alignment is necessarily applied on bilingual/multilingual corpora and can be performed at various levels, depending on the phenomena being investigated.

In computational linguistics, a corpus generally refers to a collection of spoken or written utterances. This data is typically representative of a particular genre and is usually available electronically. It is considered partially representative because a corpus is merely a sample, a finite set of data aimed at capturing the expressive power of natural language. Corpora can be classified along several dimensions, one of which is language. A corpus is monolingual if it consists of only one language, and multilingual if it involves multiple languages.

A parallel corpus is needed for alignment but not all bi/multilingual corpora are parallel corpora. In a parallel corpus, segments of a source-language document are aligned with corresponding segments in different languages [8]. In bilingual parallel corpora, these aligned segments are sometimes referred to as bitexts [9].

2.1.1.1 Sentence Alignment

In sentence alignment, paragraphs and sentences are typically used as segmentation units. Segments in the source language are matched with corresponding segments in the target language. Although these links are often not 1-to-1 (as illustrated by the sample bitext pairs in Table 1[1]), the mapping process is monotonic. This means that the direction of information flow at the sentence level is assumed to be consistent between the source and target languages. Consequently, the alignment links created do not cross.

Table 1: Sentence alignment in one of the TED talks: Al Gore: Averting the climate crisis / Al Gore: Al Gore iklim krizine çözüm buluyor.

Relation Type	**English**	**Turkish**
1:1	Then I remembered it could be a bunch of things.	Sonra bir sürü şey olabileceği aklıma geldi.
1:2	(Laughter) But what it turned out to be was that my staff was extremely upset because one of the wire services in Nigeria had already written a story about my speech, and it had already been printed in cities all across the United States of America.	(Kahkahalar) Fakat çalışanlarımdan birisi oldukça üzgündü çünkü Nijerya'daki haber ajanslarından birisi konuşmamla ilgili bir hikaye yazmışlardı. Ve bu haber ABD'nin tüm şehirlerine yayılmıştı.
2:1	(Laughter) Now, I know that you wanted some more bad news about the environment – I'm kidding. But these are the recapitulation slides, and then I'm going to go into new material about what you can do.	Çevreyle ilgili daha kötü haberler beklediğinizin farkındayım – Şaka yapıyorum – bunlar özet slaytları, sonrasında ise neler yapabileceğinizle ilgili yeni bilgileri paylaşacağım.

To automatically align sentences, several methods have been proposed. Length-based models exploit the correlations between the lengths of corresponding sentences [11]. Conversely, dictionary-based models rely on the distribution correspondences between lexical units [12]. Additionally, hybrid approaches combine these two methods and may also incorporate document structure.

Automatic sentence alignment is generally a well-established task, achieving high accuracy rates of above 90%. However, this performance is highly dependent on the quality of the corpora used. Accuracy may decline in the case of noisy parallel corpora that has poor or incomplete translations.

2.1.1.2 Word Alignment

Word alignment involves linking corresponding words and phrases in parallel corpora. Unlike sentence alignment, these links are non-monotonic, making word alignment a more challenging task due to linguistic discrepancies such as differences in morphology and word order. It is not always possible

[1] Taken from the English-Turkish TED talk corpus constructed in [10]

to assume that identical lexical concepts in the source and target languages will behave similarly in a given context. However, if natural languages are compositional, there must be translational correspondences between the source and target languages, and word alignment methods rely on this notion. One valuable outcome of the word alignment process is the extraction of bilingual lexicons.

The goal of aligning all lexical items in the source language to their corresponding target language items often results in fuzzy translation relations due to morphological and grammatical differences. Additionally, inconsistencies in translations and typographical errors further complicate this process (see Table 2 and Figure 1 for an example of word alignment(produced using Eflomal[13]), where the numbers in the third row of the table indicate word indices in the bitext; the first number before the '-' represents the source index and the second one the target index, with indexing starting from zero. This format is known as pharaoh output format).

Table 2: Eflomal word alignment output for TED Talk Id: 1927 Chris McKnett: The investment logic for sustainability/Sürdürülebilirlik için yatırımın mantığı

English	And they are really complex and they can seem really far off, that the temptation may be to do this: bury our heads in the sand and not think about it.
Turkish	Gerçekten de karmaşık ve uzak görünebilirler, ki bu da şunu yapmamızı cazip kılabilir: Kafamızı kuma gömüp, bunun hakkında düşünmemek.
Eflomal Output	3-0 4-2 5-3 7-1 8-5 10-4 11-4 11-5 12-6 13-7 14-8 14-9 15-12 15-13 16-10 17-9 19-11 20-10 21-14 22-17 23-15 24-15 27-16 28-18 30-21 31-20 32-19 33-22
Eflomal Output Text Form	really-Gerçekten, complex-karmaşık, and-ve, can-de, seem-görünebilirler, far-uzak, off-uzak, off-görünebilirler, ,-,-, that-ki, the-bu, the-da, temptation-cazip, temptation-kılabilir, may-şunu, be-da, do-yapmamızı, this-şunu, :-:, bury-gömüp, our-Kafamızı, heads-Kafamızı, sand-kuma, and-,, think-düşünmemek, about-hakkında, it-bunun, .-.

Several approaches for word alignment are well-documented in the literature. The first of these is the association or heuristic approach, which aligns words using various correspondence measures or heuristics (e.g., co-occurrence, string similarity) [14]. In these approaches, lexical segmentation or boundary identification is performed first. Using certain association criteria, possible translation pairs are then gathered with weighted scores to form an association dictionary. Finally, a best-first search algorithm, constrained by linguistic criteria, aligns the lexical items. The bilingual translation lexicon extracted in this final step is generally more reliable than the initial association dictionary.

The second approach is known as the estimation or statistical alignment method [15]. The Statistical Machine Translation (SMT) approach originates from information theory [16], calculating the probability of a target string being an appropriate translation of a source string in a noisy channel transformation. The architecture of these models depends on the initial IBM statistical models of word alignment. Prominent examples include the widely known GIZA++ [14](and its faster version MGiza[17]), Fast-align [18], and Eflomal[13].

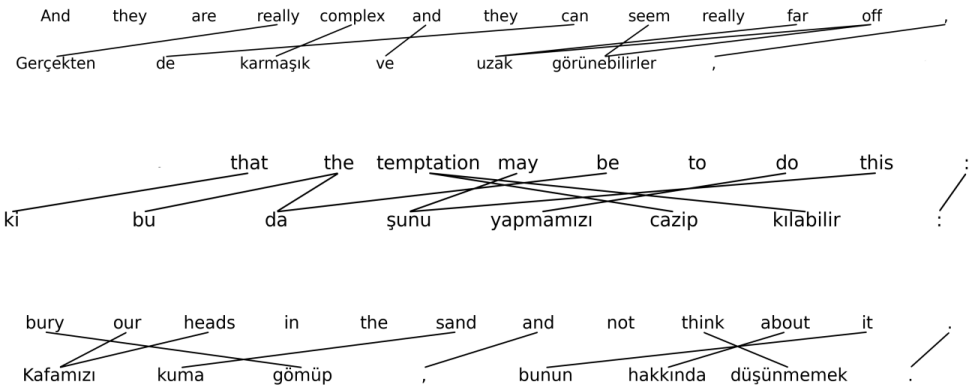


Figure 1: Eflomal word alignment output for TED TalkId: 1927 Chris McKnett : The investment logic for sustainability/Sürdürülebilirlik için yatırımın mantığı

Recently, with the rise of sentence embedding models in machine translation, word alignment models using multilingual sentence embeddings like BERT have also emerged. Examples include SimAlign [19] and AwesomeAlign [20]. While SimAlign does not rely on pre-trained parallel data, AwesomeAlign offers the flexibility of training the underlying models for improved performance. Although these models often outperform statistical ones, they require larger amounts of data to do so.

Traditional statistical word alignment models do not account for morphology and struggle with allomorphy, a frequently observed phenomenon in languages with complex morphology, such as Turkish, which is agglutinative. In allomorphy, a single morpheme (as in the case of plural formation, case suffixes, verb conjugation, etc., in Turkish) can take different forms depending on the phonological or morphological context [21]. In statistical word alignment literature [22], this is incorrectly described as a limitation, specifically data sparsity, where there are many different forms with low individual frequencies. However, in many NLP tasks, including statistical machine translation, it has been shown that this is an inaccurate use of the terminology. Indeed, when properly defined and used, morphology provides a rich source of information that enhances task performance. For example, [23] shows that rich morphology provides more information in information retrieval tasks. [24] emphasizes that morphology can provide additional contextual clues in semantic disambiguation.

In statistical word alignment or machine translation models, several methods are applied to handle morphological variants. Tools like Eflomal incorporate parameters such as prefixes and suffixes for the source and target languages. These parameters specify the number of characters to be removed from the beginning or end of lexical items to reduce variations in lemmas caused by inflections. However, this approach is insufficient, as it results in information loss rather than exploring patterns in the data. [25] suggests that truly incorporating linguistic features as additional layers of meaningful information can improve word alignment performance. Such attempts have been made by several studies in the literature: [26] incorporated parts-of-speech (POS) tags, while [27], [28], [29], [30], and [31] incorporated morphologically analyzed data in word alignment or statistical machine translation models.

2.1.2 Projection

Since the early 2000s, a method known as projection, cross-lingual transfer learning or annotation transfer, has been successfully used as a data-driven approach to create monolingual annotated data for various linguistic features, including morphological, grammatico-syntactic, and semantic information. For instance, [32] and [33] projected annotations from English to both closely related and distant languages to develop resources for part-of-speech tagging, noun phrase chunking, named-entity tagging, and morphological analysis. Similarly, [34], [35], and other researchers explored projection for tasks such as dependency parsing, temporal annotation, word sense disambiguation, information extraction, FrameNet construction, syntactic tree translation, spoken language understanding, and coreference chains. However, the projection of discourse connectives remains a relatively understudied area. [36] and [37] made attempts in this direction by disambiguating German connectives and projecting discourse annotations from English to French texts, respectively.

The annotation projection process typically involves a bilingual pair where one segment is in the source language (SL, usually English) and the other is its translation in a target language (TL, often a low-resource language). Annotations are then added manually or automatically to the SL segment for a specific paradigm (e.g., semantic roles, syntax, discourse connectives), and these annotations are projected onto the TL segment. Annotation Projection can be used to facilitate manual annotation in the target language (TL). After transferring the annotations from the source language (SL) to the TL, manual corrections and additional annotations are made on the TL. Alternatively, the projected annotations on the TL segment can be used as training data to develop an automatic parser for the paradigm. Figure 2 depicts a typical explicit discourse connective annotation projection. In the English sentence, explicit discourse connectives are highlighted in red color; whereas in Turkish, blue color is used. First, in projection, manually or automatically discourse connective annotations are done on the source sentence. Later on through word alignment, these annotations are projected onto the target language. However, as the example shows discourse annotation projection task is not a feasible way to align discourse relations. In Figure 2, Turkish discourse connective gömüp was aligned to *bury* rather than *and*. [36] and [37] propose the usage of external resources like discourse connective lexicons or additional heuristics to filter non-connective word alignments. Also, in the Example provided in Figure 2, including morphology into Eflomal word alignment model may partially increase the projection performance.

This method is effective when there is a certain level of consistency between the SL and TL for the paradigm. For example, [38] report the limited effectiveness of cross-lingual transfer in creating an English-Chinese parallel corpus annotated for negation, as negation is handled quite differently in English and Chinese. In the case of discourse connectives, both languages should have similar explicit and implicit connective annotations. Sometimes, annotation projection is combined with other resources to enhance accuracy, such as looking up each projected annotation in a discourse connective lexicon in the TL.

2.1.3 Linking

Over the years, linked data, which refers to structured data published on the web, has gained widespread acceptance, with new techniques and methods continuously evolving to store, connect, and represent this data in formats that can be understood by both machines and humans [39], [40], [41]. This con-

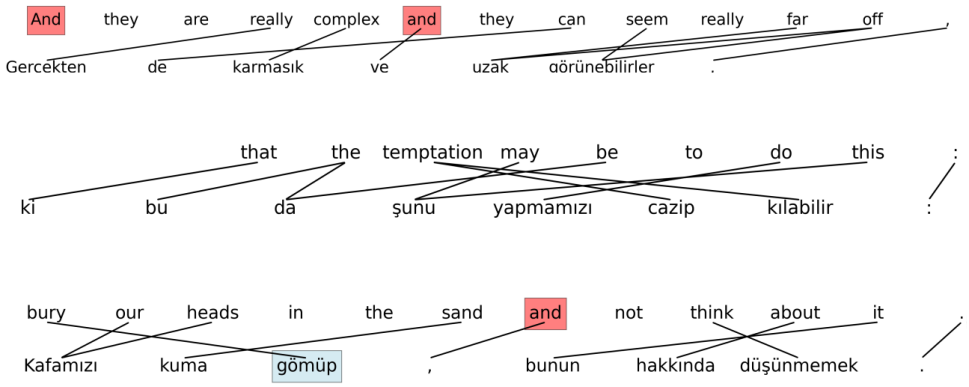


Figure 2: Projection Example on Eflomal word alignment output for TED TalkId: 1927 Chris McKnett : The investment logic for sustainability/Sürdürülebilirlik için yatırımın mantığı

cept is a crucial component of the Semantic Web, transforming the internet into a global database as highlighted by the Linked Open Data (LOD) cloud[2]. The notion of linking resources and data across various websites has captured the interest of researchers in the Semantic Web and Natural Language Processing (NLP) communities, leading to the emergence of Linguistic Linked Open Data (LLOD) as a distinct subset of LOD[42].

The primary objective of LLOD researchers is to address similar challenges faced by the LOD community, such as storing, connecting, combining, and representing linguistic data on the web in a linked data format to facilitate linguistic research. Linguistic data, including machine-readable dictionaries, ontologies, annotated linguistic corpora, and semantic knowledge bases, are regarded suitable for inclusion in the LLOD cloud as long as they adhere to Linked Data (LD) principles, which include:

- Assigning a Unique Identifier (URI) to each entity (e.g., an entry in a lexicon, a document or token in a corpus, annotation labels/data categories) for unambiguous identification.
- Using HTTP URIs to access these entities.
- Adhering to web standards for data representation and querying, such as the Resource Description Framework (RDF) for generic data modeling, metadata for data representation, and SPARQL for querying linked data.

  RDF employs labeled directed graphs to depict data structures, with three key components: a property/relation/predicate (represented as a labeled edge) connecting a subject (a resource) to its object (another resource). This relation between the subject and object serves to link the two entities. An RDF graph is illustrated in Figure 3 using the example sentence: "Burkhard Jung is the mayor of Leipzig" (as extracted from [40]).
- In order to facilitate effective querying of the data, creating unambigous links between different, possibly distributed data sources so that identical senses would be linked across different lexico-semantic resources, equivalent annotation labels will be linked to their corresponding data categories in different repositories, etc.

[2] http://lod-cloud.net/state/

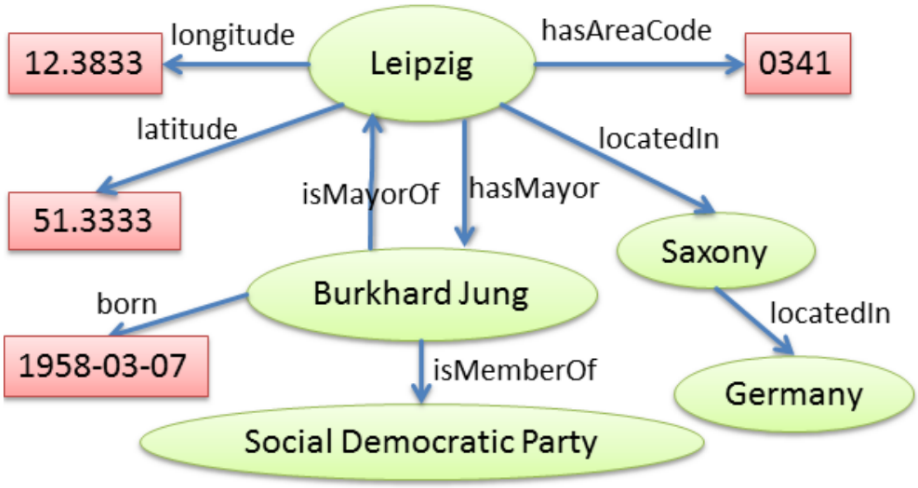


Figure 3: RDF Representation of the Sentence "Burkhard Jung is the mayor of Leipzig"

These principles align with the core objective of the LLOD researchers. By presenting linguistic data on the web, web-based information processing is enabled, making such data accessible to everyone. This approach brings about both structural uniformity (consistency in meta-data) and conceptual uniformity (common vocabulary) in how linguistic data is represented, thereby allowing for the augmentation of data and the creation of rich linguistic resources. It also facilitates the development of generic, reusable, and evolving technologies for processing linguistic data. The implementation of the LLOD framework is inspired by theories related to human lexical memory, focusing on how concepts are stored, accessed, and linked in an efficient manner.

While English dominates the majority of resources in the LLOD cloud, with most resources comprising monolingual RDF datasets, the aim of LLOD is to promote multilinguality in data. This involves providing lexical information in multiple languages and enabling cross-lingual mapping[43]. Despite the abundance of studies that apply LLOD principles to create lexico-semantic resources or annotated corpora for various linguistic layers (such as parts-of-speech, syntax, nominal/verbal chunks, constituent syntax, and WordNet), the development of linked discourse-level resources is a recent evolution. One notable endeavor is highlighted in [44] (also [45]), where distributed discourse connective lexicons and annotations were converted into a machine-readable and linked format. To standardize the annotation schemas across different language resources and discourse relation representations (e.g., PDTB[7], RST[46]), the OLiA (The Ontologies of Linguistic Annotation) Discourse Extensions were introduced[47]. Within this architecture, an OLiA Annotation Model is defined for each language-specific annotated resource, each linked to a generic OLiA Reference Model. The discourse connectives considered are limited to lexical markers comprising one or more lexemes. The definition of a discourse relation, following the structural connective definition of PDTB, involves a discourse connective linking two arguments, with the type of relation representing the sense of that relation. The primary goal is to establish a discourse marker lexicon containing a list of discourse connectives and their associated discourse relation senses. To populate this lexicon, they amalgamated various mono and multilingual discourse marker resources like DimLex, PDTB, LICO, TED-MDB, and others. In most cases, entries adhere to the guidelines of the DimLex-XML data structure. For resources diverging from this convention, such as TED-MDB, a conversion to Dim-Lex-XML format is conducted.

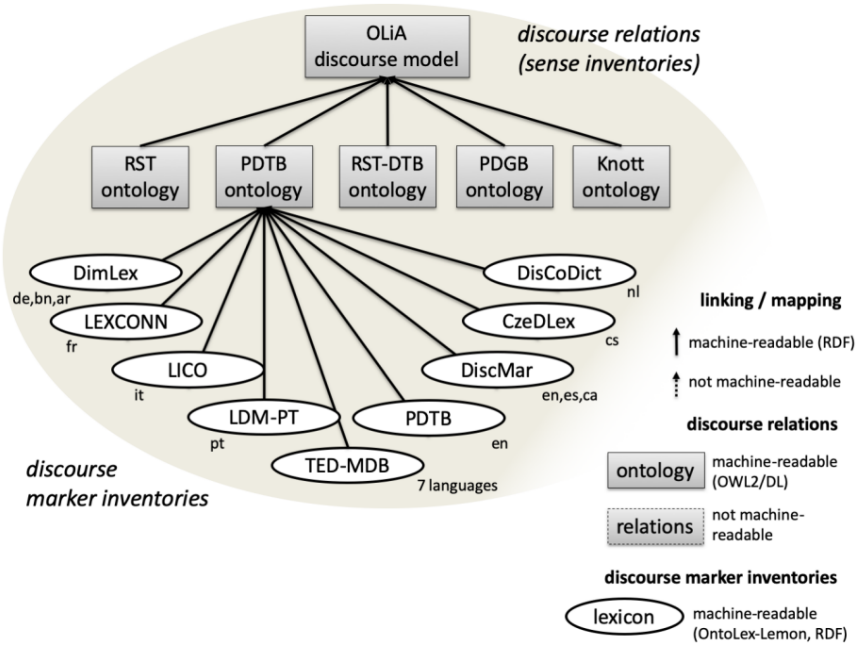


Figure 4: LLOD Framework for Linked Discoure Connective Annotations/Lexicons

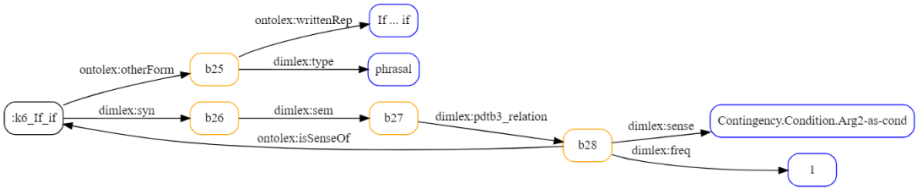


Figure 5: Example OntoLex Discourse Marker

Linking specific Dim-Lex discourse connective entries to designated OLiA annotation models necessitates a conversion from Dim-Lex to OntoLex-Lemon ([48]), which serves as the lexicon model for ontologies defined by and incorporated within the LLOD community. Their framework is illustrated in the accompanying in Figure 4. Using such a framework allows for the listing of discourse markers represented by a similar PDTB sense. Additionally, it is possible to list potential mappings between discourse relation senses from different theoretical frameworks. For example, pdtb:Cause ontology can be mapped to rst:Evidence via olia_discourse:Cause.

In Figure 5, an example of an OntoLex discourse marker and its corresponding sense entry are graphically depicted. The OntoLex-Lemon representation consists of three main elements: ontolex:LexicalEntry serves as the unit of analysis for the discourse connective lexicon, grouping one or more forms/senses. ontolex:Form represents the written form of the connective, and ontolex:LexicalSense denotes the word sense of the entry represented by ontolex:LexicalEntry. The OntoLex dimlex:sense labels are matched with the relevant entries in the OLiA PDTB ontology. [3] Linking numerous distributed resources in a standardized representation is a significant and highly beneficial endeavor. However,

[3] Source: http://purl.org/olia/discourse/discourse.PDTB.owl

restricting the connective set to those with discourse relation types that are Explicit or Alternative Lexicalization (detailed information on DR types will be provided in Chapter 3) would offer an incomplete depiction of the discourse grammar in the provided text. In line with this limitation, the PDTB 3.0 annotation manual ([7]) notes that within the same text span, there may exist more than one relation, often comprising pairs of both Explicit and Implicit relations. Since Implicit discourse relations are not integrated into the LLOD framework, the platform cannot display such connections. Furthermore, while these pairs may not occur within the exact same text span, certain discourse relations (like Explicit comparison and implicit contingency, Explicit comparison and implicit expansion, etc.) are observed to appear consecutively in neighboring sentences at a frequency significantly higher than expected by chance ([49]). Therefore, for any discourse parser seeking to resolve discourse relation senses, it is crucial to include Implicit discourse relations.

Another limitation of the framework is the exclusion of morphologically expressed discourse connectives, also known as converbs. Given that Turkish is a language rich in morphology, and temporal-sense discourse connectives are often expressed through suffixes, excluding this subset of connectives means overlooking a substantial portion of the discourse context.

2.2 Discourse Relation Alignment-Our Methodology

In our three-partite framework (details of the procedure will be provided in Chapter 3), the first step involves aligning discourse relation annotation labels in the TED-MDB, which is an unaligned corpora. To narrow down the search space for alignment, we merged discourse annotation labels with the raw text and performed sentence alignment for English-target languages (TLs). English annotation labels were then projected onto the existing annotation labels in the TLs. Discourse annotation projection works for Explicit and partially for Implicit Discourse Relations (DRs). Yet, within the current framework, all annotations are projected or mapped over to the existing TL annotation labels. However, our use of linking differs from that used by the LLOD.

At present, our framework serves as a pre-processing stage to supply multilingual data to the LLOD cloud. After formal conversions are made to the output of our linked annotation data set, it could potentially be merged with existing LLOD resources, which is a topic for future research. With a source language (SL) and target language (TL), both annotated for discourse connectives based on the rules of the same sense hierarchy and methodology (Penn Discourse Treebank 3.0 for TED-MDB), annotation labels over the discourse connectives are linked (see the example DR alignment in Figure 6). Our ultimate goal is to process multilingual data annotated with discourse relations and convert it into a format that is semantically interpretable and interoperable, ideal for discourse processing.

In summary, in this chapter, I have introduced the major aspects of my approach involving the methods known as alignment, projection and linking in the literature. In the next chapter, I will introduce how I integrated these procedures to solve the problem of discourse relation linking.

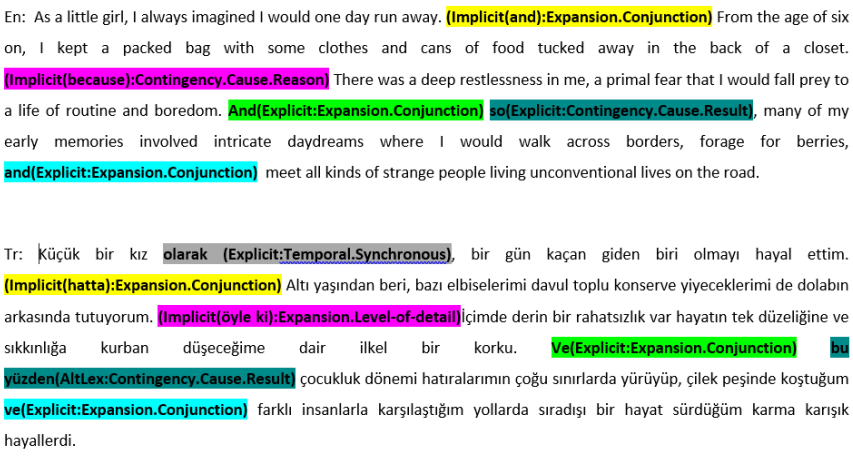


Figure 6: Discourse relation alignments provided by the methodology described in this thesis ([1]) output for part of the TED TalkId: 2009 Kitra Cahana: A glimpse of life on the road/Evsizlerin ve saklananların hikayeleri for English-Turkish language pairs. Discourse relations which are linked are highlighted in the same colour. As depicted in the example, one DR in Turkish segment is unlinked: 'olarak'

CHAPTER 3

DEVELOPMENT OF A DISCOURSE RELATION LINKING METHODOLOGY

In this chapter, I introduce the development of a discourse relation alignment methodology, focusing on the data resources and the processes implemented. The primary data resource is the TED Multilingual Discourse Bank (TED-MDB), which contains TED talk transcripts initially delivered in English and translated into other languages. TED talks provide a rich source of linguistic data due to their diverse topics and structured presentations.

The chapter details the systematic approach to annotating discourse relations, guided by the Penn Discourse Treebank (PDTB) framework, which categorizes relations into types such as Explicit, Implicit, AltLex, EntRel, and NoRel. By extending PDTB to multiple languages, TED-MDB facilitates cross-lingual discourse analysis.

An alignment algorithm that employs multilingual embeddings to associate similar semantic representations of discourse units across languages was developed. This method calculates composite scores based on semantic similarity, sense levels, and relation types. It is thoroughly tested on language pairs involving English and target languages such as Turkish, Portuguese, and Lithuanian, and is evaluated using metrics like Precision, Recall, and F-score.

The chapter further discusses the challenges in aligning discourse relations across languages, including discrepancies in argument spans, different realizations of relations, and translation variations. An error analysis is conducted to pinpoint areas for improvement.

Finally, the results of the DR alignment are published in XML format[1], structured to aid further research and application. The data schema accommodates both linked and unlinked relations, ensuring the accessibility and utility of the corpus for various linguistic and computational studies.

3.1 Data Resource: TED Multilingual Discourse Bank

TED talks are scripted presentations delivered in English before a live audience. These presentations are recorded and shared online along with English subtitles translated into various languages by volunteers and reviewed by experts. The subtitles exclude most interruptions in speech, like pauses and hesitations, while retaining key discourse markers such as *well* [50]. The extensive range of topics

[1] https://github.com/MurathanKurfali/Ted-MDB-Annotations

covered in TED talks across multiple languages makes them a valuable resource for linguistic analysis and comparative studies on prepared spoken text. Furthermore, they are readily accessible on TED's website[2] and web inventories[3]. The raw data used for the TED-MDB was sourced from the Web Inventory of Transcribed and Translated Talks (WIT3) by [51].

TED-MDB corpus is a publicly accessible collection of annotated data resources[4]. The transcripts stored in TED-MDB contain the original English text and translations into six other languages: German, Polish, Russian, European Portuguese, and Turkish. Later on, Lithuanian was also added [6, 52]. These talks, which cover diverse subjects, are conducted by native English speakers, as outlined in Table 3. Also, Table 4 shows the total count of sentences for each language.

Table 3: The list of the TED talks annotated in TED-MDB

ID	Author	Title
1927	Chris McKnett	The investment of logic for sustainability
1971	David Sengeh	The sore problem of prosthetic limbs
1976	Jeremy Kasdin	The flower-shaped starshade that might help us detect Earthlike planets
1978	Sarah Lewis	Embrace the near win
2009	Kitra Cahana	A glimpse of life on the road
2150	Dave Troy	Social maps that reveal a city's intersections and separations

Table 4: Sentence counts in each talk of TED-MDB

TalkID	EN	DE	PL	LT	RU	PT	TR
Talk 1927	114	127	117	122	122	128	117
Talk 1971	27	26	30	31	26	28	28
Talk 1976	88	89	86	96	87	85	100
Talk 1978	82	81	95	88	85	83	83
Talk 2009	30	31	32	32	31	31	31
Talk 2150	44	58	58	45	65	57	62

The TED-MDB corpus was developed collaboratively by an international team of researchers, adhering to the guidelines of the Penn Discourse Treebank (PDTB). The PDTB is a comprehensive resource designed to represent discourse structure by annotating discourse relations in Wall Street Journal texts. By extending the PDTB framework to TED talks, the TED-MDB systematically annotates multiple languages using a unified descriptive model. This initiative aims to facilitate discourse analysis across English and various translated languages within spoken contexts. The resulting corpus serves as a valuable tool for linguists, computational linguists, discourse analysts, translation professionals, and educators seeking pedagogical resources.

[2] https://www.ted.com/

[3] https://www.ted.com/talks/subtitles/id/1927/lang/tr : Turkish Transcript of the TED talk 1927 can be accessed from here.

[4] https://github.com/MurathanKurfali/Ted-MDB-Annotations

3.1.1 Annotation Schema

Discourse markers and connectives serve as linguistic instruments for indicating coherence relations. While discourse connectives are distinct from discourse markers in that they consistently convey a binary relationship between text segments, with each segment having an abstract object interpretation (eventualities, propositions, facts) [53], discourse markers are restricted to associating with just one abstract object. The Penn Discourse Treebank (PDTB) specifically focuses on discourse connectives, viewing them as lexical and syntactic indicators that signify the existence of a discourse relationship [54]. These text segments, known as arguments, constitute the foundation of discourse relations, with coherence emerging through the deduction of a semantic connection between these arguments.

3.1.1.1 Discourse Relation Types

The PDTB identifies six ways in which these discourse relations manifest.

1. Explicit: Similar to example (1), a clearly communicated discourse relation is indicated and emphasized by discourse connectives such as subordinating or coordinating conjunctions, or adverbs. These connectives form a limited set of items to convey the relationship.

   (1) But *no other experience has felt as true to my childhood dreams* as **living amongst and documenting the lives of fellow wanderers across the United States**[5].
   [Explicit, Comparison:Similarity] (English, TED Talk no. 2009)

   Ama *hiçbir başka deneyim çocukluk rüyalarımı yaşayanlar arasında olmak* kadar **ve Birleşik Devlet boyunca gezgin arkadaşların arasında yaşamak kadar gerçek hissettirmedi**.
   [Explicit, Comparison:Similarity] (Turkish, TED Talk no. 2009)

2. Implicit: The concept of implicit discourse relations is rooted in the idea of linearity. An implicit discourse relation can be deduced simply by the closeness of discourse units, with linguistic elements in clauses or sentences providing clues regarding the intended discourse relation ([55], [56]). Also, within the PDTB framework, an implicit discourse relation can be clarified by using a discourse connective known as 'implicit connectives'; for example, a sentence like (2) could be rephrased using *and* in English and *ve* in Turkish.

   (2) *By day, they hop freight trains, stick out their thumbs, and ride the highways with anyone from truckers to soccer moms.* (And) **By night, they sleep beneath the stars, huddled together with their packs of dogs, cats, and pet rats between their bodies**.
   [Implicit, Expansion:Conjunction] (English, TED Talk no. 2009)

[5] The examples are taken from TED-MDB. In all the examples, the discourse connective or AltLex is underlined, Arg1 is rendered in italics and Arg2 in bold type. As in the PDTB, Arg2 is the discourse segment hosted by the discourse connective or AltLex, while Arg1 is the other discourse unit.

Gün başlarken, yük trenlerinden atılıyorlar, baş parmaklarını kaldırıyorlar, ve kamyoncudan futbolcu annelerine kadar kim gelirse onlarla yolculuk ediyorlar. (Ve) **Akşam, yıldızların altında birlikte köpeklerine, kedilerine ve evcil farelerine sokularak uyuyorlar**.
[Implicit, Expansion:Conjunction] (Turkish, TED Talk no. 2009)

3. AltLex: In contrast to implicit relations, AltLexes are alternative ways of lexicalizing discourse relations that are not members of closed class items ([57]). This makes it redundant for a reader to insert an explicit discourse connective, as is evident in Example (3). AltLexes encompass a variety of expressions, ranging from lexically fixed to both syntactically and lexically unrestricted forms.

   (3) *Franz Kafka saw incompletion when others would find only works to praise*, so much so **that he wanted all of his diaries, manuscripts, letters, and even sketches burned upon his death**.
   [AltLex, Contingency:Cause:Result] (English, TED Talk no. 1978)

   Diğerleri çalışmalarını yalnızca övgüye değer bulurken, Franz Kafka bitmemiş olarak gördü, o kadar ki **bütün günlüklerinin, el yazılarının, mektuplarının ve hatta taslaklarının öldükten sonra yakılmasını istedi**.
   [AltLex, Contingency:Cause:Result] (Turkish, TED Talk no. 1978)

4. EntRel: In this type of relation, coherence is provided by the use of the same noun phrase (NP) referents (e.g., the same person, object, or entity) in subsequent clauses ([58], [59]). The key difference between implicit and entity relations is that while implicit discourse relations can be rephrased using a discourse connective, adding an extra connective in entity relations can disrupt the natural coherence of the text (see Example (4)).

   (4) *Investors should also look at performance metrics in what we call ESG: environment, social, and governance*. **Environment includes energy consumption, water availability, waste, and pollution, just making efficient uses of resources**.
   [EntRel] (English, TED Talk no. 1927)

   Yatırımcılar ÇSY diye adlandırdığımız üç faktörün performans metriklerine de bakmalılar: Çevre, sosyal ve yönetim. **Çevre; enerji tüketimini, su bulunabilirliğini, atık ve kirliliği içeriyor, yani kaynakların etkin kullanımını sağlamakla ilgili**.
   [EntRel] (Turkish, TED Talk no. 1927)

5. NoRel: When there is neither an implicit discourse relation nor an entity-based relation between two adjacent sentences, these are labeled as 'no relation' (NoRel). It is important to differentiate these instances of non-coherence or weak coherence from others where coherence is established. In spoken genres, they are mainly used for topic shifts, either by means of an overt marker like "well," "so," "now," or without using any marker (see Example (5)).

   (5) *I believe that sustainable investing is less complicated than you think, better-performing than you believe, and more important than we can imagine*. **Let me remind you what**

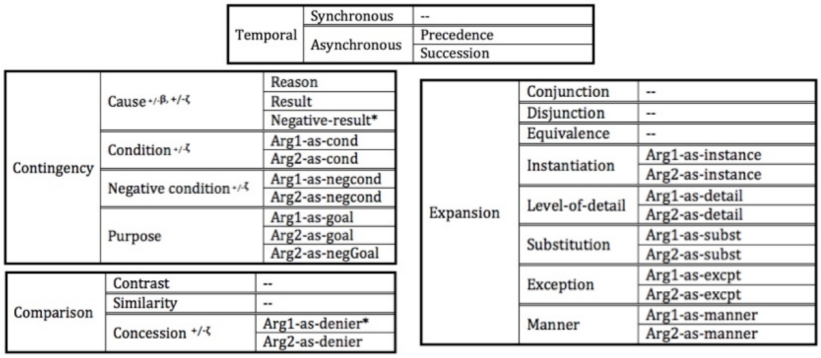


Figure 7: PDTB 3.0 sense hierarchy

we already know.
[NoRel] (English, TED Talk no. 1927)

İnanıyorum ki, sürdürülebilir yatırım düşündüğünüzden daha az karmaşık, sandığınızdan çok daha iyi performansı var ve hayal edebileceğimizden çok daha önemli. **Halihazırda bildiklerimizi tekrar hatırlatayım**.
[NoRel] (Turkish, TED Talk no. 1927)

3.1.1.2 Discourse Relation Senses

In the Penn Discourse Treebank (PDTB) sense hierarchy [7], discourse relation (DR) senses are organized into three levels, progressing from more generic to more specific (see Figure 7). At the highest level, there are four primary semantic categories known as Level-I senses: Expansion, Temporal, Contingency, and Comparison.

- **Expansion** pertains to elaboration relationships between text spans.
- **Temporal** covers time-related events.
- **Contingency** deals with causal, conditional, and purposive relations.
- **Comparison** focuses on highlighting differences and similarities between eventualities.

Each of these top-level categories is further specified at the second level, which adds more detail. At the third level, the semantic contribution of each argument within the relation is detailed, offering a nuanced understanding of how they connect and interact. In Example (6), a discourse relation is illustrated where one argument acts as the goal of the other argument. The Level-I sense is tagged as Contingency, while Level-2 details the relation as Purpose. Finally, Level-3 specifies that Arg2 is the goal ([60], [7]).

(6) **Portfolyomuzu politika beyanı yapmak** için *kullanmak istemiyoruz.*
[Explicit, Contingency:Purpose:Arg2-as-goal] (Turkish, TED Talk no. 1927)

TED talks are interactive speeches where the speaker directs questions to the audience and provides immediate answers, making the talk more lively and engaging as in the Example (7). In TED-MDB, these alternatively lexicalized relations are labeled with a new Level-I sense called Hypophora [6].

(7) **Şehir nedir**? *Sanırım, bazıları coğrafi bölge ya da sokak ve bina topluluğu diyebilir*
[AltLex, Hypophora] (Turkish, TED Talk no. 2150)

3.1.1.3 Annotation Categories, Evaluation and Tool

Different strategies are employed in annotating discourse relation (DR) types. For a summary, see Table 5. Explicit, Implicit, and AltLex discourse relations are identified both within (intra) and across (inter) sentence boundaries, and both arguments (arg1 and arg2) are tagged together, including the sense label. In contrast, EntRels and NoRels are annotated within paragraphs and between sentences, along with their arguments, but they are not assigned sense tags.

Although TED talks are a spoken genre, their transcripts include punctuation marks. These symbols are used to differentiate between inter-sentential and intra-sentential discourse relations. If the arguments of a discourse relation are separated by a period, exclamation mark, or question mark, these are accepted as delimiters, and the DR is classified as inter-sentential. On the other hand, if argument separation is done with no symbol, a comma, semi-colon, or colon, the DR is categorized as intra-sentential. These rules are also incorporated in Chapter 5.

Annotators are asked to annotate all three levels if they are confident; otherwise, they should annotate only the more generic senses. Furthermore, a DR can have two senses if the annotator believes that a discourse connective is ambiguous between two senses or conveys two senses simultaneously. For instance, in Example (8), the discourse connective *ve* (and) is annotated with the sense Expansion:Conjunction. At the same time, it is annotated with Contingency:Cause:Result as a second sense because it conveys a resultive relation also. In Example (9), the discourse connective *when* is annotated with the sense Temporal:Synchronous. However, because it is ambiguous between Temporal and Conditional senses, the annotator also annotated it with Contingency:Condition:Arg2-as-cond as the second sense.

(8) *Dünya benzeri bir gezegene sahip olduklarını* ve **yaşam barındırıyor olabileceklerini**...
[Explicit, Expansion:Conjunction] (Turkish, TED Talk no. 1976)

(9) ...when **they look at a company and decide whether to invest** *they look at financial data, metrics like sales growth, cash flow, market share, valuation – you know, the really sexy stuff...*
[Explicit, Temporal:Synchronous] (English, TED Talk no. 1927)

Table 5: TED-MDB Annotation Scheme

Relation type	Relation anchor	Arguments	Sense	Inter-Intra
Explicit	Overt discourse connective	Arg1, Arg2	Yes	Both
Implicit	Inferred discourse connective	Arg1, Arg2	Yes	Both
Alternative Lexicalization (AltLex)	Alternative way of expressing the relation	Arg1, Arg2	Yes	Both
Entity Relation (EntRel)	None	Arg1, Arg2	No	Inter
No Relation (NoRel)	None	Arg1, Arg2	No	Inter

During the annotation phase, texts in each language were annotated concurrently by native speakers of the respective languages. To ensure that the annotations accurately reflected the discourse structure of each translated language, the work was conducted independently of the original English texts. This methodology aimed to capture the unique characteristics of each language's discourse structure, resulting in different sets of annotated relations for each language. Table 6 presents the number and percentage of each type of relation (Explicit, Implicit, AltLex, EntRel, and NoRel) for each language [6].

Table 6: Distribution of discourse relation types in TED-MDB

Language	AltLex	EntRel	Explicit	Implicit	NoRel	Total
German	17 (3.04%)	59 (10.54%)	240 (42.86%)	214 (38.21%)	30 (5.36%)	560
English	46 (6.5%)	78 (11.02%)	289 (40.82%)	246 (34.75%)	49 (6.92%)	708
Lithuanian	23 (2.63%)	80 (9.15%)	414 (47.37%)	325 (37.19%)	32 (3.66%)	874
Polish	2 (0.35%)	104 (18.06%)	217 (37.67%)	201 (34.9%)	52 (9.03%)	576
Portuguese	34 (4.78%)	40 (5.62%)	283 (39.75%)	319 (44.8%)	36 (5.06%)	712
Russian	20 (3.54%)	57 (10.09%)	237 (41.95%)	221 (39.12%)	30 (5.31%)	565
Turkish	69 (8.77%)	72 (9.15%)	312 (39.64%)	282 (35.83%)	52 (6.61%)	787
Total	1945	1774	201	485	277	4682

The subtitles were annotated sequentially from start to finish, reflecting the temporal unfolding of information in TED talks. To capture the incremental comprehension of the texts by the translators, discourse relation annotation was performed in a manner that mimicked incremental processing. This approach ensured that the annotators' inferences and intuitions were also reflected in the annotations. The PDTB annotation tool was employed for this task [61]. It is a Java-based, user-friendly tool featuring three panels: annotated relations, the main annotation editor pane, and the raw text pane. Annotations are stored in a pipe-delimited format, which can be easily processed programmatically or using other file editors like spreadsheets. The tool supports various customizations through a configuration file, such as default folder descriptions and pre-defined sets of implicit discourse connectives.

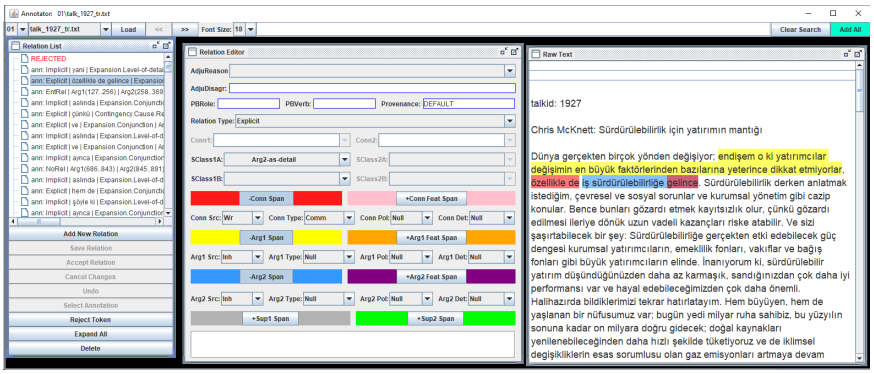


Figure 8: PDTB Annotation Tool Interface

However, there are two main limitations of the tool. Firstly, it is not open source, meaning that customization is limited to what can be accomplished via the configuration files. Secondly, the primary annotation units are tokenized words. This presents a challenge for morphologically complex languages such as Turkish, where discourse relations may be expressed through inflectional morphology, making it impossible to annotate within word boundaries.

To assess the reliability of the annotations (as explained by [6] and [52]), approximately 20% of the entire corpus—equivalent to two TED talks per language—were annotated by an independent annotator. This process adhered to the defined annotation scheme and principles outlined earlier. The inter-annotator agreement (IAA) involves two key aspects: first, determining whether the annotators identified a relation between the same discourse units, with an average F-score of 0.83 indicating agreement on relation spotting. Second, evaluating whether the identified relations were of the same type and sense, which was conducted through a simple ratio agreement using Cohen's Kappa. The IAA results show an average κ of 0.83 for type agreement and 0.84 for sense agreement on spotted relations.[6]

3.2 DR Alignment in TED-MDB

The primary focus of this thesis, referred to as relation linking (as explained in Chapter 2), is to connect the various elements of a discourse relation across different languages. This involves associating the labels for the type of relation, arguments, connectives, and their respective discourse senses. The objective is not to align words and sentences between languages, but rather to establish connections among connectives, sentences, or sentence components that form part of a discourse relation, along with their linguistic annotations. The outcome of this research is to facilitate access to discourse labels across texts in diverse languages at a structural level, thereby enabling easy navigation through the discourse structures of different languages by referencing existing labels [62].

[6] Kappa values falling within the range of 0.61–0.80 are typically classified as signifying substantial agreement, while those between 0.81–1.00 are viewed as representing nearly perfect agreement.

3.2.1 Difficulties on DR Alignment

Understanding and managing cross-lingual variations is crucial when connecting two sets of relations. Translated texts exhibit differences across multiple dimensions [63], including coherence relations. Challenges can arise at various levels.

Discrepancies in argument spans of discourse relations between languages: For instance, in Example (10), none of the arguments align perfectly. While both arguments of the English connective are full clauses, the Turkish counterpart lacks this feature. In Turkish, subordinate clauses are formed with a non-finite predicate bearing one of the subordinating suffixes. As an annotation decision , the converb suffix -Ip,when used as a simplex subordinator, annotated as a connective linking one non-finite and one finite VP with a shared subject.([64] and [65]). Therefore, in the Turkish relation, Arg2 comprises only the verb 'tut-' (meaning 'take' in this context) without an object or subject. Although the subject information is provided in the other argument, Turkish Arg2 still omits the object referring to The Hubble Space Telescope due to Turkish's ability to omit objects in certain cases. While it was possible to translate the relation into Turkish by mirroring the syntactic structure of English, the translator chose not to do so.

(10) ... *we take the Hubble Space Telescope* and **we turn it around ...**
[Explicit, Expansion: Conjunction] (English, TED Talk 1976)

Hubble Uzay Teleskobu'nu **tut**up *döndür-düğümüzü ...*
[Explicit, Expansion: Conjunction] (Turkish, TED Talk 1976)

Annotators' choice of different Discourse Relation (DR) Types for the same text span: In Example (11), while one annotator identified a lack of coherence in the English text, the other annotator marked it as an Implicit Expansion relation. Such discrepancies might adversely affect the performance of the algorithm:

(11) Now *over almost eight years, they've outperformed by about two-thirds*. So **yes, this is correlation**.
[NoRel] (English, TED Talk 2150)

Yaklaşık sekiz sene boyunca, yaklaşık üçte iki oranda daha fazla performans gösterdiler. (yani) **Evet bu korelasyon**.
[Implicit, Expansion] (Turkish, TED Talk 2150)

Structural changes in translation lead to semantically different translations: In the example below, where two connectives are similar in terms of discourse relation type and senses, they connect different abstract objects and thus should not be aligned. Any alignment algorithm sensitive to semantic content should be able to handle this. Translation quality is crucial for semantically oriented alignment algorithms.

(12) When we think about mapping cities, we tend to think about roads and streets and buildings, and the settlement narrative that led to their creation, or you might think about the bold vision

of an urban designer, but there are other ways to think about *mapping cities* and **how they got to be made**.
[Explicit, Expansion: Conjunction] (English, TED Talk no. 2150)

Şehirlerin haritalarını oluşturmayı düşündüğümüzde yollar, sokaklar, caddeler, binalar ve şehirlerin oluşumuna yol açan yerleşim hikayeleri aklımıza gelir. Ya da bir kentsel tasarımcının cesur vizyonunu düşünebilirsiniz. Ancak, şehirlerin haritalarını oluşturmayı *düşünmenin* ve **yapmanın** başka yolları da var.
[Explicit, Expansion: Conjunction] (Turkish, TED Talk no. 2150)

Omission and Inclusion of Discourse Relations in Translation: In translation, a discourse relation (DR) might lose its connective function in the target language, as seen in Example (13), where the English connective *and* loses its connective function. Conversely, the translation of the source text may require the addition of a discourse connective, as illustrated in Example (14), where an additional discourse connective *için* is included in the Turkish translation. An alignment algorithm should be able to detect and isolate these tokens from the discourse relations (whose locations were shown in the Examples (13) and (14)) which should be aligned.

(13) And(Explicit,Expansion:Conjunction) *they are really complex* and (consequently,Implicit, Contingency:Cause:Result) **they can seem really far off**, that the temptation may be to do this: (actually,Implicit,Expansion:Level-of-detail:Arg2-as-detail) bury our heads in the sand and(Explicit ,Expansion:Conjunction) not think about it.
[Explicit, Expansion: Conjunction] (English, TED Talk no. 1927)

(ve,Implicit,Expansion:Conjunction)Gerçekten de karmaşık ve uzak görünebilirler, ki bu da şunu yapmamızı cazip kılabilir: (EntRel) (şöyle ki,Implicit,Expansion:Level-of-detail:Arg2-as-detail)Kafamızı kuma gömüp(Explicit,Expansion:Conjunction), bunun hakkında düşünmemek.

(14) What if(AltLex,Contingency:Condition:Arg2-as-cond) they used that firepower to allocate more of their capital to companies working the hardest at solving these challenges or(Explicit, Expansion:Disjunction) at least not exacerbating them?

(fakat,Implicit,Expansion.Conjunction)Bu gücü **bu sorunları çözmek** için *en fazla çalışan* veya(Explicit,Expansion:Disjunction) en azından bunları kötüleştirmeyen şirketlere daha fazla sermaye aktarmak için(Explicit, Contingency.Purpose.Arg2-as-goal) kullansalar (Explicit, Contingency:Condition+SpeechAct) ne olurdu?
[Explicit, Contingency:Purpose:Arg2-as-goal] (Turkish, TED Talk no. 1927)

Linked Discourse Relations Located in Different Bitext Units: A monotonic, sentence-ordered alignment algorithm can capture relations only within bisentences. Relations occurring outside a bisentence unit cannot be aligned. In Example (15), an AltLex relation was annotated in the first sentence in English, whereas its equivalent discourse relation, an entity relation, was annotated in the second sentence. Therefore, whether the alignment algorithm should cover such cases is also an issue of concern:

(15) *Who here knows that in many cities across the United States it is now illegal to sit on the sidewalk, to wrap oneself in a blanket, to sleep in your own car, to offer food to a stranger?* **I know about these laws because I've watched as friends and other travelers were hauled off to jail or received citations for committing these so-called crimes.**
[AltLex, Hypophora] (English, TED Talk no. 2009)

Burada kim Birleşik Devletler'de birçok şehirde yol kenarına battaniyeye sarılı oturmanın yasadışı olduğunu biliyor veya arabanızda uyumanın, ya da yabancılara yiyecek önermenin? **Bu kanunları biliyorum çünkü birkaç arkadaşın ve diğer gezginlerin suç denen bu şeylerden dolayı hapishaneye götürüldüğünü veya uyarı aldığına şahit oldum.**
[EntRel] (Turkish, TED Talk no. 2009)

Multiple DRs in the Same/Similar Text Span: According to the PDTB 3.0 annotation manual, there may be more than one DR sharing arguments, known as 'linked' relations. For instance, in Example (16), the English sentence contains two relations, *and* and the implicit *then*, marked on identical text spans. In contrast, in Turkish, only one discourse relation is identified in the same text span. In Example (17), arg2 of the discourse relation in English contains all the segments of the discourse relation in Turkish and they share the Level-I sense. DR alignment algorithm should be able to identify these two cases as distinct.

(16) *they open up* and (then) **the telescope turns around**
[Explicit, Expansion: Conjunction] (English, TED Talk no. 1976)

Yapraklar açılıp genişliyor (ve) **Teleskop yön değiştiriyor.**[Implicit, Temporal:Asynchronous] (English,TED Talk no. 1976)

(17) *thats not all* (additionally) **Theyre economic issues, and that makes them relevant to risk and return**
[Implicit,Expansion:Level-of-detail:Arg2-as-detail] (English, TED Talk no. 1927)

Bunlar ekonomik sorunlar ve **bu da bu sorunları risk ve kazanç ile ilgili hâle getiriyor**.[Explicit, Expansion.Conjunction] (Turkish,TED Talk no. 1927)

In the remainder of this chapter, particularly in Section 3.2 and the following sections, the DR alignment methodology, along with its evaluation and error analysis, will be introduced primarily based on [1].

3.2.2 DR Alignment Algorithm

Discourse Relation Alignment Algorithm [1] leverages recent advancements in multilingual embeddings to assign similar representations to semantically analogous linguistic units across different languages, thereby facilitating their mapping. This method, based on a previous study [10] focusing

specifically on the English-Turkish pair within TED-MDB, begins with a preprocessing step (the method's pipeline is depicted in Figure 9).[7]

In TED-MDB, as described before, annotation files and raw data as kept as separate files, with raw files untokenized and unaligned format(see Figures 10 and 11. The Example (18) used in Section 3.2.2.2 is shown as highlighted in both English and Turkish files).

Initially, DR annotations stored in pipe-delimited files are transferred to the base text files of both languages and assigned unique IDs.(see Figures 12 and 13)

Subsequently, word and punctuation tokenization as well as sentence alignment procedures are conducted using the UPlug tool [8] [67] and corresponding TMX sentence alignment files are generated as in Figure 14. The sentence tokenization model in UPlug employs an unsupervised algorithm for sentence boundary detection [68], while the sentence alignment algorithm is inspired by Phillip Koehn's approach for aligning sentences in the Europarl corpus [69]. This method uses Gale-Church [11] sentence-length information and a dictionary; if no dictionary is provided initially, alignment occurs through two passes based on sentence length.

Relation linking within this framework operates within bitext units that consist of source and target sentences exhibiting either partial or full translation equivalence [22]. Discontinuous segments are disregarded as TED MDB exclusively focuses on adjacent argument-connective triplets [6]. In relation linking, relations from each bitext unit are paired to form relation matrices. For pairs surpassing a specific semantic similarity threshold established during training, a composite score is computed. This score takes into account agreement across all three sense levels and the type of relation present in the matched pair (disregarding relation type match if there's no Level1 sense match), along with the semantic similarity between text segments (Arg1 + connective + Arg2). Semantic similarity is calculated using cosine similarity between LASER embeddings of each relation's text segments [70]. Connective word is included into the segments for Explicit, AltLex and Implicit discourse relations.

3.2.2.1 Obtaining the semantic similarity threshold

To determine the semantic threshold parameter, training is conducted on language pairs involving the source language and three target languages in TED-MDB: Turkish, Portuguese, and Lithuanian (EN-TR, EN-PT, EN-LT). Initially, relation labels from English texts are automatically aligned with those in the texts of these three languages. Subsequently, the accuracy of this automatic process was verified manually to correct any incorrect matches. The training phase involves further evaluation

[7] In the first version of the alignment algorithm, preprocessing and DR sense-based scoring logic are identical to the new version. The main difference lies in the calculation of semantic similarity of the DR text spans. If Level1 senses of the DR pairs match, arguments in the English DR segment are translated into Turkish, and for each translated argument and the original Turkish argument (arg1EnT-arg1Tr, arg1EnT-arg2Tr, arg2EnT-arg1Tr, arg2EnT-arg2Tr), the BLEU score is calculated. BLEU [66] is a standard metric used in machine translation (MT) performance evaluation. This measure relies on word n-gram overlaps between source language (SL) and target language (TL) texts. The process is repeated by translating the Turkish arguments into English and assigning them BLEU scores. The maximum BLEU scores of each calculation are selected, summed, and added to the sense/type scores. All translations are done using the Google Translate API. Providing a small part of the DR relation for translation leads to translation errors and adds to inherent translation noise. Therefore, the algorithm was modified.

[8] https://github.com/Helsinki-NLP/Uplug

using this manually validated data, referred to as manually-corrected or semi-automatically linked data throughout the thesis.

For training purposes, the six English files are divided into training and test datasets based on the overall relation counts in the English texts. To address potential overfitting or underestimation due to limited data size, a random and equal split is made for training and test sets proportional to total relation counts. The specific ratio of train:test data per relation is 52:48.

In order to have a representative training set, four talks (IDs 1971, 1978, 2009, and 2150) are designated as training data, while the remaining two are assigned to the test set. By incrementing semantic threshold values from 0 to 0.95 at intervals of 0.05, the algorithm is iterated upon. The optimal threshold value that maximizes the F-score across all language pairs is chosen and verified using the test set before application to other language pairs.

Figure 15 illustrates how the semantic threshold impacts relation linking performance based on evaluation metrics. For clarity, figures start at 0.35; however, performance remains consistent between thresholds of 0 and 0.55 across languages. Notably, effects become noticeable around a threshold of 0.6 for all languages, with peak performance observed between ranges such as 0.6–0.7 for European Portuguese and 0.65–0.75 for Lithuanian, while Turkish's performance declines rapidly post-threshold of 0.65. Within thresholds of 0 to 0.55, average F-scores are 0.82 for Lithuanian and 0.88 for the other language pairs; yet maintaining thresholds within this range leads to an increase in False Positives due to minimal control over this parameter. The model's reliance solely on similarity at sense levels and relation types can result in erroneous linkages between English relations and target relations, as demonstrated by Example (12), where an English relation connected by *and* incorrectly links to its Turkish counterpart *ve*.

3.2.2.2 Relation Linking Algorithm

During the phase of establishing connections between relations, a scoring system is formulated based on factors such as semantic similarity, correspondence at various sense levels, and relation types. The process involves several steps:

1. Initially, the similarity score between pairs of relations is computed by analyzing their text segments. Relations failing to meet the similarity threshold (0.65), as set in a previous step (outlined in Section 4.3.1), are eliminated. Semantic similarity is calculated using cosine similarity between LASER embeddings ([70]) of the text spans for each relation. LASER, which supports 93 languages, embeds sentences into a shared space where semantically similar sentences receive similar representations, surpassing the traditional bag-of-words approach. Furthermore, semantic similarity among discourse connectives is examined to address scenarios involving multiple discourse relations within the same text segment such as the Example (17). As LASER provides vectors for the semantic representation of sentences, connective similarity is calculated using FastText word vectors [71] and checked against the semantic similarity threshold without being added to the total score.

2. The semantic similarity score is augmented by another score that reflects how well relation pairs match in each source-target language pair at different sense levels and relation types. Matched Level1 senses are given 1000 points, Level2 senses 100 points, Level3 senses 10 points, mirror-

ing the PDTB sense hierarchy where a higher level is considered more precise, and one point is given for identical relation types (such as Explicit or Implicit). If there is no Level1 sense match, scoring for other sense levels and relation types is not carried out; however, for NoRels and EntRels, in particular, matching of relation realization types is crucial due to the absence of a sense label.

3. Subsequently, the target relation with the highest combined score is linked to each source relation; this process continues iteratively until all relation pairs are exhausted from the matrices. To account for cases such as when an AltLex matches with a succeeding sentence unit in the target language (as shown in Example (15)), additional comparisons are made. In instances where a source relation does not find a match in its parallel unit in the target language, it undergoes another evaluation in the subsequent alignment unit.

The entire process is demonstrated using a provided sample text in Example (18). This text features three explicit relations in both English (EN) and Turkish (TR), indicated by the connectives (*but*, *as*, *and*) and their Turkish equivalents (*ama* for 'but', *kadar* for 'as', and *ve* for 'and'), as outlined in Example (18). Initially, all possible combinations of these relations are computed, resulting in a (3x3) matrix displayed in Table (18). Subsequently, each pair undergoes a scoring evaluation.

In the case of the Turkish language, the connective label *ama* aligns with the English connective label *but* in terms of relation realization type and sense across all levels, as depicted in the first column of Table (18). Similarly, *kadar* corresponds to *as*, and *ve* corresponds to *and*. However, when examining a non-matching scenario such as the third English relation conveyed by *and*, which has no correspondence with the first Turkish connective *ama* at any sense levels, matching in the relation realization type is not taken into consideration between *ama* and *and*. Finally, each source relation (each row) is connected to the target relation (each column) that yields the highest score during this process.

(18) Years have passed, but many of the adventures I fantasized about as a child – traveling and weaving my way between worlds other than my own — have become realities through my work as a documentary photographer. But no other experience has felt as true to my childhood dreams as living amongst and documenting the lives of fellow wanderers across the United States. (English, TED Talk no. 2009)

Yıllar geçti, ama çocuk olarak hayalini kurduğum birçok macera – benim dünyam dışındaki dünyalar arasında seyahat ederken ve yoluma dokunurken – bir belgesel fotorafçısı olarak işim aracıyla bunlar gerçek oldu. Ama hiçbir başka deneyim çocukluk rüyalarımı yaşayanlar arasında olmak kadar ve Birleşik Devlet boyunca gezgin arkadaşların arasında yaşamak kadar gerçek hissettirmedi. (Turkish, TED Talk no. 2009)

English :

- DR11[9]-Explicit-Comparison.Concession.Arg2-as-denier-DC[10]-**But**
- DR12-Explicit-Comparison.Similarity-DC-**as**
- DR13-Explicit-Expansion.Conjunction-DC-**and**

Turkish:

[9] DR stands for Discourse Relation.
[10] DC is used for Discourse Connective.

•DR13-Explicit-Comparison.Concession.Arg2-as-denier-DC-**Ama**

•DR14-Explicit-Comparison.Similarity-DC-**kadar**

•DR15-Explicit-Expansion.Conjunction-DC-**ve**

Table 7: The Relation Scoring Matrix for Example (18). The numbers refer to the calculated scores based on sense/type agreement + semantic similarity of the segments (Arg1 + Conn (if available) + Arg2).

	Ama	**kadar**	**ve**
But	**1111+0.85**	1001+0.79	0+0.72
and	0+0.69	0+0.71	**1111+0.75**
as	1001+0.8	**1111+0.85**	0 + 0.77

3.2.3 Method Evaluation and Error Analysis

Quantitatively assessing alignment can be complex due to several factors. Firstly, aligning an entire structure presents challenges in breaking it down into smaller, evaluable components. Secondly, the concept of alignment involves bisegmentation, requiring both proper segmentation into alignable units and accurate mapping of these units. The presence of unaligned elements and one-to-one mappings further complicates quantitative analysis. Establishing a definitive standard for precise alignments is challenging and varies based on the specific task and domain context. In this study, a stringent approach similar to that of [6] was not adopted for identifying gold alignments; exact argument span matching was not obligatory either. Instead, accepting the match between the endpoint of one text span and the beginning of another was regarded sufficient, with linear ordering considered as a general tradition [72].

Alignment typically serves as a foundational component within larger applications such as machine translation (MT). Performance metrics often focus on the outputs of MT systems; however, this study prioritizes evaluating discourse relation (DR) alignment links as the primary focus. Evaluation criteria draw from information retrieval concepts such as precision (Equation 1), recall (Equation 2), and F-score (Equation 3) to accurately measure the quality of data linking. The β parameter is used as a weighting factor between precision and recall; in this study, it is set to 1, as both measures are essential for the DR alignment task.

Precision denotes the proportion of correct matches among assigned links (true positives), recall measures how accurately true matches are identified, and accuracy reflects the correct identification across valid matches and non-matches. The F-score provides a comprehensive measure by combining precision and recall performance [73], [22], [74], [72].

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{F1 Score} = \frac{(1 + \beta^2) \times \text{Precision} \times \text{Recall}}{(\beta^2 \times \text{Precision}) + \text{Recall}} \tag{3}$$

Striking a balance between precision and recall is crucial; while accuracy may be less effective when dealing with numerous non-linking points in source and target datasets, it remains essential for relation linking tasks like ours. Understanding non-linking relations offers valuable insights into linguistic nuances, machine translation dynamics, and annotation quality assessments.

Discourse Relation (DR) alignment algorithm will be evaluated against established gold DR alignments for English-Lithuanian, English-European Portuguese, and English-Turkish, as provided by the TED-MDB annotators specific to each language pair. Alignment performance for each language pair will be detailed using precision, recall, and F-score metrics. Each DR alignment is assessed by determining if a source language's DR corresponds to a target language's DR as captured by the alignment algorithm: True Positives (TP) indicate correct matches, while False Positives (FP) denote incorrect pairings. True Negatives (TN) occur when an unmatched source DR remains unpaired in the target language, whereas False Negatives (FN) signify incorrect connections between non-matching source DRs and target DRs.

Evaluations are conducted uniquely for each language pair, considering both source-to-target and vice versa directions, to ensure a comprehensive analysis across the varied relations present in the alignment data. For this study, unaligned target language (TL) DR tokens are as essential as unaligned source language (SL) tokens.

In Table 8, evaluation results are presented. Overall, DR alignment performance is quite good (>0.9) for all three language pairs. In the next section, I will focus on the specific alignment error cases.

Table 8: Quality metrics for each language obtained in two test files

Lang. Pair	TP	FN	FP	TN	Accuracy	Precision	Recall	F-Score
EN LT	288	2	15	41	0.95	0.95	0.99	**0.97**
LT EN	288	1	15	66	0.96	0.95	1	**0.97**
EN PT	273	5	37	31	0.88	0.88	0.98	**0.93**
PT EN	273	2	37	17	0.88	0.88	0.99	**0.93**
EN TR	279	10	37	20	0.86	0.88	0.97	**0.92**
TR EN	279	12	37	38	0.87	0.88	0.96	**0.92**

3.2.3.1 Error Analysis

When analyzing error cases, it is essential to recognize that they can stem from both data-specific and language-specific origins. TED translators often translate texts in fragments alongside videos to adhere to translation guidelines. However, this approach can occasionally result in a lack of overall coherence and alterations in the intended meaning within the discourse of the target language. This discrepancy poses challenges for discourse relation alignment algorithms as they struggle to align such relations across different languages.

One common issue arises when there are disparities in argument span lengths between the source and target relations during translation, especially with free translations or partial overlaps in argument spans. These scenarios lead to decreased performance, characterized by an increase in either False Negatives (refer to Example (19)) or False Positives (refer to Examples (20) and (21)). These fluctuations significantly impact various performance metrics.

Varying Argument Span Lengths: As outlined in Section 3.2.1, discrepancies in argument structure across languages can lead to alignment mismatches. For instance, Example (19) illustrates a case where a Lithuanian relation could not be mapped to its English equivalent due to a longer Arg2 span annotation in Lithuanian.

(19) Now these initiatives create a more mobile workplace, and *they reduce our real estate footprint*, and **they yield savings of 23 million dollars in operating costs annually**, and avoid the emissions of 100,000 metric tons of carbon.
[Explicit, Expansion: Conjunction] (English, TED Talk no. 1927)

To rezultatai šiandien – mobilesnės darbo vietos, mažinančios mūsų nekilnojamojo turto pėdsaką, o **tai leidžia sutaupyti 23 milijonus dolerių kasmetinių veiklos išlaidų ir sumažinti anglies dioksido išmetimą 100 000 metrinių tonų**.
[Explicit, Contingency: Cause: Result] (Lithuanian, TED Talk no. 1927)

Different Realizations of Discourse Relations: The process of translation can result in diverse interpretations of discourse relations. For instance, in Example (20), the semantic link conveyed by the term *and* in the English sentence is absent in Turkish. However, this English connection is subsumed within Arg2 of the Turkish relationship in a loosely translated form. Unfortunately, our relation alignment models mistakenly associate the English term *and* with the Turkish term *sanki*, which translates to 'as if'.

(20) Lord, grant that I desire more than I can accomplish, Michelangelo implored, as if to that Old Testament God on the Sistine Chapel, and he himself was that Adam *with his finger outstretched* and **not quite touching that God's hand**.
[Explicit, Expansion:Conjunction] (English, TED Talk no. 1978)

Tanrım, bana başarabileceğimden daha fazla-sını istemeyi bahşet, diye yakarmıştı Michelangelo, sanki **Sistina Şapeli'ndeki Eski Ahit Tanrısı'na ve kendisi de uzattığı parmağı Tanrı'nın eline tam değmeyen Âdem'di.**
[Explicit, Comparison:Similarity] (Turkish, TED Talk no. 1978)

Partially Overlapping Argument Spans: In certain instances, even when there is no direct alignment between source and target relations within a bitext unit, the argument spans may partially overlap. For instance, in Example (21), although the English implicit discourse relation indicated by *as a result* lacks an equivalent in the Portuguese sentence, an explicit relationship is expressed by the term *sem* ('without') in Portuguese, which partially overlaps with the Arg2 of the English discourse relation. Consequently, these two relations are incorrectly aligned.

(21) And *I saw that gave her more tenacity*, (implicit = as a result) **and she went after it again and again**.
[Implicit, Contingency:Cause:Result] (English, TED Talk no. 1978)

E vi que isso deu-lhe mais persistência, e *continuou, continuou*, sem **parar**.

[Explicit, Expansion: Conjunction] (Portuguese, TED Talk no. 1978)

3.2.4 Publishing the Linked Relations

The resulting DR alignment data is shared publicly in the form of XML files to support further studies[11]. Each language pair (English-Language X) has its connected and unconnected relationships stored in distinct XML files. An English-Turkish sample file is presented in Figure 16. It demonstrates the structure of the XML-formed linking data. Linked relations are categorized under $\langle linked\_relations \rangle$ while the unlinked ones are listed under $\langle non\_paired\_relations \rangle$.

Linked relations are encapsulated within the $\langle relation\_pair \rangle$ element tag. Under this element, SL and TL linked DRs are presented within the $\langle relation \rangle$ nodes, comprising $Arg1$, $Arg2$, $Conn$ (if applicable), and five attributes denoting the sense, and type details of the relationship, along with metadata such as its language, source TED Talk ID, and unique relation ID. The $\langle relation\_pair \rangle$ tag has only one attribute, which is the alignment score signifying the alignment method's confidence level in linking these discourse relations.

Unlinked relations lack $\langle relation\_pair \rangle$ tags and are composed of $\langle relation \rangle$ nodes only. However, we consider both linked and non-linked relations equally valuable for understanding discourse structures across languages. It is important to acknowledge that these links are generated automatically; thus, there might be some inaccuracies present which might require manual correction.

3.3 Summary

This chapter presented the core work of this thesis, focusing on the TED-MDB corpus, which serves as the primary data source for this study. TED-MDB is a publicly accessible collection of annotated discourse relations from TED talks, originally delivered in English and translated into six other languages: German, Polish, Russian, European Portuguese, Turkish, and later, Lithuanian. As these transcripts cover a broad range of topics, they provide an invaluable resource for linguistic analysis and comparative studies.

The TED-MDB was developed by an international team of researchers following the guidelines of the Penn Discourse Treebank (PDTB). This approach extends the PDTB framework to systematically annotate discourse relations in TED talks across multiple languages using a unified descriptive model. Consequently, the corpus serves as a significant tool for linguists, computational linguists, discourse analysts, translation professionals, and educators.

The chapter outlined the annotation schema used in TED-MDB, focusing on different types of discourse relations—Explicit, Implicit, AltLex, EntRel, and NoRel—and described the hierarchical sense system defined in PDTB. It also introduces the Discourse Relation (DR) alignment algorithm developed to link equivalent relations across translations in different languages.

[11] https://github.com/MurathanKurfali/Ted-MDB-Annotations

The algorithm employs multilingual embeddings to assign similar representations to semantically equivalent linguistic units across languages, using cosine similarity of LASER embeddings. The alignment process assigns composite scores based on semantic similarity, sense levels, and relation types. This procedure is trained and tested on language pairs involving English and three target languages (Turkish, Portuguese, and Lithuanian), and evaluated based on information retrieval metrics such as Precision, Recall, and F-score.

The chapter also analyzed challenges and errors in DR alignment, such as discrepancies in argument spans, different realization of discourse relations, partially overlapping argument spans, and the pitfalls arising from free translations. Error analysis identifies factors impacting the algorithm's performance, demonstrating the complexities inherent in aligning discourse relations across languages.

Finally, the chapter discussed the publication of the DR alignment data in XML format, detailing the structure of the linked and unlinked relations within XML files. The XML schema is designed to facilitate further studies and applications, acknowledging that the automatic generation of these links might require manual correction due to potential inaccuracies.

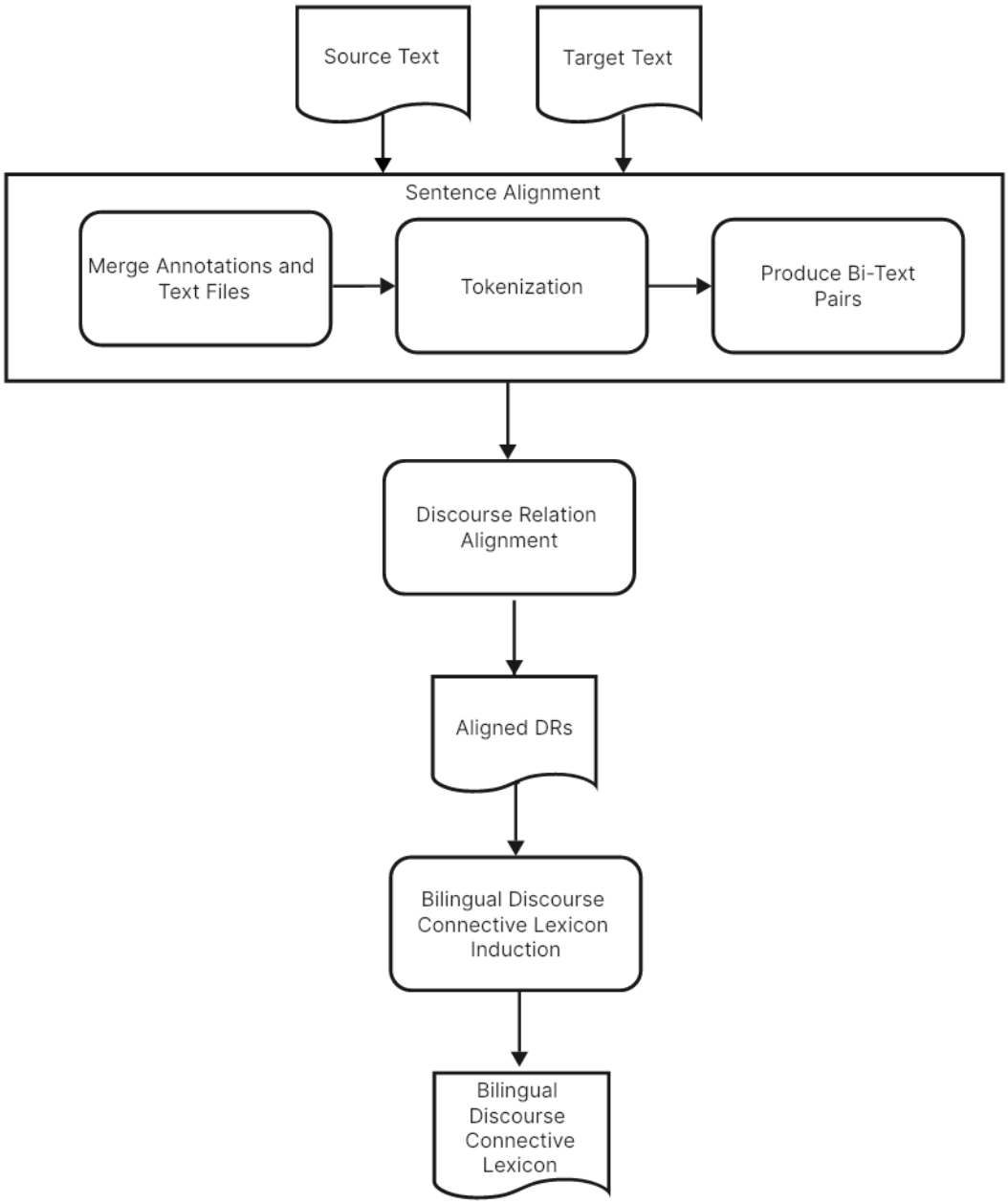
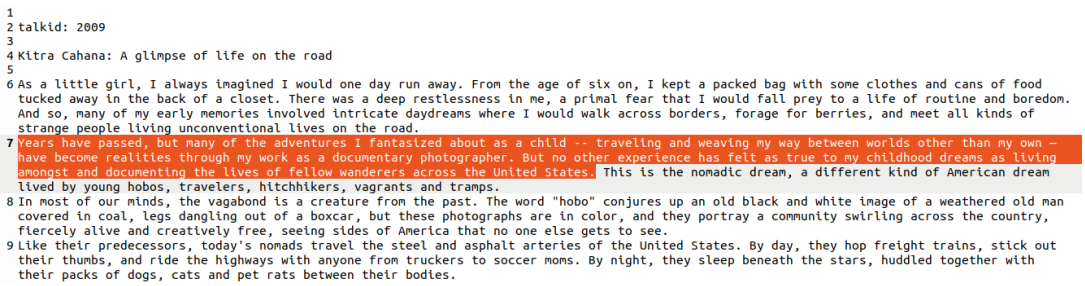
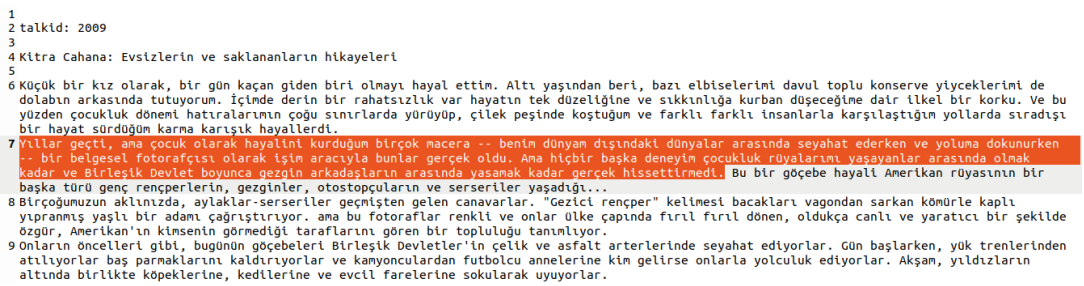


Figure 9: DR Alignment Pipeline



```
1
2 talkid: 2009
3
4 Kitra Cahana: A glimpse of life on the road
5
6 As a little girl, I always imagined I would one day run away. From the age of six on, 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.
  And so, many of my early memories involved intricate daydreams where I would walk across borders, forage for berries, and meet all kinds of
  strange people living unconventional lives on the road.
7 Years have passed, but many of the adventures I fantasized about as a child -- traveling and weaving my way between worlds other than my own —
  have become realities through my work as a documentary photographer. But no other experience has felt as true to my childhood dreams as living
  amongst and documenting the lives of fellow wanderers across the United States. This is the nomadic dream, a different kind of American dream
  lived by young hobos, travelers, hitchhikers, vagrants and tramps.
8 In most of our minds, the vagabond is a creature from the past. The word "hobo" conjures up an old black and white image of a weathered old man
  covered in coal, legs dangling out of a boxcar, but these photographs are in color, and they portray a community swirling across the country,
  fiercely alive and creatively free, seeing sides of America that no one else gets to see.
9 Like their predecessors, today's nomads travel the steel and asphalt arteries of the United States. By day, they hop freight trains, stick out
  their thumbs, and ride the highways with anyone from truckers to soccer moms. By night, they sleep beneath the stars, huddled together with
  their packs of dogs, cats and pet rats between their bodies.
```

Figure 10: A sample raw file segment from English TED Talks no:2009



```
1
2 talkid: 2009
3
4 Kitra Cahana: Evsizlerin ve saklananların hikayeleri
5
6 Küçük bir kız olarak, bir gün kaçan giden biri olmayı hayal ettim. Altı yaşından beri, bazı elbiselerimi davul toplu konserve yiyeceklerimi de
  dolabın arkasında tutuyorum. İçimde derin bir rahatsızlık var hayatın tek düzeliğine ve sıkkınlığa kurban düşeceğime dair ilkel bir korku. Ve bu
  yüzden çocukluk dönemi hatıralarımın çoğu sınırlarda yürüyüp, çilek peşinde koştuğum ve farklı farklı insanlarla karşılaştığım yollarda sıradışı
  bir hayat sürdüğüm karma karışık hayallerdi.
7 Yıllar geçti, ama çocuk olarak hayalini kurduğum birçok macera -- benim dünyam dışındaki dünyalar arasında seyahat ederken ve yoluma dokunurken
  -- bir belgesel fotorafçısı olarak işim aracıyla bunlar gerçek oldu. Ama hiçbir başka deneyim çocukluk rüyalarımı yaşayanlar arasında olmak
  kadar ve Birleşik Devlet boyunca gezgin arkadaşların arasında yasamak kadar gerçek hissettirmedi. Bu bir göçebe hayali Amerikan rüyasının bir
  başka türü genç rençperlerin, gezginler, otostopçuların ve serseriler yaşadığı...
8 Birçoğumuzun aklınızda, aylaklar-serseriler geçmişten gelen canavarlar. "Gezici rençper" kelimesi bacakları vagondan sarkan kömürle kaplı
  yıpranmış yaşlı bir adamı çağrıştırıyor. ama bu fotoraflar renkli ve onlar ülke çapında fırıl fırıl dönen, oldukça canlı ve yaratıcı bir şekilde
  özgür, Amerikan'ın kimsenin görmediği taraflarını gören bir topluluğu tanımlıyor.
9 Onların öncelleri gibi, bugünün göçebeleri Birleşik Devletler'in çelik ve asfalt arterlerinde seyahat ediyorlar. Gün başlarken, yük trenlerinden
  atlıyorlar baş parmaklarını kaldırıyorlar ve kamyoncular dan futbolcu annelerine kim gelirse onlarla yolculuk ediyorlar. Akşam, yıldızların
  altında birlikte köpeklerine, kedilerine ve evcil farelerine sokularak uyuyorlar.
```

Figure 11: A sample raw file segment from Turkish TED Talks no:2009

```
  Explicit|755..758|Wr|Comm|Null|Null|||Comparison.Concession.Arg2-as-denier||||||566..753|Inh|Null|Null|Null||759..907|Inh|Null|-
Null|Null|||||||755..758|DEFAULT||755|DR11|

Explicit|819..821|Wr|Comm|Null|Null|||Comparison.Similarity||||||759..818|Inh|Null|Null|Null||822..907|Inh|Null|Null|Null|||||||-
819..821|DEFAULT||819|DR12|

Explicit|837..840|Wr|Comm|Null|Null|||Expansion.Conjunction||||||822..836|Inh|Null|Null|Null||841..907|Inh|Null|Null|Null|||||||-
837..840|DEFAULT||837|DR13|


 Years have passed (DR6) , (DR7) but many of the adventures I fantasized about as a child -- (DR8) traveling (DR9) and weaving my
way between worlds other than my own — have become realities (DR10) through my work as a documentary photographer. (DR11) But no
other experience has felt as true to my childhood dreams (DR12) as living amongst (DR13) and documenting the lives of fellow
wanderers across the United States.
```

Figure 12: A sample raw file segment from English TED Talks no:2009 with annotations imported. Relevant annotation lines are provided also.

```
  Explicit|760..763|Wr|Comm|Null|Null|||Comparison.Concession.Arg2-as-denier||||||694..758|Inh|Null|Null|Null||764..927|Inh|Null|-
Null|Null|||||||760..763|DEFAULT||760|DR13|

Explicit|831..836|Wr|Comm|Null|Null|||Comparison.Similarity||||||837..927|Inh|Null|Null|Null||764..836|Inh|Null|Null|Null|||||||-
831..836|DEFAULT||831|DR14|

Explicit|837..839|Wr|Comm|Null|Null|||Expansion.Conjunction||||||785..836|Inh|Null|Null|Null||840..906|Inh|Null|Null|Null|||||||-
837..839|DEFAULT||837|DR15|


 (DR8) Yıllar geçti, (DR9) ama çocuk olarak hayalini kurduğum birçok macera -- benim dünyam dışındaki dünyalar arasında seyahat
(DR10) ederken (DR11) ve yoluma dokunurken -- bir belgesel fotorafçısı (DR12) olarak işim aracıyla bunlar gerçek oldu. (DR13) Ama
hiçbir başka deneyim çocukluk rüyalarımı yaşayanlar arasında olmak (DR14) kadar (DR15) ve Birleşik Devlet boyunca gezgin
arkadaşların arasında yasamak kadar gerçek hissettirmedi.
```

Figure 13: A sample raw file segment from Turkish TED Talks no:2009 with annotations imported. Relevant annotation lines are provided also.

```
    <tu>
      <tuv xml:lang="en"><seg> ( DR11 ) But no other experience has felt as true to my childhood dreams ( DR12 ) as living amongst ( DR13 )
and documenting the lives of fellow wanderers across the United States .  </seg></tuv>
      <tuv xml:lang="tr"><seg> ( DR13 ) Ama hiçbir başka deneyim çocukluk rüyalarımı yaşayanlar arasında olmak ( DR14 ) kadar ( DR15 ) ve
Birleşik Devlet boyunca gezgin arkadaşların arasında yasamak kadar gerçek hissettirmedi .  </seg></tuv>
    </tu>
```

Figure 14: A sample TMX file segment(TED Talks no:2009) which shows the sentence aligned text segment.

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

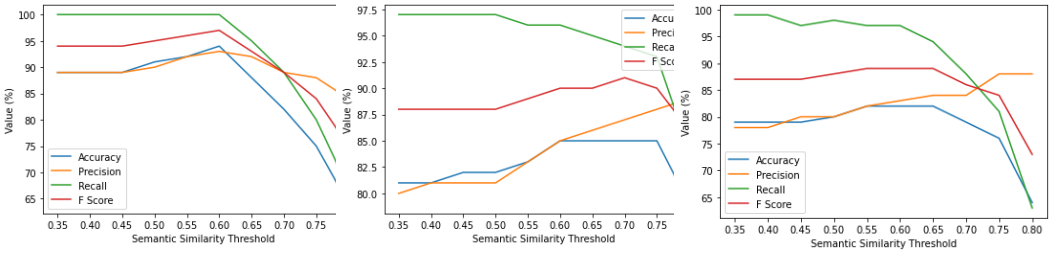


Figure 15: The evaluation metrics (Accuracy, Precision, Recall, and F-Score) vary across semantic threshold values. The thresholds range from 0 to 0.95 in increments of 0.05, but for clarity, only the data between 0.3 and 0.85 is presented to enhance visualization.

```
▼<doc>
  ▼<linked_relations>
    ▼<relation_pair alignment_score="1111.85">
      ▼<relation RelID="DR64" TalkID="talk_1978" Language="en" Type="Explicit" Sense="Contingency:Cause:Reason">
          <Arg1>I stayed</Arg1>
          <Arg2>I realized I was witnessing whats so rare to glimpse, that difference between success and mastery</Arg2>
          <Conn>because</Conn>
        </relation>
      ▼<relation RelID="DR58" TalkID="talk_1978" Language="tr" Type="Explicit" Sense="Contingency:Cause:Reason">
          <Arg1>orada kaldım</Arg1>
          <Arg2>çok nadir olarak görülecek bir şeye tanıklık ettiğimi anladım, başarı ve ustalık arasındaki o farka</Arg2>
          <Conn>çünkü</Conn>
        </relation>
      </relation_pair>
    </linked_relations>
  ▼<non_paired_relations>
    ▼<relation RelID="DR16" TalkID="talk_1978" Language="en" Type="Implicit" Sense="Expansion:Level-of-detail:Arg2-as-detail">
        <Arg1>this is the thing</Arg1>
        <Arg2>What gets us to convert success into mastery</Arg2>
        <Conn>actually</Conn>
      </relation>
    </non_paired_relations>
  </doc>
```

Figure 16: A sample file showing the structure of the adopted XML schema for the published relation alignments

CHAPTER 4

BILINGUAL DISCOURSE CONNECTIVE LEXICON INDUCTION

4.1 Introduction

Discourse connective lexicons are crucial in various fields, including automatic annotation of discourse connectives [4], second language learning, and machine translation. Variations exist in how languages convey discourse relations through connectives. From a syntactic perspective, certain connectives are restricted to sentence-initial positions, may have multiple POS tags based on context, or are expressed through morphology such as converbs in Turkish. Additionally, some connectives are predominantly used in formal contexts. Semantically, a discourse relation in a language can be articulated using multiple connectives. Even though TED-MDB is a small corpus consisting of only six texts, and considering that texts in languages other than English are translations of the English text, significant findings can still be uncovered by examining the DR alignment data. For instance, in TED-MDB, the Expansion.Conjunction relation is represented by 5 distinct connectives in English, 21 in Turkish, and 6 in European Portuguese and Lithuanian datasets. Focusing on the connective 'and', it appears 124 times in the English portion of the data. However, in 52% of instances, it was translated as 've' in Turkish, while it was omitted in 31% of cases. Notably, in the Turkish data, 8% of occurrences of 've' do not have a corresponding equivalent in the English text. Despite 'and' and 've' being accepted as translational equivalents, their usage patterns vary in natural language data. Furthermore, a single connective can convey multiple meanings based on the context. For example, 'Ancak' can indicate Expansion:Conjunction (Example (1)) or Comparison:Contrast (Example (2)). Given these complexities, it is not surprising that discourse connectives pose challenges for both first and second language learners ([5], [75], and [76]), as well as for human translators and machine translation systems ([77]).

(1) *Seviyorum, çünkü iklim değişikliği mevzusunun iki yönüyle de dalga geçiyor* (Ancak) **bununla ilgili gerçekten sevdiğim şey, bana Mark Twainin söylediği birşeyi hatırlatıyor olması, Gelecek için planla, çünkü orası hayatının geri kalanını geçireceğin yerdir.**
[Implicit, Expansion:Conjunction](English, TED Talk no. 1927)

(2) *Dünyadaki genel müdürlerin yaklaşık yüzde 80i sürdürülebilirliği inovasyonda artış ve endüstrilerinde rekabet avantajına erişmenin kaynağı olarak görüyor* (Ancak) **yüzde 93ü ÇSYyi gelecek olarak veya şirketlerinin geleceği için önemli olarak görüyor.**

[Implicit, Comparison:Contrast](Turkish, TED Talk no. 1927)

Only by examining bilingual discourse connective lexicons, it becomes challenging to conduct cross-lingual comparisons regarding the realization of discourse relations. Given all the aforementioned circumstances, a multilingual discourse connective lexicon derived from naturally occurring data, which considers co-occurrence frequencies, provides usage examples, and establishes relevant connections to other languages, would be advantageous across various domains such as human/machine translation [78], language learning/teaching [79], generating discourse-annotated corpora [80], as well as in shallow discourse parsing tasks like connective identification and sense classification [81]. This chapter will start with a concise overview of existing discourse connective lexicons. Subsequently, in Section 4.3, a bilingual lexicon induction algorithm will be presented. The chapter will conclude with evaluation of the induced lexicons, their limitations, and some statistics on the them. Lexicon data and the procedure presented in this chapter are based on the works in [82] and [1].

4.2 Background

Following the release of the PDTB 2.0 corpus [60] and the subsequent introduction of the TextLink project,a network dedicated to standardizing and improving the linking of cross-text and corpus linguistic data, there has been growing interest in developing resources annotated for discourse-level elements such as discourse relations and pronouns. The expansion of lexicon entries for discourse connectives has also witnessed a notable increase. The majority of existing discourse connective lexicons are monolingual, encompassing languages from diverse language families, including German [83], French [84], Italian [85], Czech [86], Portuguese [87], Turkish [88], and Bangla [89].

Despite the limited availability of multilingual discourse connective resources, several prominent alternatives are present. One notable resource is the Connective-Lex database [90], currently supporting 17 languages and accessible online at no cost. This database classifies discourse connectives based on linguistic characteristics such as spelling, grammatical category (e.g., coordinating conjunction, adverb, subordinating conjunction), and their conveyed functions (e.g., contrast, temporal, concession).

While the database accommodates multiple languages, its linking capability remains bilingual; for instance, the Portuguese lexicon is linked to English DiMLex, and the Italian lexicon is associated with German DiMLex. Backend storage of lexicons follows an XML format with standardized structuring, facilitating the inclusion of new lexicons. Other notable bilingual discourse connective lexicons include the Italian-German contrastive/concessive connective lexicon [78], GeCzLex, Anaphoric Connective Lexicon for Czech and German [91], and TED-MDB lexicons for Turkish-English and Portuguese-English originating from aligned relations (Tr-EnConnLex, Pt-EnConnLex) [82].

Furthermore, by applying Linked Lexical Open Data (LLOD) principles [92], discourse connectives have been derived from lexical knowledge graphs. Despite the value offered by these resources, certain limitations persist concerning covered discourse relation types, machine-readability status, multilinguality, scope of sense inventory, etc. Hence, further advancement of comprehensive multilingual discourse connective lexicons in standardized formats is imperative to enhance linguistic research and extend the benefits from monolingual to multilingual contexts.

4.3 Populating lexicon entries automatically

Developing a bilingual discourse connective lexicon poses a challenging task ([78] and [91]). The initial phase of this task typically involves extracting translation candidate tables from a substantial parallel corpus comprising a minimum of 2 million parallel sentence pairs. These candidates then undergo filtration to retain translations that align with the same sense as the source language connective, guided by the monolingual discourse connective lexicons of the respective languages [91].

Although such resources may not always be readily accessible for all languages within datasets like TED-MDB (e.g., Lithuanian, Russian, Polish), leveraging multilingual resources such as TED-MDB annotated for discourse connectives enables the extraction of lexicon entries from discourse relation annotations. This method of using annotated resources at the discourse level for inducing lexicons is well-documented in the current literature ([93] and [94]). By compiling bilingual discourse connective lexicons from annotated resources that elucidate their contexts and usages through discourse relations, researchers can access comprehensive information on the senses, argument structures, and relation types of these connectors across languages. In Chapter 3 where discourse relation alignment algorithm is introduced, a general workflow of the procedure is presented in Figure 9. In this pipeline, input to the bilingual lexicon induction process is aligned discourse relations.

- DR alignment data initially undergoes filtration to encompass solely Explicit or Implicit discourse relation types. Following this, a secondary filtering step eliminates pairs that do not share the same sense to minimize translation-related noise within the dataset. To broaden lexicon coverage, Implicit discourse connectives are also added. These Implicit connectives are chosen from a predefined list within the annotator tool and subsequently validated by annotators [60]. An included Implicit connective is regarded as the most appropriate overt indicator for a specific implicit relation. Thus, akin to explicit connectives derived from DR alignment data being considered valid entries, a parallel reasoning can be applied to implicit discourse relation connectives. Nonetheless, to enhance organizational clarity and support future research endeavors, separate entries are established for explicit and implicit usages within each bilingual lexicon. Also, there are instances where incorrect annotation span selection for discourse connectives in explicit discourse relations led to the inclusion of punctuation marks in the connective annotations. For such erroneous cases, the discourse connectives are stripped of punctuation marks. All connectives are converted to lowercase to eliminate case variations.

  Within Turkish-English discourse connective lexicons, morphological variations of identical connectives are grouped together (see Table 9). This grouping is done with regular expressions and also checked manually.

- Potential meanings for each connective in the source language (SL) are enumerated. For example, for Turkish connective *Böylece*, there is only one sense type: *Contingency:Cause:Result*

- For every SL connective, corresponding translation pairs in the target language (TL) are identified using the DR alignment data. The identified translations are then categorized under the previously listed senses. This categorization is structured in sequential order based on the frequency of occurrence of each translation candidate with that SL connective under the specified sense. Separate entries are created for each conveyed sense of the SL connective to reflect its polysemous nature. For example, in Tr-En, the "but/ama" pair is categorized under Comparison:Concession:Arg2-as-denier, Comparison:Contrast sense, and Expansion: Conjunc-

Table 9: List of Suffixal Connective Types and Their Example Tokens

Connective	**Examples**
-(y)ArAk	azaltarak, edilerek, yaparak, uzaklaşarak, saçılarak, düşünerek, bakarak, olarak, sokularak, çalışarak, düşerek
-(y)ArAk..-(y)ArAk	yiyerek..uyuyarak
-(y)HncA	düşününce
-(y)HncA dA	yapınca da
-(y)Hp	çalışıp, biriktirip, gömüp, bakıp, doğup, açılıp, çevirip, alıp, tutup, fotoğraflayıp, kaldırıp, yapıp, inceltip, olmayıp, alıp, durup, geçip, olup, karıştırıp, yürüyüp
-dA	olduğumuzda, kaçtığımızda, sorduğunda, baktığımızda, karşılaştırıldığında, yaptığımızda, gördüğümde, gittiğimde, düşündüğümüzde
-dAn	keşfetmemizden
-dAn kaynaklanmak	keşfetmemizden kaynaklanıyor
-ken	izlerken, üzereyken, gülümserken, yaklaşırken, giderken, dururken, kaldırırken, görürken, bulurken, başlarken
-sA	olursa, görürse, istiyorsan, karşılaştırırsak, ederlerse, kullansalar, yapsak, abartılmışsa, büyütülmüşse, değilse, şeyse, yaşıyorsanız, yapabilirseniz, gösteriyorsa, tabi alabilirse, ne kadar..olursa olsun, değilse, yaparsak, edebilirsek, yapabilirsek, edebilirsek, götürebilirsek, yumuşatabilirsek, istesem, istesem, yapamazsan, bilmeseniz, devam edersek, düşünürseniz, deniyorsak, sahipsek, başlayabilirsek, aşırılılıklarındansa
-sA bile	olsaydık bile
-sA dA	gelse de
-sA..-sA	koyarsak..getirirsek

tion senses arranged by frequency of appearance. For the Turkish connective *Böylece*, there are three translation candidates in English, with the last two being of the implicit type: *so* (frequency = 3), *as a result* (frequency = 2), and *consequently* (frequency = 1). Since all three candidates share the sense *Contingency:Cause:Result*, they are listed under this sense and ordered by their frequency of occurrence.

4.3.1 Lexicons

TED-MDB lexicons are structured consistently and a template-based approach was implemented. This method facilitated the automated production and maintenance of the lexicon web pages, and ensured standardization in their representation through the use of common CSS files. Each connective entry created as web pages from templates in a cascaded manner and were shared on a Java-based website[1] for public accessibility. Section 4.3.1.1 describes the template structure. Following this, the information presented to the end user in the front-end is detailed in Section 4.3.1.2.

[1] http://metu-db.info/mdb/ted/resources.jsf

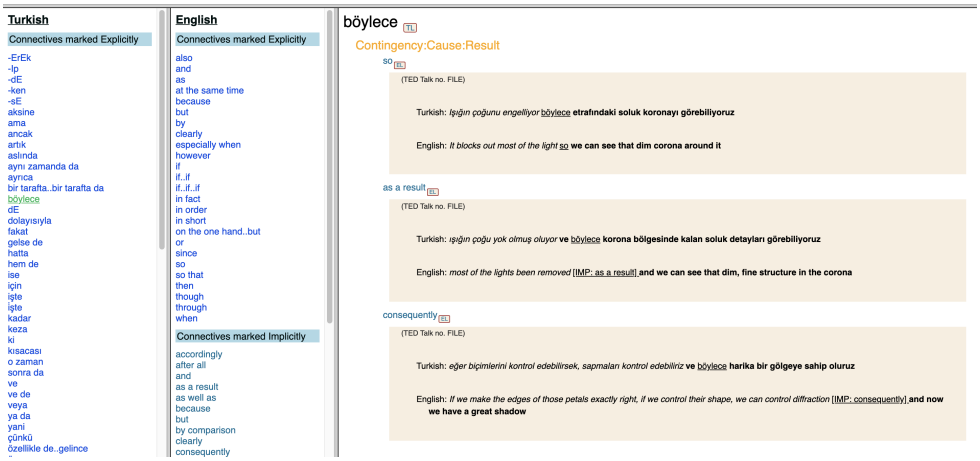


Figure 17: A screenshot showing the entry for "böylece" in the Turkish-English lexicon.

4.3.1.1 Template Structure

Template files were created in a hierarchical manner. Figure 18 illustrates the general structure of the lexicon web page. There is an n-to-1 relationship from inner templates to outer templates. This means that a ‘conlist.template‘ class file(see Figure 19) may contain multiple ‘conn.template‘ classes(see Figure 20), and a ‘conn.template‘ may contain multiple ‘conn.usage.template‘ classes (see Figure 21), and so forth.

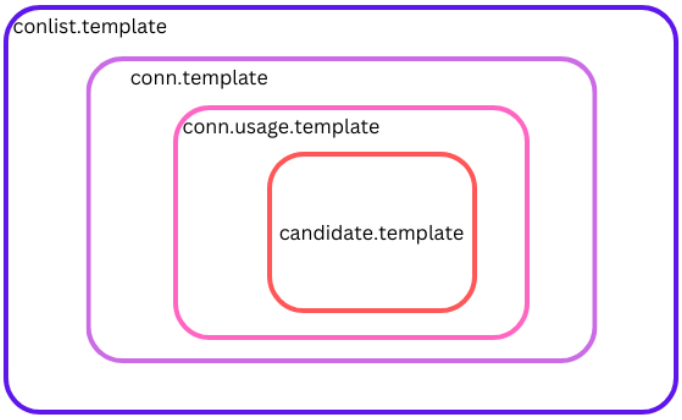


Figure 18: General Template Structure of the Lexicon Web Page

In each template, the variables that need to be filled automatically are highlighted in the `{{ }}` format. For each connective, a ‘conn.template‘, which serves as the main HTML page for each entry, is constructed. For each distinct sense of the entry, a corresponding ‘conn.usage.template‘ is created and placed within the `{{main_content}}` variable of the ‘conn.template‘.

```
<meta http-equiv="Content-Type" content="text/html; charset=utf-8">
<link href="../css/lexStyle.css" rel="stylesheet" type="text/css" />
</head>
<body>
    <div class="list_header">{{lang}}</div>
    <div class="sublist_heading_exp">Explicit Connectives</div>
    {{explicit_con_list}}

    <br>
    <div class="sublist_heading_imp">Implicit Connectives</div>
      {{implicit_con_list}}
    <br>
</body>
</html>
```

Figure 19: General Template Structure of the Connective Lists

```
    <title>TrEnConnLex 1.0</title>
     <meta http-equiv="Content-Type" content="text/html; charset=utf-8">
     <link href="../../../../css/lexStyle.css" rel="stylesheet" type="text/css" />
  </head>
  <body>
     <span class="lemma">{{eng_connective}}</span>
     <span class="lexlink"><a
       href="{{src_dimlex_link}}"
       target="_blank">{{src_dimlex_link_abb}}</a></span>
     <div class="conn_usage_div">
       {{main_content}}
     </div>
  </body>
</html>
```

Figure 20: General Template Structure of Connective Entry

```
<span class="conn_usage">{{eng_sense}}</span>
<div class="complex_forms_div">
    {{main_content}}
</div>
```

Figure 21: General Template Structure of a Single Sense Representation for an Entry

For each translation candidate associated with each sense, a ‘candidate.template‘ (see Figure 22) is created and placed within the `{{main_content}}` variable of the ‘conn.usage.template‘. Finally, the ‘conn.template‘ is saved as an HTML file named after the connective, and its URL link is added to either the `{{explicit_con_list}}` or `{{implicit_con_list}}` lists, depending on whether the main entry’s discourse relation type is Explicit or Implicit.

Final HTML code of *’Böylece’* can be seen in Figure 23

4.3.1.2 Lexicon GUI Structure

Each entry within the lexicon comprises the following components:

```
<span class="complex_form"><a href="{{conn-for-url}}{{other-type}}.html">{{other-conn-generic}}</a>
    <span class="exlink"><a href="{{dimlex_link}}" target="_blank">{{dimlex_link_abb}}</a></span>
    <span class="example"> <span class="exampleId"> (TED Talk no. {{file}}) </span>
    <span class="sent">
      {{lang}}: {{example_sentence}}
    </span>
    <span class="sent">
      {{other-lang}}: {{example_sentence_other}}
    </span>
    </span>
</span>
```

Figure 22: General template structure of a translation candidate representation for an entry

Figure 23: A screenshot showing HTML code for the entry "böylece" in the Turkish-English lexicon.

- **Connective**: Every lexicon entry, depicted in lowercase letters, is linked to a discourse connective. These connectives can take various forms including single-word (e.g., 'aksine'), multi-word (e.g., 'hem de'), discontinuous (e.g., 'özellikle de..gelince'), or in suffix form (e.g., 'dE', '-sE').

- **DiMLex link**: The annotations within TED-MDB, encompassing the lexicons, do not contain additional details such as syntactic/orthographic elements or Parts of Speech (PoS) pertaining to discourse connectives. To address this gap and establish a connection between bilingual and monolingual lexicons, each discourse connective along with its translations is linked to their respective entry on connective-lex[2] platform. However, not all languages(Lithuanian, Polish, Russian) within TED-MDB possess this linkage.

- **Sense list**: The entries include a comprehensive list of observed senses (based on the PDTB 3.0 sense hierarchy) associated with the primary connective in TED-MDB.

- **List of translation candidates**: Under each observed sense, potential translation candidates in the target language are presented. These candidates have dedicated entries within the lexicons accessible via URLs directly.

- **Example sentence**: To offer contextual insights into the occurrences of discourse connectives, every translation candidate is accompanied by an example relation pair extracted from TED-MDB. Additionally, the TED Talk ID is provided for further reference; furthermore, arguments

[2] http://connective-lex.info/.

within each example sentence are formatted distinctly (e.g., *arg1* in italics and **arg2** in bold) to emphasize argument-related information.

A model lexicon entry is illustrated in Figure 17.

Details regarding the coverage of each lexicon can be investigated in Table 10. Within this table, the Exp and Imp columns indicate the quantities of connectives from Explicit and Implicit relations. The total count of connectives is calculated by separately counting explicit and implicit ones (Total) as well as combining them (Unique). The Min, Max, and Avg columns denote the minimum, maximum, and average number of (i) discourse senses per connective; (ii) translation equivalents available for each connective in the lexicons, such as an English connective being maximally represented by 9 Turkish equivalents and 2.5 Turkish connectives on average.

4.3.2 Evaluation

An in-depth analysis of the automatically extracted translation pairs reveals that not all pairs serve as seamless translation equivalents, with the conveyed discourse relation sense primarily reliant on the arguments. Variations in construction between the source language (SL) and target language (TL) texts can result in disparities where, despite maintaining the relational sense, the translated discourse connective significantly differs from its SL counterpart. For example, in Example (3), the alignment algorithm associated the implicit connective *clearly* with the implicit Turkish connective *işte* based on their shared sense (Expansion: Level-of-detail: Arg2-as-detail) and argument spans. According to [7], the sense of *Expansion: Level-of-detail: Arg2-as-detail* is employed "when Arg2 elaborates on the situation in Arg1." Hence, sub-sense information derives from the arguments themselves rather than solely from the connectives. Moreover, connectives in the current example are Implicit and were chosen by the annotator not the translator himself/herself.

(3) *In blue, we see the performance of the 500 largest global companies, and in gold, we see a subset of companies with best practice in climate change strategy and risk management* (clearly) **over almost eight years, theyve outperformed by about two thirds**.
[Implicit, Expansion:Levelof-detail:Arg2-as-detail] (English, TED Talk no. 1927)

..dünyadaki en büyük 500 şirketin performansını görüyoruz ve altın rengi olarak iklim değişikliği stratejisi ve risk yönetiminde en iyi uygulamalara sahip şirketlerin alt kümesini görüyoruz (işte) **Yaklaşık sekiz sene boyunca, yaklaşık üçte iki oranda daha fazla performans gösterdiler**.
[Implicit, Expansion:Levelof-detail:Arg2-as-detail] (Turkish, TED Talk no. 1927)

To tackle such scenarios, there is a necessity to explore an automated approach. Similar to the methodology in [91], each translation pair underwent evaluation against comprehensive bilingual dictionaries like Treq [95] and the OPUS word alignment database[3]. Nonetheless, these resources were found ineffective for the current study, resulting in the exclusion of numerous translation options. The challenge arose from the absence of specific candidates in these references or their low likelihood of occurrence,

[3] http://opus.nlpl.eu/lex.php

which prevented the establishment of a sufficient threshold to distinguish unacceptable translations from acceptable ones. Furthermore, the inclusion of suffixal connectives in Turkish further restricted the applicability of these resources for lexicographic assessment. Therefore, as delineated in [82], a manual evaluation was conducted on two bilingual discourse connective lexicons for Turkish-English and European Portuguese-English. This assessment unveiled 9 instances of improper usage in European Portuguese and 3 cases in Turkish (See the example entries in (4)), constituting a relatively small proportion within alignment dataset's translation pairs. Substantially, upon detailed examination, the majority of these pairs involve at least one implicit discourse relation type categorized under Expansion: Level-of-detail: Arg2-as-detail, aligning with a rationale similar to the example presented in Example (3).

(4) **Pt-En:** e - *rather*, e - *for that matter*, enquanto - *and*, assim - *that is*, de facto - *specifically*, e - *as well as*, e - *lastly*, isto é - *clearly*, assim - *specifically*
Tr-En: özetle - *clearly*, yani - *clearly*, işte - *clearly*

In this study, conducting a manual examination of all language pairs was considered impractical. To evaluate the lexicons, an intrinsic approach was utilized rather than relying on an extrinsic evaluation method. F-scores were computed for Explicit and Implicit discourse relations alignments, serving as the basis for the lexicon evaluation. The Precision, Recall, and F-score metrics for English-Turkish, English-European Portuguese, and English-Lithuanian are outlined in Table 11, as gold alignments are available exclusively for these three language pairs. A mean F-score of 0.94 was achieved. Furthermore, the lexicons constructed from automatically generated relation links cover, on average, 97.46% of entries compared to those derived from manually-verified links. This result implies that the F-score outcomes can be reliably extended to alignment data concerning German, Polish, and Russian languages.

Table 10: Statistics regarding the generated lexicons.

Language	Connectives			Senses			Translations		
	Exp	Imp	Total (Unique)	Min	Max	Avg	Min	Max	Avg
English	26	26	52 (44)	1	3	1.25	1	6	1.79
German	29	20	49 (43)	1	3	1.24	1	8	1.90
English	27	32	59 (51)	1	5	1.20	1	9	2.27
Lithuanian	33	35	68 (59)	1	5	1.38	1	4	1.97
English	17	22	39 (33)	1	4	1.18	1	7	2.21
Polish	31	25	56 (51)	1	4	1.25	1	3	1.54
English	28	34	62 (53)	1	3	1.23	1	6	1.84
Portuguese	27	27	54 (44)	1	6	1.46	1	6	2.11
English	22	20	42 (35)	1	3	1.10	1	5	1.76
Russian	31	12	43 (43)	1	3	1.12	1	5	1.72
English	25	33	58 (48)	1	4	1.29	1	9	2.50
Turkish	39	40	79 (67)	1	5	1.43	1	4	1.84

Table 11: The performance of method II on only Implicit and Explicit relations

Language Pair	TP	FN	FP	TN	Accuracy	Precision	Recall	F Score
EN LT	205	2	11	36	0.95	0.95	0.99	0.97
LT EN	205	3	9	60	0.96	0.96	0.99	0.97
EN PT	197	0	26	27	0.9	0.88	1	0.94
PT EN	197	1	21	24	0.91	0.9	0.99	0.95
EN TR	188	6	30	20	0.85	0.86	0.97	0.91
TR EN	188	6	28	40	0.87	0.87	0.97	0.92

4.4 Limitations and Conclusion

In this section, a technique for constructing bilingual discourse connective lexicons using aligned DR annotations was explained. By employing a fully automated approach, several high-quality bilingual discourse connective lexicons were produced. These lexicons were published online as HTML web pages[4]. Despite the inclusion of implicit discourse relations in the lexicons, the resources generated from TED-MDB are limited due to its small size and require improvement. Moreover, there is uncertainty regarding whether the selected connectives truly represent their respective contexts since implicit discourse relations are chosen from the annotator tool's configuration file. To address this issue, enhancing the annotator's inventory of discourse connectives is recommended. Unlike prior studies on bilingual lexicons, monolingual connective lexicons or dictionaries were not used in the induction and evaluation of connective lexicons, as not all languages in the corpus have such resources. Nevertheless, these lexicons can be verified, rectified, and converted into gold standards by researchers proficient in the respective languages.

Translating a discourse connective accurately to convey its sense naturally in the target language text is a challenging task, as discussed in Section 4.1. Unlike standard lexical entries, understanding discourse connectives demands knowledge of both the target language's culture and the semantic nuances of the specific contexts. Therefore, even bilingual dictionaries may not always suffice for translators. Hence, bilingual discourse connective lexicons developed in this study based on authentic data hold significance. Despite their limited volume, these prepared lexicons are valuable additions to cross-lingual and machine translation studies and serve as beneficial resources for both human translators and second language learners alike.

[4] http://metu-db.info/mdb/ted/resources.jsf

CHAPTER 5

ANNOTATION IMPROVEMENT

5.1 Introduction

Aligning Discourse Relations (DRs) and refining existing annotations is an alternative to the translation spotting approach [96]. In annotation schemes with fine-grained, detailed sense levels, as exemplified by the English Penn Discourse Treebank (PDTB) [7], there is a risk of lower annotator agreement if each monolingual text is annotated independently of the translated texts. To overcome this problem, the translation spotting technique has been proposed. This technique involves annotating the sense of discourse relations in the source text by comparing it with its translation. It speeds up the annotation process and provides a means to spot translational differences at the outset. However, this approach results in annotations specific to source-target language pairs. If the connective in the target text is ambiguous or has multiple senses, it does not aid in identifying the sense of the discourse relation in the source text. Furthermore, any discourse relations added in the target text due to translational variations are discarded, which can provide an incomplete picture of the cross-lingual realization of discourse relations.

Our methodology, which involves first annotating independently, then aligning and checking the annotations, removes the language pair specificity problem and does not solely rely on original text annotations. This improves the quality of annotations in both the source and target texts. However, it should be noted that the shortcoming of our approach is that it slows down the annotation process.

In the annotation of the TED Multilingual Discourse Bank (TED-MDB), each monolingual team annotated texts in their native language without considering the annotations in the original language [6]. Although this approach ensured language independence, the resulting target language (TL) discourse relation (DR) annotations might contain missing or nonexistent tokens compared to the original source language (SL) annotations. This discrepancy arises both from cross-lingual and translational differences between SL and TL, as well as from the annotators' annotation strategies. Therefore, TED-MDB annotations provide a perfect case for refinement through alignment.

Due to the availability of semi-gold DR alignment data for English and its translations into Lithuanian, European Portuguese, and Turkish, enhancement efforts have been focused on these specific datasets. Also,the inter/intra sentential property, connective POS tags and subtitle location have been introduced as part of these enhancements for further analysis and examination.

5.2 Addition to Annotations: Inter/Intra Sentential Property

The investigation into the realization of discourse relations encompasses not only semantic means but also the representation of meaning within or across sentence boundaries. A new feature, referred to as the inter/intra sentential property, has been added to the dataset. Discourse relations that connect independent sentence pairs are classified as inter-sentential (as demonstrated in example sentence (1)), while those that connect segments within a single sentence boundary are categorized as intra-sentential (as shown in example sentence (2)).

(1) *Its not causation*. But **it does illustrate that environmental leadership is compatible with good returns...**
[Explicit, Comparison:Concession:Arg2-as-denier] (English, TED Talk no. 1927)

(2) ...if **the returns are the same or better and the planet benefits** *wouldnt this be the norm...*
[Explicit, Contingency:Condition+SpeechAct] (English, TED Talk no. 1927)

Building on the work of [55], it is important to note that information density differs between inter-sentential and intra-sentential discourse relations. This difference is significant as it reveals how information conveyed by the discourse relation is transferred, either across two separate sentences or within the same sentence boundary during translation. Analyzing the change in inter/intra-sentential properties in translation allows for a comparison of languages in terms of information packaging. Manually classifying these properties would be error-prone, so an automatic classification based on several predefined rules is performed:

- Entity relations, No relations and Hypophora Level-I sense are always across independent sentence boundaries: inter-sentential
- Existence of colon or semi colon after arg1: intra-sentential
- Existence of period or question mark after arg1 or an upper-cased connective: inter-sentential
- Existence of comma before discourse connective or after arg1: intra-sentential
- No punctuation after arg1: intra-sentential

Among a total of 3,081 discourse relations, only 28 were classified incorrectly. However, overall performance is quite high (see Table 12). An inspection of the errors reveals that they mainly originate from exceptional punctuation or annotations. For example, in the discourse relation annotation in (3), the discourse connective is intra-sentential; however, the incorrect placement of a period causes it to be classified as inter-sentential.

(3) *Burada kim Birleşik Devletlerde birçok şehirde yol kenarına battaniyeye sarılı oturmanın yaşadışı oldupunu biliyor*. veya **veya arabanızda uyumanın, veya yabancılara yiyecek önermenin...**
[Explicit, Expansion:Disjunction] (Turkish, TED Talk no. 2009)

Table 12: Rule-based Inter-Intra Labeling Performance

Language	Total_Gold	Total_Predicted	F1-Score
en	708	705	1.00
lt	874	871	1.00
pt	712	700	0.98
tr	787	777	0.99

5.3 Structural vs Anaphoric Connectives: Arg1 Span, Location of the Connective and POS Tagging

[97] distinguishes between structural (subordinating and coordinating conjunctions) and anaphoric discourse connectives. Structural connectives link adjacent sentences as arguments [98] and typically have fixed positions, although there are rare instances in Turkish where their positions may vary [99]. In contrast, discourse adverbials are presuppositional, linking the internal argument (arg2) anaphorically to its external argument (Arg1), similar to how an anaphoric expression is resolved with its antecedent as proposed by [100]. Arg1 in discourse adverbials may not be adjacent to arg2 and can span more than one sentence. Unlike structural connectives, discourse adverbials have flexible positioning within arg2. To investigate the differences between structural and anaphoric connectives, additional features are required. It's important to note that in TED-MDB, currently, only adjacent sentences are annotated for discourse relations [6]. Therefore, firstly, the Arg1 span must be identified for long-distance resolution, whether it comprises one sentence or multiple sentences. For this purpose, the Punkt tokenizer, developed by [68] and integrated into the Natural Language Toolkit (NLTK) library in Python, was used. Secondly, the location of the connective with respect to Arg1—whether it is initial, within, or at the end—was specified. Finally, part-of-speech (POS) tag information was obtained for the discourse connectives.

For POS tagging, three different libraries were utilized: the UD (Universal Dependencies) Pipeline [101], Spacy [102], and the Stanza POS tagger[103]. These three libraries were employed to enhance accuracy. The results from Stanza were primarily used in the analysis. However, in cases where Stanza did not provide an output, the outputs from the other two libraries were used. The discourse relation segment (Arg1 + connective + Arg2) was provided as input to all three pipelines.

UD-Pipe [101] is a versatile tool capable of performing tokenization, POS tagging, lemmatization, and dependency parsing. It is built upon the Universal Dependencies (UD) project, which is a cross-linguistically consistent treebank providing annotations of morphological and syntactic structures across various languages. It supports a wide range of languages. The output from UD-Pipe is produced in a revised version of the CoNLL-X format, known as CoNLL-U (Conference on Computational Natural Language Learning - Universal Dependencies), which standardizes the annotation of linguistic data across different languages. Example output is provided in Figure 24

SpaCy [102] is a Python NLP library designed for a variety of tasks including POS tagging, named entity recognition, dependency parsing, and more. It offers pre-trained models for multiple languages, facilitating quick and easy implementation of complex NLP tasks. In this context, SpaCy is employed

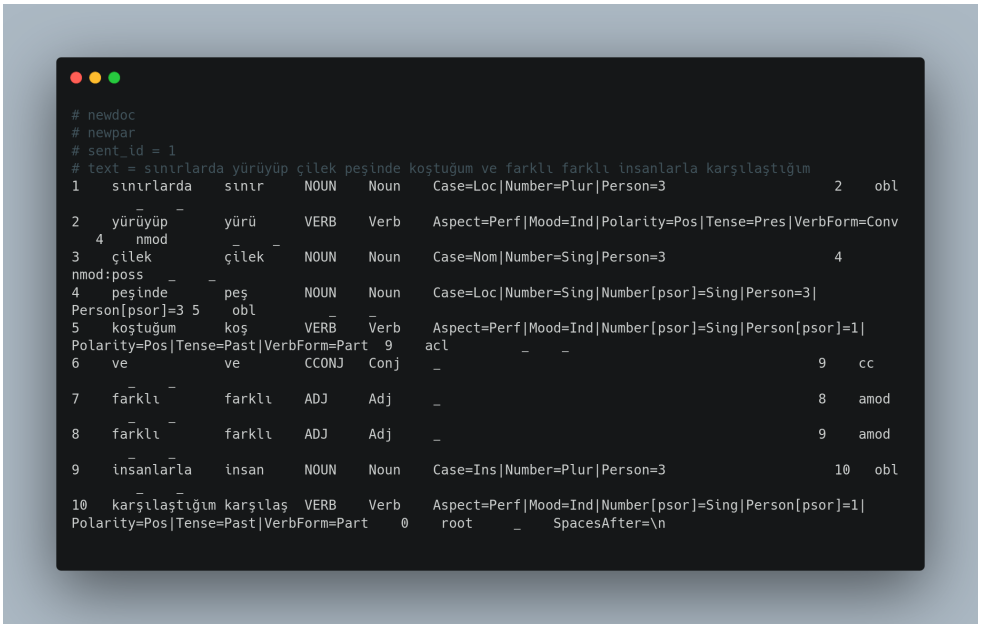


Figure 24: A screenshot showing the output of UD-Pipe for a DR from Turkish, TED Talk no. 2009

for POS tagging only in Turkish data. While SpaCy itself does not currently support POS tagging for Turkish out of the box, there are other Turkish models compatible with SpaCy [104].[1]

Stanza[103] is an NLP library developed by the Stanford NLP Group. This library provides a wide range of language processing tools, such as tokenization, POS tagging, lemmatization, dependency parsing and Named Entity Recognition. Stanza is a PyTorch NLP library which not only provides state of the art pre-trained neural models for 66 (as of now) languages but also frameworks to easily construct high-performance analysis pipelines. Spacy and Stanza each use their own object oriented structure to represent their outputs.

5.4 Subtitling Information

The literature proposes ten subtitling strategies: expansion, paraphrase, transfer, imitation, transcription, dislocation, condensation, decimation, deletion, and resignation [105]. Among these, condensation, decimation, and deletion are identified as reduction strategies, also referred to as omission-implicitation. Conversely, expansion, paraphrase, transfer, imitation, transcription, and dislocation are categorized as addition-explicitation strategies. Subtitling generally involves reduction. According to [106], several factors contribute to reductions in subtitling. First, reductions may occur due to changes in medium, channel, and code, such as transitioning from a spoken to a written register, leading to the omission of spoken features in the source text. Second, reductions may arise from selection criteria inherent to subtitling, like the necessity for text compression because of time and space constraints. Subtitles are typically limited to two lines, necessitating text reduction based on available time, audience reading speed, and the speed of the source text. Third, reductions may result from translators

[1] https://huggingface.co/turkish-nlp-suite/tr_core_news_lg

working solely with scripts and not viewing the actual video content.[2] Although subtitling strategies can manifest at various linguistic levels—such as the simplification and shortening of phrases, enumeration generalization, the use of shorter near-synonyms or equivalent expressions, and the preference for simple tenses over compound tenses—the current thesis focuses specifically on the explicitation and implicitation of discourse connectives.

For each TED Talk, subtitle information can be obtained in JSON file format from the following URL: https://www.ted.com/talks/subtitles/id/X/lang/Y, where "X" represents the TED Talk ID and "Y" indicates the language abbreviation. For instance, the Turkish subtitle for TED Talk 1927 is found at this URL: https://www.ted.com/talks/subtitles/id/1927/lang/tr. In the JSON file, each subtitle line includes features such as id, duration, content, startOfParagraph, and startTime. When extracting discourse relations (DRs) for each language, subtitle information was also recorded, specifically noting whether the DR annotation appears at the beginning of a subtitle line, at the start of a paragraph or within the subtitle. This information was used to investigate the distribution of DRs in relation to the subtitle content.

5.5 Extending the Annotations for 3 Languages Using Missing Alignments

Using the first version of discourse relation alignments for English-Turkish, English-European Portuguese, and English-Lithuanian language pairs, English discourse relations that are not aligned with their target languages were identified in the annotation files. Since the PDTB annotation tool [61] is not open source, a solution was implemented to highlight missing tokens by marking them as REJECTED in the English annotation files. These REJECTED annotations are treated separately by the tool and displayed as REJECTED in the annotations panel.

Subsequently, using the PDTB annotation tool, cases that were either missed or annotated incorrectly were corrected through new annotations or updates to the existing ones. Additionally, several duplicate annotations were detected and rejected in the files. These changes necessitated the reproduction of the alignments. The modifications in the annotation files led to changes in the distribution of discourse relation types and senses. However, these changes are provided only descriptively here, with a more detailed interpretation to be provided in the final chapter (6) for the new dataset.

5.5.1 Total Count of Discourse Relations

In the first version of the data, an average of 611 discourse relations (DRs) in English were aligned with target languages. In the tuned data, this number increased to 660.3, with the DR alignment percentage rising in each language (see Table 13). This is significant because the total number of DRs in English actually decreased due to the removal of duplicate annotations in the new dataset.

[2] https://www.ted.com/participate/translate/guidelines

Table 13: Comparison of Old and New Data

	Old Data				New Data			
	EN	LT	PT	TR	EN	LT	PT	TR
Total DRs	716	821	680	760	708	874	712	787
Linked DRs	-	627 (76%)	597 (87.8%)	608 (80%)	-	686 (78.5%)	641 (90%)	654 (83.1%)

5.5.2 DR Type Distribution Across Four Languages

In both datasets, except for European Portuguese, there is a trend toward explicitness for all three languages, which is found to be statistically significant using the Chi-square Pearson statistics test (for the first dataset: χ^2=10.67, p=0.01<0.05, and for the second dataset: χ^2=11.45, p=0.01<0.05). When this issue is investigated for each language with the Chi-Square goodness-of-fit test, the difference between Explicit and Implicit DRs is statistically significant in the old dataset for Lithuanian (χ^2=5.56, p=0.02<0.05) and Turkish (χ^2=4.49, p=0.03<0.05). However, in the second dataset, although the same data pattern continues to exist for Lithuanian (χ^2=10.72, p=0.001<0.05), it loses statistical significance for Turkish (see Figure 25). Lighter colors indicate the old dataset, whereas darker colors belong to the new datasets. Additionally, Tables 14 and 15 show the distribution of DR Types in the old and new data for all four languages.

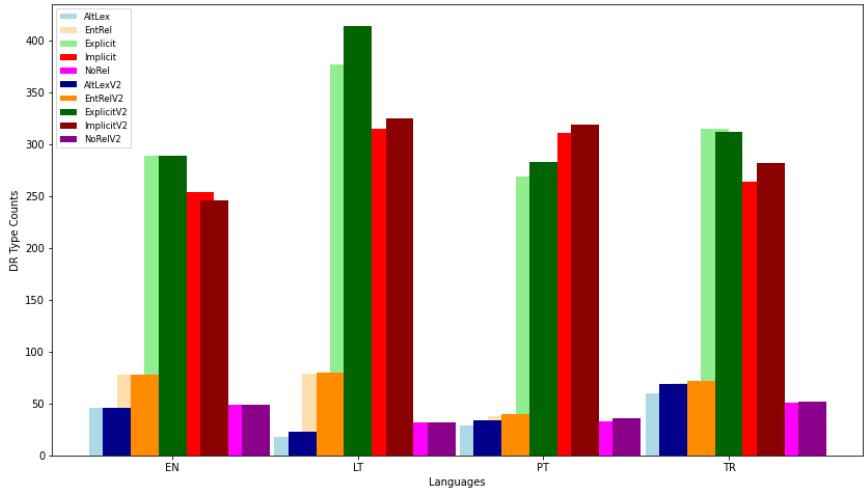


Figure 25: DR Type Distribution in the Old and New DR Alignment Data Sets

In this chapter, I presented how the DR alignment methodology can be used to improve existing annotations on multilingual parallel corpora as an alternative to the translation spotting approach. I also described how we included other DR properties: inter/intra sentential, POS tags and subtitle location. Following that, old and new alignment data sets were compared briefly for English-Lithuanian, English-Porteguese and English-Turkish.

Table 14: Distribution of DR Types Across 4 Languages in the Old Data Set

DRType	EN	LT	PT	TR	Total
AltLex	46.0 (6.42%)	18.0 (2.19%)	29.0 (4.26%)	60.0 (7.89%)	153
EntRel	78.0 (10.89%)	79.0 (9.62%)	38.0 (5.59%)	70.0 (9.21%)	265
Explicit	289.0 (40.36%)	377.0 (45.92%)	269.0 (39.56%)	315.0 (41.45%)	1250
Implicit	254.0 (35.47%)	315.0 (38.37%)	311.0 (45.74%)	264.0 (34.74%)	1144
NoRel	49.0 (6.84%)	32.0 (3.9%)	33.0 (4.85%)	51.0 (6.71%)	165
Total	716	821	680	760	2977

Table 15: Distribution of DR Types Across 4 Languages in the New Data Set

DRType	EN	LT	PT	TR	Total
AltLex	46.0 (6.5%)	23.0 (2.63%)	34.0 (4.78%)	69.0 (8.77%)	172
EntRel	78.0 (11.02%)	80.0 (9.15%)	40.0 (5.62%)	72.0 (9.15%)	270
Explicit	289.0 (40.82%)	414.0 (47.37%)	283.0 (39.75%)	312.0 (39.64%)	1298
Implicit	246.0 (34.75%)	325.0 (37.19%)	319.0 (44.8%)	282.0 (35.83%)	1172
NoRel	49.0 (6.92%)	32.0 (3.66%)	36.0 (5.06%)	52.0 (6.61%)	169
Total	708	874	712	787	3081

CHAPTER 6

DISCUSSION

Parallel corpora have played a crucial role in advancing research in cross-linguistic studies and translation. However, the lack of parallel corpora annotated for discourse relations in both languages has limited previous cross-lingual investigations to specific aspects, such as the implicitation of discourse markers. Typically, these studies relied on parallel data with manual annotations on only one side. By aligning discourse relations, it has become possible to conduct a more comprehensive multilingual analysis.

This chapter centers on a descriptive analysis of cross-linguistic discourse structures and the expression of discourse relations in TED-MDB languages. It explores the nuances of discourse realization in various languages by examining the aligned discourse relations. By addressing key questions regarding differences in expression, semantic shifts, inter-sentential encoding patterns, and the effect of genre and contextual effects in connective translation (relative and translation entropy), the chapter sheds light on the variances in discourse structure between English and other target languages within the TED-MDB corpus. The analysis underscores how aligning discourse relations in different languages enhances multilingual understanding and facilitates in-depth cross-linguistic comparisons.

By analyzing relation types, discourse senses, inter- and intra-sentential encoding patterns, and subtitle location encoding patterns, the chapter provides insights into how languages diverge in aspects like implicitation, explicitation, inter- and intra-realization, and the prepared speech genre. The findings highlight patterns of discourse coherence, implicitation-explicitation frequency, and the distribution of discourse senses across languages, offering critical insights into the complexities of cross-linguistic discourse analysis and translation studies within the TED-MDB corpus. A more theoretical discussion of the previous version of the data for EN-LT, EN-PT, and EN-TR language pairs is available in [107].

6.1 Realization of Discourse Relations in TED-MDB languages

Rest of the chapter offers a broad overview of the variances in discourse structure between English and other target languages. Subsequent chapters adopt a systematic approach to explore variations in discourse relations by focusing on five key questions:

- How do discourse relations differ in their expression across various languages (e.g., explicit or implicit)?
- How do the semantics of relations between identical text segments change across languages?

- Are the translations of discourse relations affected by sentence-encoding patterns, specifically, whether information is dispersed across sentences or concentrated within a single sentence (i.e., inter-sententially or intra-sententially)?
- Does implicitation/explicitation differ with respect to the discourse relations' location in the subtitle line?
- Are there cross-lingual commonalities or divergences in the translation of frequent English connectives? The final question directs the analysis towards examining individual connectives, taking into account other connectives that convey the same sense in both the source and target languages..

Linked relations will be examined to explore these questions. The alignment data for EN-LT, EN-PT, and EN-TR language pairs presented in this chapter are the result of the annotation refinement process described in Chapter 5, manually checked and not published elsewhere. For the remaining language pairs, as no new annotation was added, annotations are based on the automatic alignments released previously. In order to ensure the credibility of the analysis, references to gold alignments will be made, if available, for Lithuanian, Turkish, and Portuguese; otherwise, automatically aligned data will be used as stated. Due to potential inaccuracies in automatically aligned links, it is crucial to approach the analysis with caution. Nonetheless, the high F-scores obtained in capturing the semi-automatically formed links (see Table 8) indicate that the findings reported closely reflect the distribution within the semi-automatically linked data.

In the following sections, similarity between contingency matrices will be measured with the Hellinger distance, which indicates the similarity between two probability distributions. The contingency matrices (or tables) used in this chapter represent the frequency distribution of aligned discourse relation types, senses, etc. In order to use the Hellinger distance, frequency counts were converted into probability distributions, as described in the following Section 6.1.1.

6.1.1 Hellinger Distance Calculation

To calculate the Hellinger distance between two contingency matrices A and B, following steps were followed:

1. Contingency matrices were normalized as shown in Equations 4 and 5: A and B were converted into probability matrices P and Q respectively:

$$p_{ij} = \frac{a_{ij}}{\sum_k \sum_l a_{kl}} \tag{4}$$

$$q_{ij} = \frac{b_{ij}}{\sum_k \sum_l b_{kl}} \tag{5}$$

2. Hellinger Distance was calculated: The Hellinger distance between matrices P and Q is defined as shown in Equation 6:

$$H(P,Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_i \sum_j \left(\sqrt{p_{ij}} - \sqrt{q_{ij}}\right)^2} \tag{6}$$

Hellinger distance was chosen because it is symmetrical (the distance between distribution P and distribution Q is the same as the distance between Q and P) and bounded (between 0 and 1). Distance values closer to 0 indicate perfect similarity, whereas values closer to 1 indicate a more significant difference.

6.1.2 Variations in Relation Types Across Languages:

To address the initial question, a comparative examination of relation types from each connected relation is undertaken. The heat-map visualizations in Figure 26 a-e illustrate confusion matrices for relations across all language pairs. In these matrices, the rows represent English relations, while the columns correspond to target languages. Row-wise normalization is implemented, with each cell indicating the percentage of English relations transformed into the corresponding discourse relation type label in the target language. It is important to note that bold highlights signify confusion matrices derived from manually corrected links. The row entries depict English relations and demonstrate how frequently they are expressed as specific labels on the X-axis. For example, as illustrated by the statement "15% of English explicit discourse relations are realized implicitly in German" in Figure 26-a, cells are color-coded to show agreement levels. Lighter colors indicate lower levels of agreement, transitioning into redder tones as agreement increases (a detailed color legend is provided within each figure). An ideal match scenario would exhibit red diagonal cells only, with off-diagonal cells appearing as white or gray. Difference between each contingency table pairs are measured via Hellinger Distance. Differences in distance measures (<=0.33) suggest that target languages display largely similar patterns in discourse relation type conversion. Russian and German show the closest similarity in DR type translation, while Turkish and Polish exhibit the greatest disparity. The observation concerning these discourse relation (DR) type conversions among the languages is quite interesting. The resemblance and diversity in DR type translations can shed light on the similarities or differences in the linguistic structures and discourse patterns of these languages.

The similarity in DR type conversion patterns between Russian and German indicates that they may share certain structural or syntactic characteristics that enable similar expressions of discourse relations. Despite belonging to different branches of the Indo-European language family (Germanic for German and Slavic for Russian), they might still exhibit similar ways of structuring discourse.

On the other hand, distance differences in DR type conversion between Turkish and Polish,compared to other language pairs are likely due to their distinct language families and structures. Turkish, being a Turkic language, features unique syntactic and morphological features different from Polish, a Slavic language like Russian. (see Table 16).
Inspection of the heatmap data indicates that 72-79% of explicit relations and 0.67-0.79 of implicit relations are retained by target languages. Considering all discourse relation type conversions, from English to 6 target langauges, three main type conversions are observed. On average 31% of AltLex relations become Explicit; 29% of EntRels become Implicit and finally 16% of Explicit relations become Implicit (also named as implicitation).

With respect to the first type shift, since alternative lexicalizations are types that are bound to the linguistic properties of each language, it is common to observe type shifts in the AltLex group. AltLexes in English are represented either as Explicits, Implicits, or AltLexes in the target languages.

Table 16: Hellinger Distance Between DR-Type Contingency Tables

Language Pair-I	Language Pair-II	Hellinger Distance
en_ru	en_de	0.14
en_pt	en_ru	0.17
en_pt	en_de	0.18
en_pl	en_ru	0.20
en_pl	en_de	0.20
en_ru	en_lt	0.20
en_pt	en_tr	0.22
en_de	en_lt	0.23
en_pt	en_lt	0.24
en_tr	en_lt	0.24
en_pl	en_lt	0.24
en_tr	en_de	0.27
en_tr	en_ru	0.28
en_pt	en_pl	0.29
en_tr	en_pl	0.33

The second most observed type shift is from EntRel to Implicit. Differentiating between these two discourse relation types has been reported to be difficult even within the same language [108]. Moreover, in implicit discourse relation recognition tasks, EntRels are used as Implicit Expansion relations in the literature, with the aim of increasing the training data [109, 110]. The interchangeable nature of EntRels and Implicit Expansion relations is also observed in the aligned discourse relations we produced. On average, 78.68% of English EntRels that become Implicit in the target language are annotated with the Expansion Level-I sense.

The phenomenon of implicitation, which involves omitting a connective and expressing it via an implicit discourse relation in the target text during translation from the source text, emerges as the third most common shift in relation types. This trend persists even when excluding conversions from EntRel to Implicit, which are viewed as somewhat interchangeable, as stated before. Notably, Portuguese has a tendency for implicitness, as indicated by the Implicit column in Figure 26-d. In the analysis of all language pairs in TED-MDB, it is noted that at least 13% of English relations are expressed implicitly.

Transitions from AltLex to Explicit or from Explicit to Implicit may primarily stem from language-specific factors (see Example (2)), translational influences (see Example (1)), or differences in annotation strategies. In contrast, the conversion from EntRel to Implicit is driven by annotation methodology, indicating that it is influenced by the annotator's decisions (see Example (3)). For instance, in Example (1), the explicit discourse connective *so* is rendered implicit in the translation.

(1) *Thats the equivalent of taking 21,000 cars off the road.* So **awesome, right**.
[Explicit,Contingency:Cause+SpeechAct:Result+SpeechAct] (English, TED Talk no. 1927)

Bu, 21.000 aracı trafikten çıkarmaya eşdeğer. (İşte) **Çok harika, değil mi**.
[Implicit, Contingency:Cause+SpeechAct:Result+SpeechAct] (Turkish, TED Talk no. 1927)

In Example (2), alternative lexicalization *the same way* is expressed via a discontinuous Explicit connective in Turkish *tıpkı..gibi* due to syntactic differences between English and Turkish discourse relation segments.

(2) *the light waves of the light and waves diffracts around that screen* the same way **it did in the telescope**.
[AltLex, Comparison:Similarity] (English, TED Talk no. 1976)

Bunun nedeni, ışık dalgalarının tıpkı teleskobun içinde olduğu gibi *gölgelikten sapma yapmaları*.
[Explicit, Comparison:Similarity] (Turkish, TED Talk no. 1976)

In Example (3), The first annotator decided that discourse relation was provided by the use of the same entity *prosthetic socket*, so annotated the relation as EntRel. However, for Turkish, the annotator preferred to link the two sentences via an Implicit *ve*.

(3) *The prosthetic socket is the part in which the amputee inserts their residual limb, and which connects to the prosthetic ankle*.**Even in the developed world, it takes a period of three weeks to often years for a patient to get a comfortable socket, if ever**.
[EntRel] (English, TED Talk no. 1971)

Protez soketi, uzvu kesilmiş kişinin kesik uzvuna taktığı ve böylece uzvu protez ayağa bağladığı parçadır. (ve) **Gelişmiş ülkelerde bile bir hastanın rahat bir soket alabilmesi 3 haftadan yıllara kadar çıkabiliyor, tabi alabilirse**.
[Implicit, Expansion :Conjunction] (Turkish, TED Talk no. 1971)

Implicitation of explicit discourse relations raises the question of which Level-I sense is more prone to implicitation. Notably, implicitation is most commonly observed in Expansion discourse relations (see the Table 17), where AltLexes are also taken into account. In this table, percentage of implicitated values are shown in parenthesis. Inline with the literature [111] and [76], discourse relations represented by more frequent senses (here Expansion)(see Figure 28), has more implicitation counts. In the table, Turkish seems to have the highest number in Temporal sense. However, a closer inspection of the alignment data shows that two out of five cases in Temporal sense are discourse relations aligned to Turkish discourse relations with a type change of Expansion as in Example (4). Higher percentages in Hypophora sense are closely related to the fact that they are represented by alternative lexicalizations which are quite prone to implicitation.

(4) *Youll see the petals unfurl* . Now **youre seeing it deploy**.
[AltLex, Temporal:Asynchronous] (English, TED Talk no. 1976)
Yapraklarının açıldığını göreceksiniz. (Ve) **Şimdi yerleştirilmesini görüyorsunuz**.
[Implicit, Expansion:Conjunction] (Turkish, TED Talk no. 1976)

Table 17: The sense distribution of the English relations that are implicitated in the target language.

Language	Comparison	Contingency	Expansion	Hypophora	Temporal
German	2 (3.92%)	10 (12.99%)	23 (13.53%)	1 (10.00%)	4 (14.81%)
Lithuanian	4 (7.84%)	10 (12.99%)	43 (25.29%)	0 (0.00%)	6 (22.22%)
Polish	5 (9.80%)	3 (3.90%)	37 (21.76%)	3 (30.00%)	3 (11.11%)
Portuguese	3 (5.88%)	11 (14.29%)	36 (21.18%)	3 (30.00%)	4 (14.81%)
Russian	4 (7.84%)	12 (15.58%)	21 (12.35%)	1 (10.00%)	2 (7.41%)
Turkish	3 (5.88%)	4 (5.19%)	27 (15.88%)	0 (0.00%)	5 (18.52%)

In contrast, explicitation(see the Table 18) tends to occur more frequently within Contingency discourse relations on average, although they do not demonstrate the same level of frequency as implicitly realized Expansion discourse relations.

Table 18: The sense distribution of the English relations that are explicitated in the target language.

Language	Comparison	Contingency	Expansion	Hypophora	Temporal
German	1 (4.17%)	20 (24.69%)	14 (8.86%)	2 (20.00%)	0 (0.00%)
Lithuanian	6 (25.00%)	30 (37.04%)	20 (12.66%)	2 (20.00%)	3 (15.79%)
Polish	6 (25.00%)	23 (28.40%)	10 (6.33%)	0 (0.00%)	0 (0.00%)
Portuguese	1 (4.17%)	19 (23.46%)	9 (5.70%)	0 (0.00%)	2 (10.53%)
Russian	1 (4.17%)	21 (25.93%)	6 (3.80%)	4 (40.00%)	2 (10.53%)
Turkish	3 (12.50%)	24 (29.63%)	14 (8.86%)	0 (0.00%)	1 (5.26%)

6.1.3 Variation in Level-I Discourse Senses Across Languages:

Figure 27 a-e illustrate the Level-I sense shift from English to other languages. The rows represent English relations, while the columns represent the target languages. The matrices are normalized row-wise, with each cell indicating the percentage of English relations transformed into the corresponding label in the target language. It is important to note that bold highlights indicate confusion matrices derived from manually corrected links. Difference between each contingency table pairs are measured via Hellinger Distance. Distance differences (<=0.19) indicate that the target languages exhibit similar patterns in Level-I sense transitions of discourse relations. This means that within this range, the methods these languages use to express primary discourse relations—like cause-effect, temporal, and contrast connections—are closely aligned. (see Table 19.

In contrast to relation types, as the Hellinger Distance results in Table 19 indicate, Level-I senses of the discourse relations remain relatively consistent across different languages. On average, 85% of English discourse relations maintain their top-level senses when translated into target languages. Line graph in Figure 28 also shows the tendency to preserve higher level sense types. This analysis excludes EntRel or NoRel types as they lack sense assignments.

Manual inspection of the data shows that the majority of the sense shifts are due to translation changes. In Example (5), translation of the sentence in English into Turkish results with meaning change, which reflects itself on the Level-I sense shift from Temporal to Expansion.

Table 19: Hellinger Distance Between Level-I Sense Contingency Tables

Language Pair-I	Language Pair-II	Hellinger Distance
en_pt	en_lt	0.05
en_pt	en_tr	0.09
en_pt	en_de	0.10
en_de	en_lt	0.10
en_tr	en_lt	0.11
en_pl	en_lt	0.13
en_pt	en_pl	0.14
en_tr	en_de	0.14
en_pl	en_ru	0.15
en_pl	en_de	0.15
en_ru	en_de	0.15
en_tr	en_pl	0.16
en_pt	en_ru	0.17
en_tr	en_ru	0.17
en_ru	en_lt	0.19

(5) *It was taken by the Voyager spacecraft in 1990* when **they turned it around as it was exiting the solar system to take a picture of the Earth from six billion kilometers away**.
[Explicit,Temporal:Synchronous] (English, TED Talk no. 1976)

Voyager uzay aracı tarafından 1990da çekildi. (şöyle ki) **Araç güneş sisteminden çıkmak üzereyken, onu bu tarafa çevirip, 6 milyar kilometre öteden Dünyanın fotoğrafını çektirdiler**.
[Implicit, Implicit,Expansion:Level-of-detail:Arg2-as-detail] (Turkish, TED Talk no. 1976)

6.1.4 Inter-Intra Sentential:

Figure 29 a-e show the distribution of discourse relation types grouped by inter-intra sentential property. It should be reminded that as described in Chapter 5, annotations for Turkish, European Portuguese and Lithuanian were refined in accordance with the annotations in English which also resulted in new intra-sentential Implicit discourse relation annotations. In the bar graphs, percentage of the DR in that language is represented over each bar, whereas inter-intra percentage in that type is written over the respective bar stack. For example, in Figure 29-a, 6.5% of all DRs are AltLex type. Among these, 34.78% are realized as intra-sentential whereas 65.22% being inter-sentential. Majority of Explicit DRs are realized as intra-sententially.AltLex and Implicit DRs are mostly realized as inter-sentential. In all languages studied, both EntRel and NoRel discourse relations are consistently classified as inter-sentential. For Polish and German, the AltLex relation types are also entirely inter-sentential.

Translation of intra-sentential and inter-sentential DR types, as shown in the contingency tables in Figure 30, inter-sentential DR types are mostly translated as inter-sentential DR types (except the heat map for English-Lithuanian). However, 9-30% of intra-sentential DRs in English become inter-sentential

DRs in the dataset. Regarding the inter-intra exchange, translators often prefer to convey information through separate sentences during the translation process. This preference is understandable, considering the time and space constraints characteristic of the subtitling genre. Difference between each contingency table pairs are measured via Hellinger Distance. Distance differences (<=0.14) indicate that the target languages exhibit similar patterns in the inter-intra expression of discourse relations. This suggests that within this threshold, how these languages manage both internal (within sentences) and external (between sentences) discourse relations are closely aligned. Russian and German show perfect similarity (see Table 20).

Table 20: Hellinger Distance Between Inter-Intra Contingency Tables

Language Pair-I	Language Pair-II	Hellinger Distance
en_ru	en_de	0.00
en_pl	en_ru	0.02
en_pl	en_de	0.02
en_pt	en_tr	0.04
en_pt	en_lt	0.05
en_tr	en_lt	0.09
en_pt	en_pl	0.10
en_tr	en_pl	0.10
en_pt	en_ru	0.12
en_pt	en_de	0.12
en_tr	en_ru	0.12
en_tr	en_de	0.12
en_pl	en_lt	0.12
en_ru	en_lt	0.14
en_de	en_lt	0.14

Figure 31 depicts the Implicitation-Explicitation of discourse relations in English, grouped by inter-intra sentential property. For explicitation, all target languages explicitate intra-sentential discourse relations more than inter-sentential discourse relations in English. For implicitation, the resulting pattern is mixed.In German, inter and intra sentential relations are treated equally for implicitation. In Russian, inter-sentential relations are implicitated more, for the other languages, a reverse pattern is observed.

6.1.5 Subtitling Structure:

Figure 32 a-e illustrate the distribution of discourse relation types across three subtitle positions for each language investigated. In this context, "Both" refers to the start of both a paragraph and a subtitle, "Subtitle" signifies the start of a subtitle line, and "Within Subtitle" denotes a location within the subtitle line. At the top of each bar, percentage of relation type is written in highlighted form. In each bar, distributions of subtitle locations are shown and percentage values are written on each bar stack. For instance, as illustrated in Figure 32-a, 6.5% of the discourse relations in English are classified as AltLex type. Of these, 2.17% are found at the beginning of paragraphs, 52.17% at the start of sentences, and 45.65% are positioned within the subtitle lines. Upon examining the Figure 32 a-e,

it becomes apparent that, in nearly all languages, Explicit and Implicit discourse relations are rarely used at the beginning of a new paragraph. Conversely, NoRel discourse relations are most commonly positioned at the beginning of a new paragraph, either to introduce a new discussion or to initiate a new topic. Consistent with these observations, Figure 33 depicts the implicitation and explicitation of discourse relations across six languages and the subtitle location of the DR in English. Across almost all languages investigated, a similar pattern emerges. Discourse relations that appear at the beginning of subtitles are more susceptible to implicitation, and to a lesser extent, explicitation. However, it's important to note that these results are reported descriptively.

6.1.6 Discourse Connectives:

In this section, the analysis will focus on the individual properties of explicit discourse connectives. Section 6.1.6.1 investigates the data related to connectives that have anaphoric references. Section 6.1.6.2 presents an analysis of the translation candidates for the most frequently used English connectives in TED-MDB. This chapter also explores the translation alignment entropy for these connectives. In this section, the specificity values for individual connectives are also quantified and analyzed. When a more specific connective is translated into a less specific or low-information-value connective, this process is referred to as *underspecification*. Conversely, the reverse condition is termed *specification* (c.f. [112] and [113]).

6.1.6.1 Anaphoric Connectives:

Out of a total of 4,782 discourse relations, only 33 met the criteria of being inter-sentential with an Arg1 span covering multiple sentences. The position of the connective is nearly always fixed in the data, with the exception of a single instance in European Portuguese involving *assim*, which can be translated as 'so'. A list of connectives is provided in the Appendix A. In addition to adverbials, the list also includes CCONJ such as 've' and 'fakat'. However, as noted by [99], discourse relations(see examples in he Appendix A for Turkish) are still realized structurally, with coherence depending on the adjacency of sentences. [114] investigated long-distance anaphoric relations in the Prague Dependency Treebank 3.0. They discovered that half of the connectives taking non-adjacent arguments are coordinating conjunctions, attributing this to the diachronic development of some adverbs or their 'syntactic stretching'. However, unlike their study, due to the limited data available, the investigation of the differential realization of structural versus anaphoric connectives in TED-MDB is deferred for future research. This would necessitate the annotation of additional non-adjacent data.

6.1.6.2 Discourse Connective Translations

Figures 34 to 39 illustrate the cross-lingual correspondence of explicit connectives between English and other languages using contingency tables. Each contingency table displays the distribution of English connectives (shown in rows) and their translations (shown in columns) in a normalized format. Darker shades of blue indicate more frequent translations. Regardless of the translation candidates, the ten most frequent English connectives were selected. Specific translation candidates in the target language were included, while instances where English connectives are represented as implicit connectives are grouped under the _implicit_ label. Instances where English connectives are omitted,

meaning aligned to a null token, are grouped under _omissions_. Less frequent translation occurrences are grouped under the _others_ label.

From these contingency tables, it can be observed that some English connectives have one or two dominant translations (e.g., English:if-Turkish:-sA), whereas majority of other English connectives are translated into multiple connectives in the target languages(e.g., English:But). All contingency tables suggest that English connectives are translated into a broader range of target language connectives.

Moreover, Figure 40 a-f shows the distribution of the ten most frequent English connectives grouped by the entropy of their translation alignments in each target language. English connectives and target language connectives with fewer than ten occurrences were excluded to avoid skewed distributions due to limited sample size. The translation alignment entropy is calculated using Shannon's entropy formula [115], as described by [113]. The formula for Shannon entropy $H(X)$ of a discrete random variable X is given Equation 7:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i) \tag{7}$$

where:

- X is a discrete random variable with possible values $x_1, x_2, \ldots, x_n$,
- $p(x_i)$ is the probability mass function of X.

For example, the English connective *but* has Turkish translation candidates *ama* (frequency = 17, $p = 0.47$), *ancak* (frequency = 12, $p = 0.33$), *fakat* (freq = 3, $p = 0.083$), *-sa da* (frequency = 1, $p = 0.027$), *aksine* (frequency = 1, $p = 0.027$), *sadece* (frequency = 1, $p = 0.027$) and *ve* (frequency = 1, $p = 0.027$). The alignment entropy for *but* is 1.91, given the provided alignment probability distributions. The translation contingency tables and alignment entropy bar plots indicate similar trends. For instance, in Figure 39, the English connective *But* has more translation variants in Turkish compared to other English connectives, resulting in one of the highest translation alignment entropy value shown in Figure 40-f. In contrast, the English connective *If* is 94 % of the time translated into the Turkish connective *-sA*, thus its translation alignment entropy in Figure 40-f is lower, 0.33.

Connectives whose senses are more predictable in a corpus are classified as strong and more specific. Additionally, [116] states that the specificity of a connective depends on the alternative connectives available in that language. For this reason, rather than approaching specificity from a qualitative perspective like [112], the specificity of each explicit connective in each language was measured via its relative entropy. This was calculated by dividing the entropy of its sense distribution by the entropy of all explicit relations. For the sense distribution, only Level-I Senses were taken into consideration. Each relative entropy value was rounded to one decimal point to enhance readability and facilitate comparison. A list of all English explicit connectives and their translation candidates in the target languages, along with their relative entropy (specificity), individual and alignment frequency values, is provided in Appendix B, Table 22. The average relative entropy values for each language are as follows: 0.018 for German, 0.062 for English, 0.11 for Lithuanian, 0.086 for Polish, 0.039 for Portuguese, 0.029 for Russian, and 0.092 for Turkish.

When a connective with higher specificity (high information value and lower relative sense entropy) is translated into one with lower specificity (low information value and higher relative sense entropy),

it is termed *underspecification*. Conversely, translating a less specific connective into a more specific one is called *specification*. In terms of underspecification, German, Russian and European Portuguese exhibits a slightly different pattern compared to Polish, Lithuanian and Turkish. In German, Russian and European Portuguese, English connectives are often translated into an equal or higher information value connective. In other target languages, however, English connectives are commonly translated into equal or lower information value connectives, with Lithuanian being the most underspecified. For individual connectives and their Level-I senses, patterns in the data are not consistent for all languages and sense groups. Furthermore, with respect to the specificity measure and Implicitation-Explicitation of connectives in English, it is quite naive to make definitive claims on whether they are language-dependent or translation-inherent, as we only have data where English is the source language. For an in-depth analysis, it is necessary to analyze texts translated from the target languages into English.[117] [113]

6.2 Summary

In this chapter, a comprehensive descriptive analysis of cross-linguistic discourse structures and the expression of discourse relations within TED-MDB languages were provided. By examining aligned discourse relations, the distinctions in discourse realization between various languages were explored. Addressing critical questions regarding differences in expression, semantic shifts, inter-sentential encoding patterns, and the effects of genre and contextual factors in connective translation, this chapter explains the variances in discourse structure between English and other target languages present in the TED-MDB corpus.

Through this analysis, significant distinctions in discourse structure and realization have been identified. Aligning discourse relations has enabled the recognition of patterns in discourse coherence and the frequency of implicitation and explicitation across different languages. Discourse relations exhibit varied expressions across languages, largely influenced by language-specific factors, translation and annotation methodologies. Among the notable shifts, AltLex relations frequently become Explicit, while EntRels tend to transform into Implicit discourse relations.

The chapter finds a high consistency in maintaining the top-level senses of discourse relations across languages. Although most sense shifts stem from translation processes, they affect how discourse relations are expressed in target languages. The consistent cross-lingual interpretation of discourse senses may suggest that translators have some flexibility in adapting the grammar of the source material to their respective languages; however, these formal variations do not significantly impact the semantic content, as the discourse senses are largely preserved.

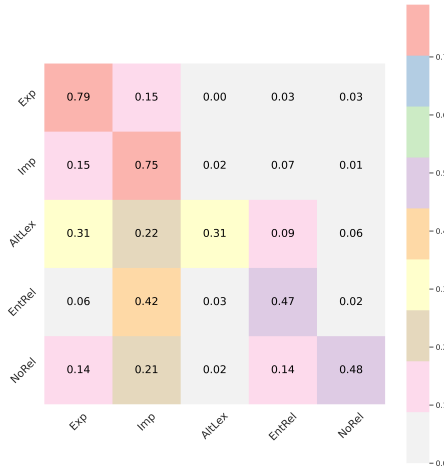
Regarding inter- and intra-sentential distinctions, it was observed that most discourse relations maintain their inter- or intra-sentential nature when translated. However, there is a notable tendency for intra-sentential relations in English to become inter-sentential in the target languages.

The structure of subtitled translations also plays a crucial role in the realization of discourse relations. Discourse relations appearing at the beginning of subtitle lines are more prone to implicitation and, to a lesser degree, explicitation. Explicit and implicit discourse relations are rarely positioned at the start of new paragraphs. Data presented here shows that subtitle boundaries might be effective in the implicitation of discourse relations.

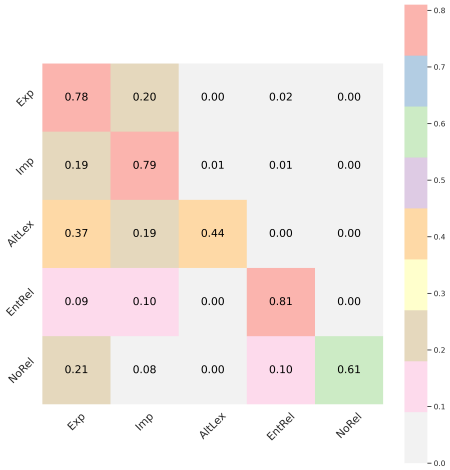
A detailed investigation of discourse connectives reveals important insights. By analyzing the most frequently used English connectives, the study assesses their translation alignment entropy and evaluates their specificity. This specific analysis highlights patterns of underspecification and specification depending on the target language. Furthermore, the chapter also investigates the anaphoric properties of connective structures in the TED-MDB corpus.

In conclusion, this thorough examination into the realization of discourse relations across TED-MDB languages emphasizes the value of aligned discourse relations in advancing multilingual understanding and facilitating cross-linguistic comparisons. The study's findings contribute crucial insights into the complexities of cross-linguistic discourse analysis and translation studies, paving the way for the development of more accurate and contextually sensitive translation methodologies. This work builds upon existing theoretical discussions and provides a refined understanding of discourse relations in a diverse multilingual setting. Nonetheless, this analysis primarily focuses on a descriptive examination, with a more thorough linguistic investigation reserved for future research studies.

(a) English - German

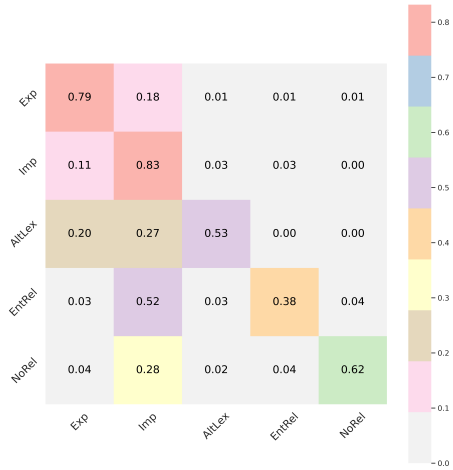


(b) **English - Lithuanian**

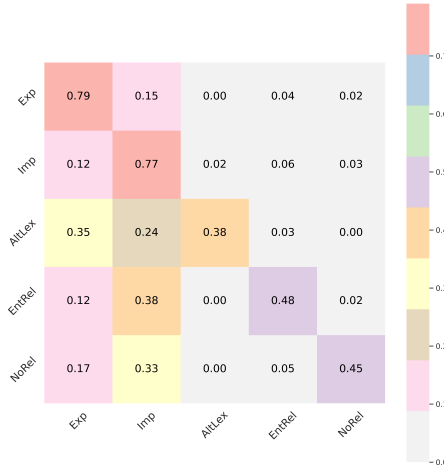


(c) English - Polish

(d) **English - Portuguese**



(e) English - Russian



(f) **English - Turkish**

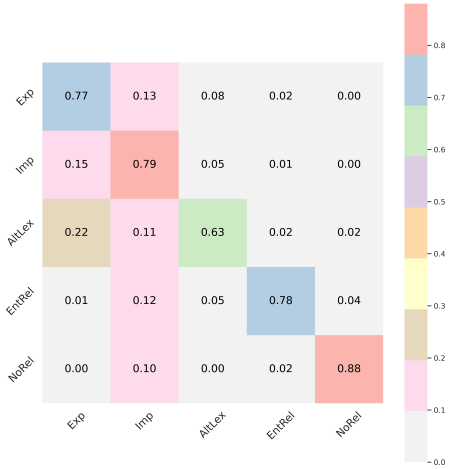
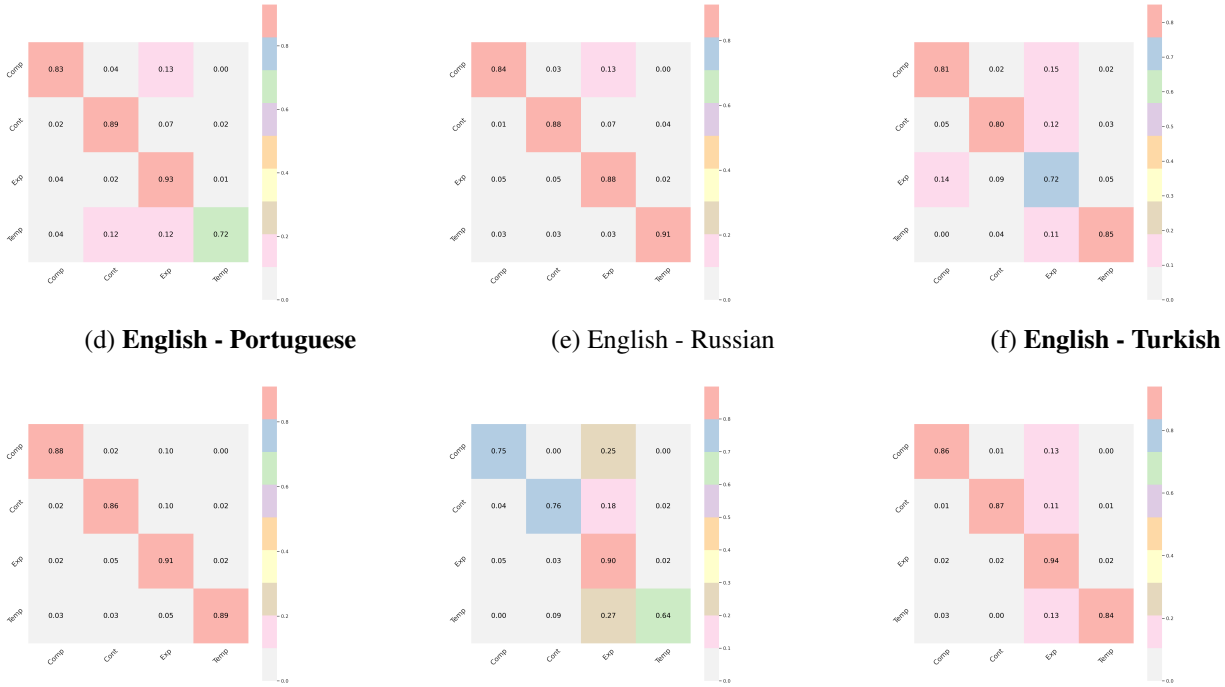


Figure 26: Heatmap visualizations of the confusion matrices for relation type of the linked discourse relations.

(a) English - German (b) **English - Lithuanian** (c) English - Polish



(d) **English - Portuguese** (e) English - Russian (f) **English - Turkish**

Figure 27: Heatmap visualizations of the confusion matrices for the top level senses of linked discourse relations.

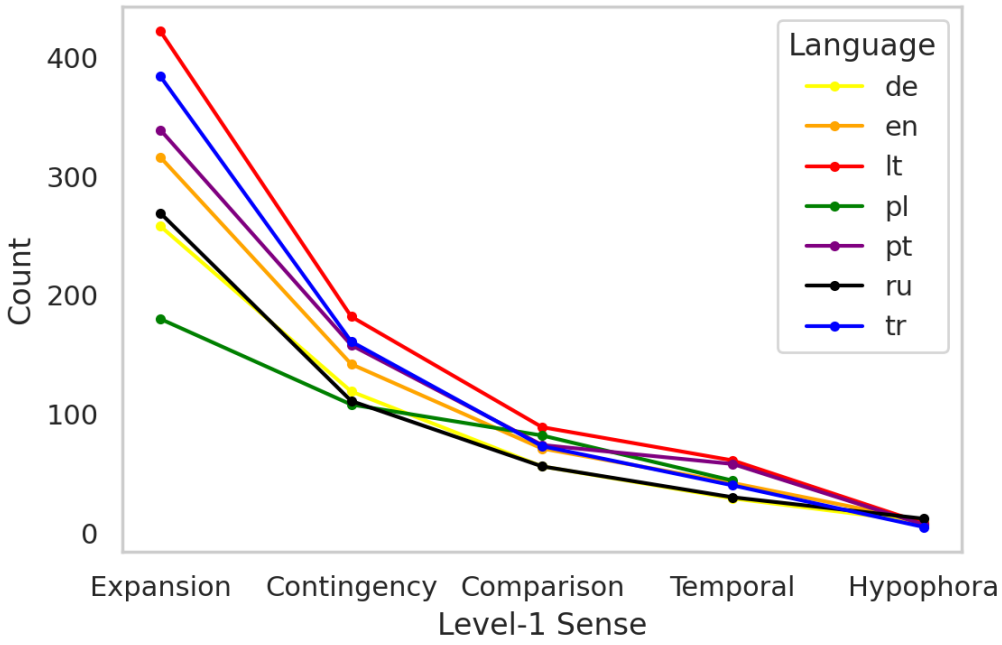
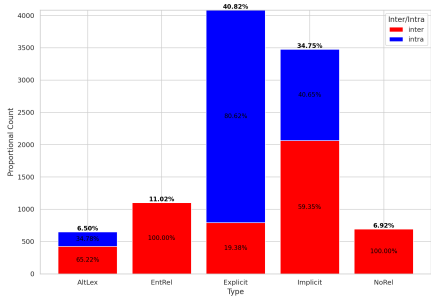
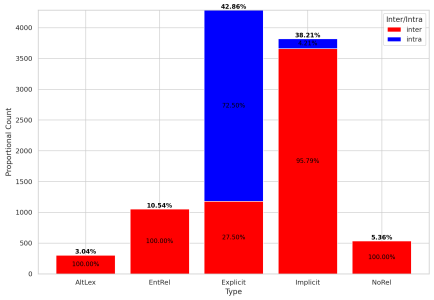


Figure 28: Level-I Sense Distribution in the New DR Alignment Data Set

(a) **English**

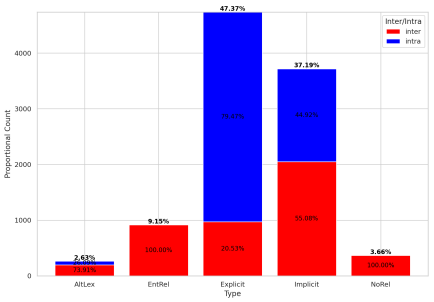


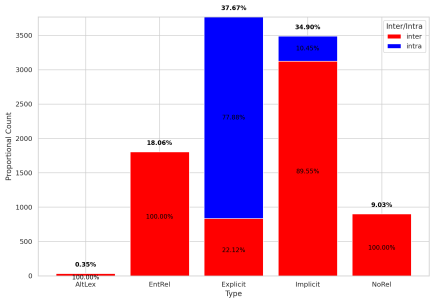
(b) German



(c) **Lithuanian**

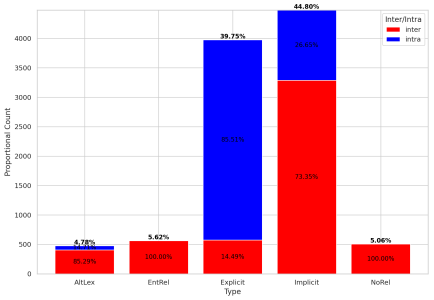
(d) Polish

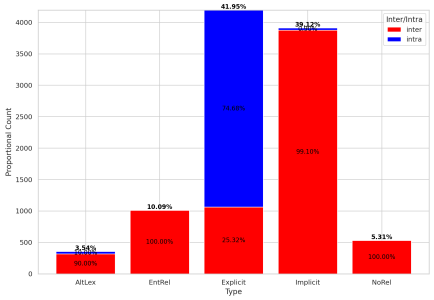




(e) **Portuguese**

(f) Russian





(g) **Turkish**

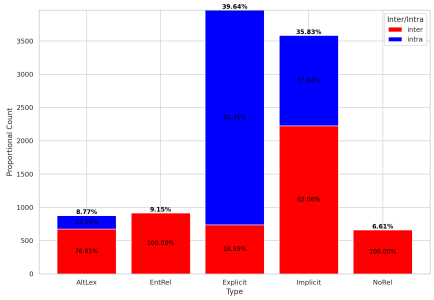
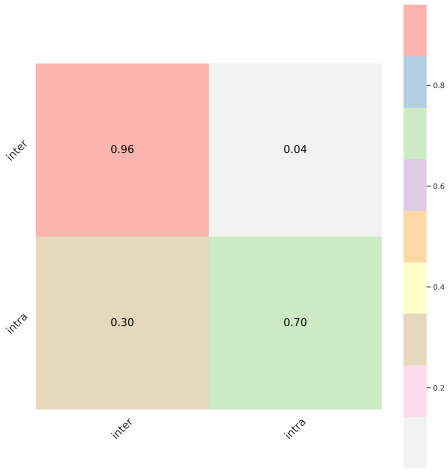
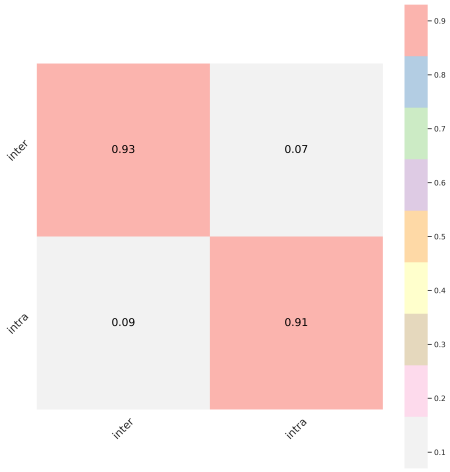


Figure 29: Stacked Bar Chart of DR Type Distribution by Inter-Intra Sentential Distinction

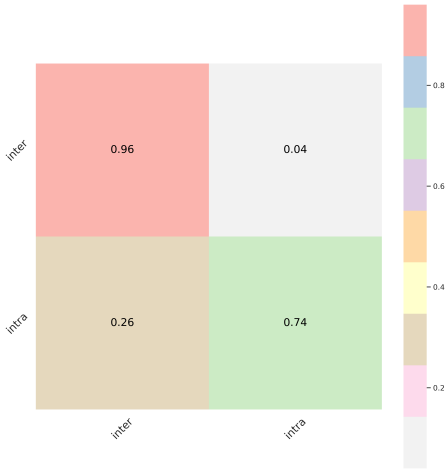
(a) English - German



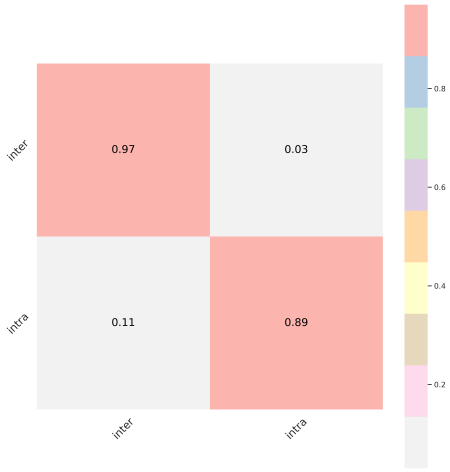
(b) **English - Lithuanian**



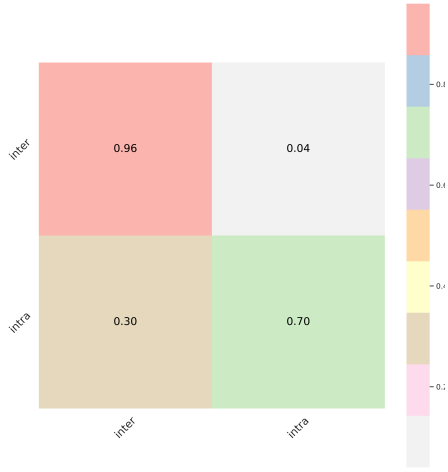
(c) English - Polish



(d) **English - Portuguese**



(e) English - Russian



(f) **English - Turkish**

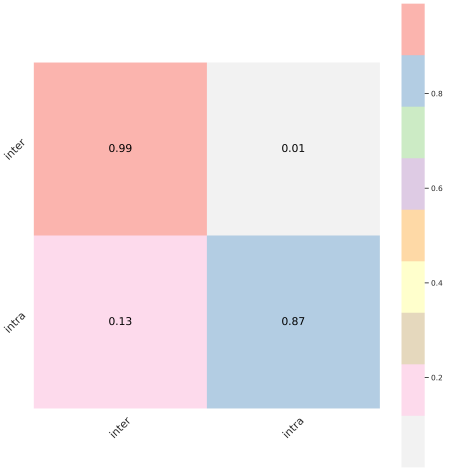


Figure 30: Heatmap visualizations of the confusion matrices for intra-inter distinction across linked discourse relations.

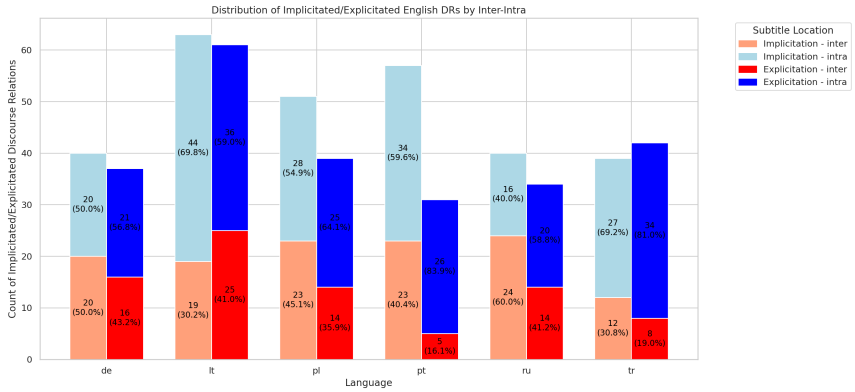
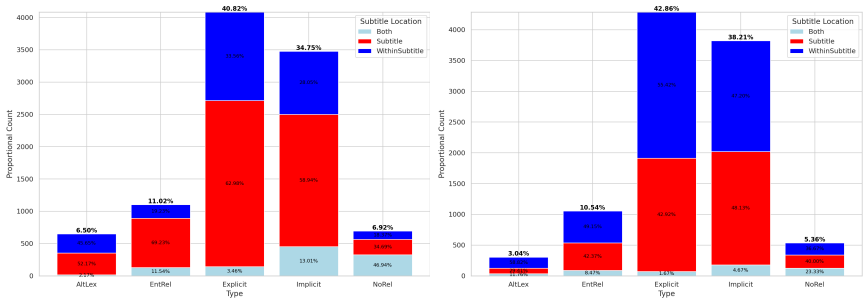


Figure 31: Explicitation versus Implicitation of English Inter-Intra Sentential DRs Across Different Languages

]

(a) **English** (b) German



(c) **Lithuanian** (d) Polish

(e) **Portuguese** (f) Russian

(g) **Turkish**

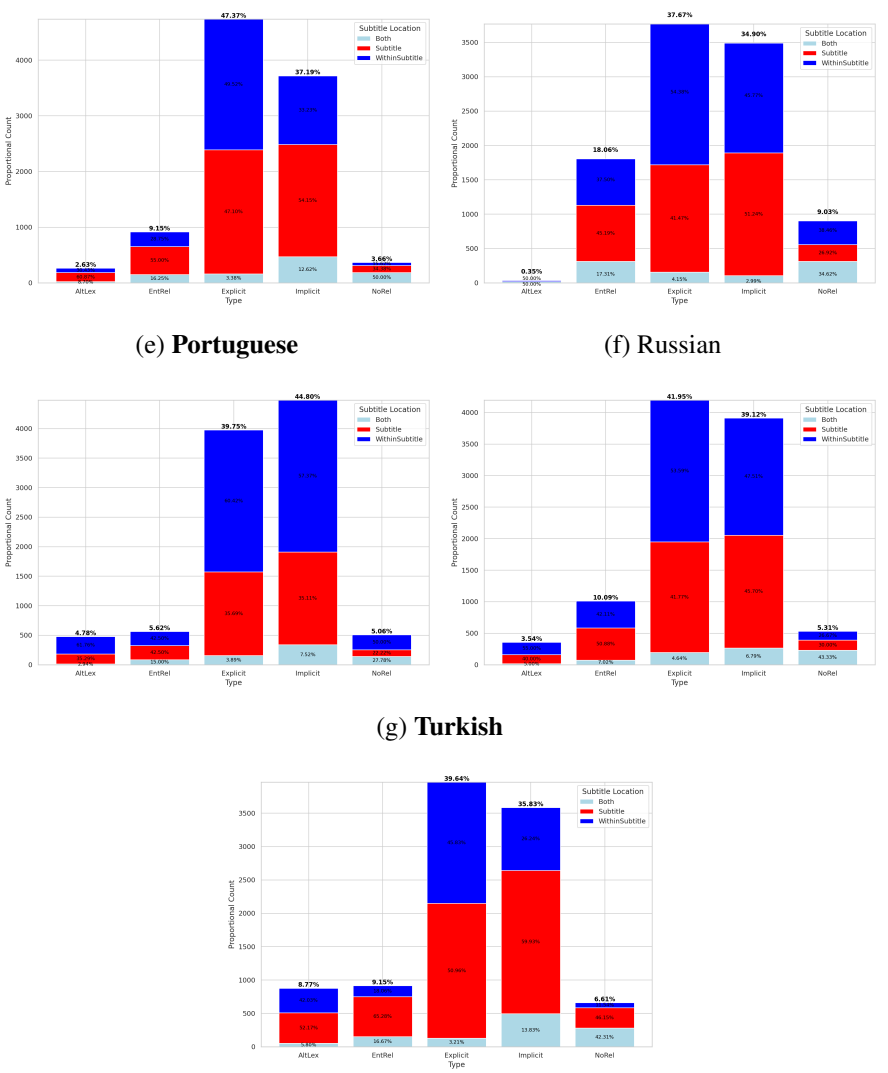


Figure 32: Stacked Bar Chart of DR Type Distribution Across Different Subtitle Line Locations

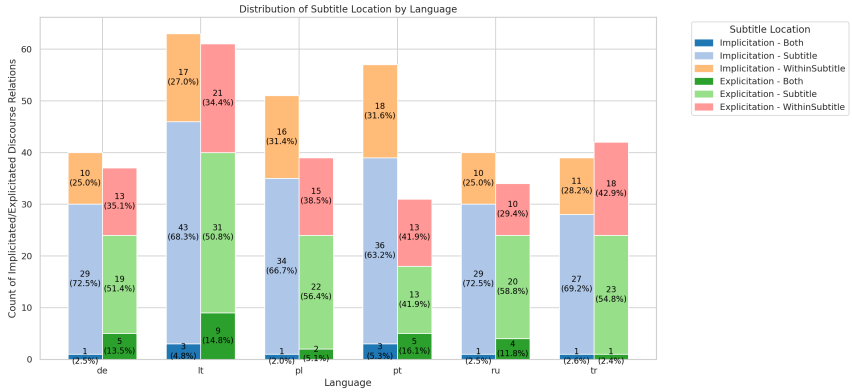


Figure 33: Explicitation versus Implicitation of DRs Across Different Subtitle Locations and Languages

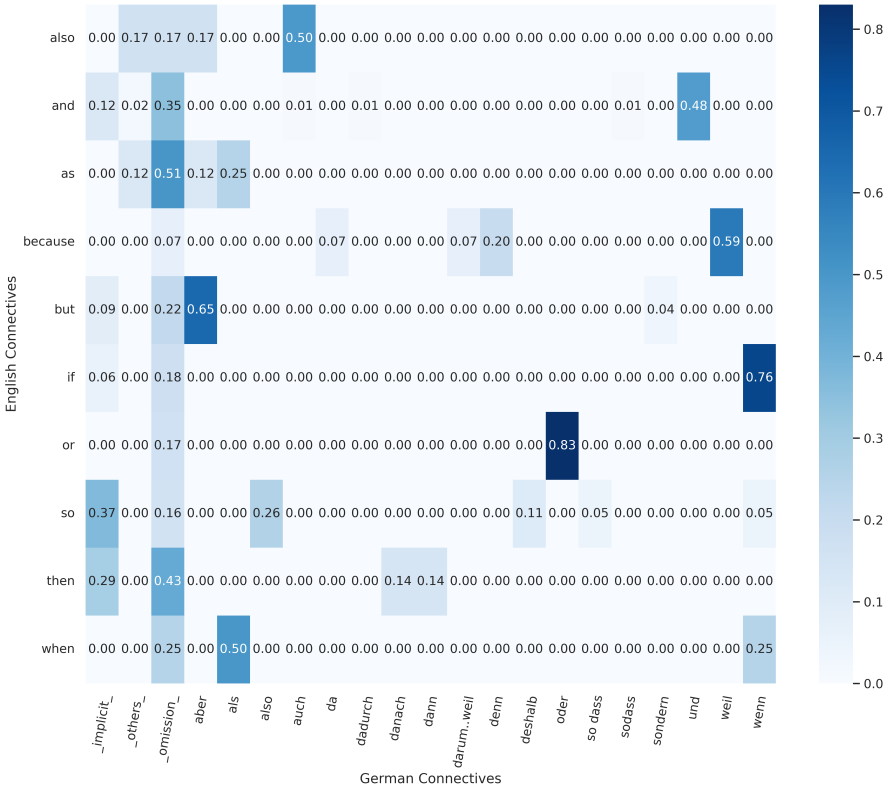


Figure 34: Translation of 10 most frequent English connectives to German Connectives

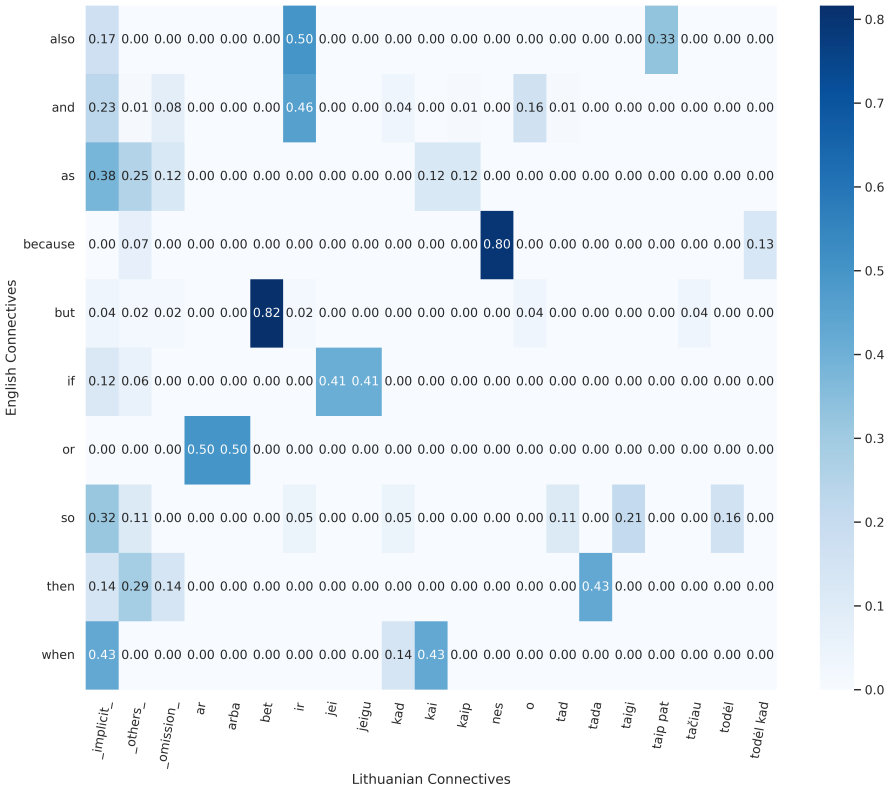


Figure 35: Translation of 10 most frequent English connectives to Lithuanian Connectives

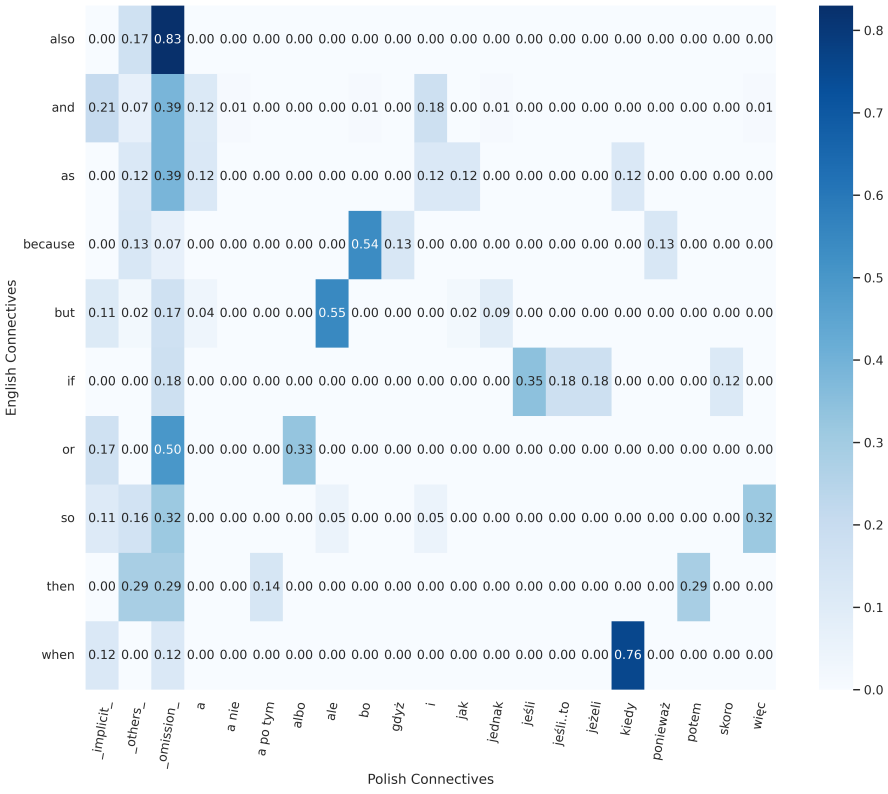


Figure 36: Translation of 10 most frequent English connectives to Polish Connectives

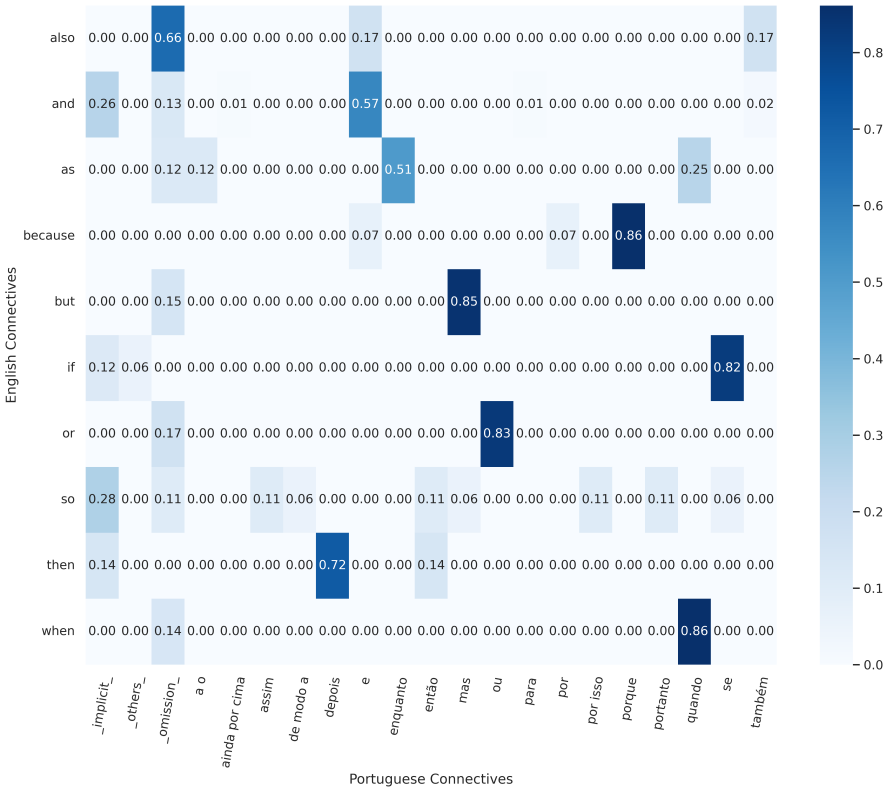


Figure 37: Translation of 10 most frequent English connectives to Portuguese Connectives

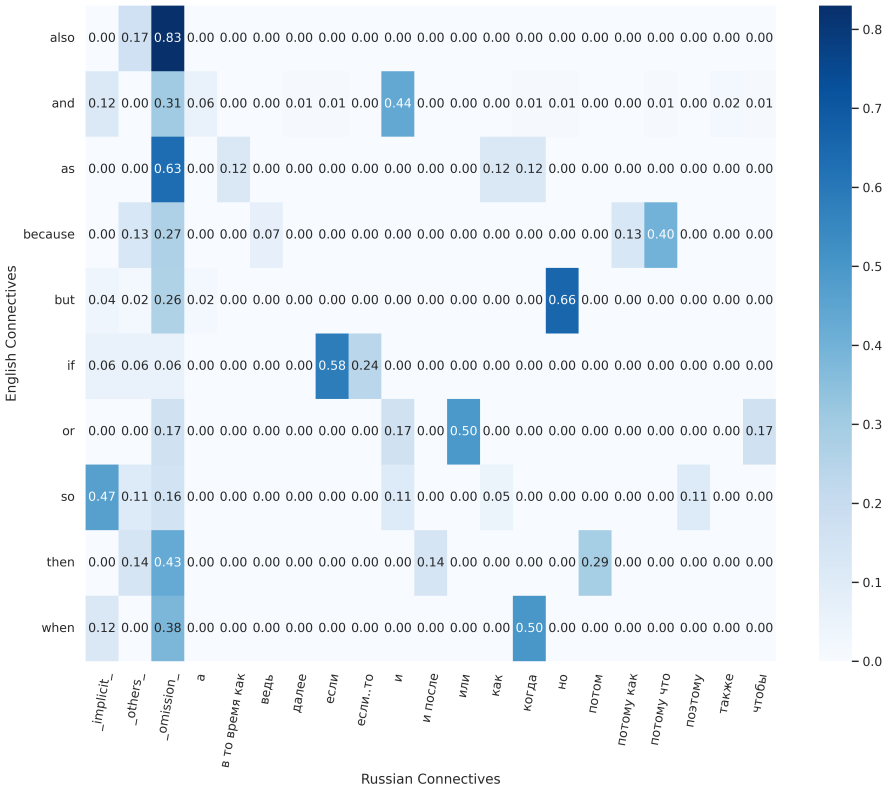


Figure 38: Translation of 10 most frequent English connectives to Russian Connectives

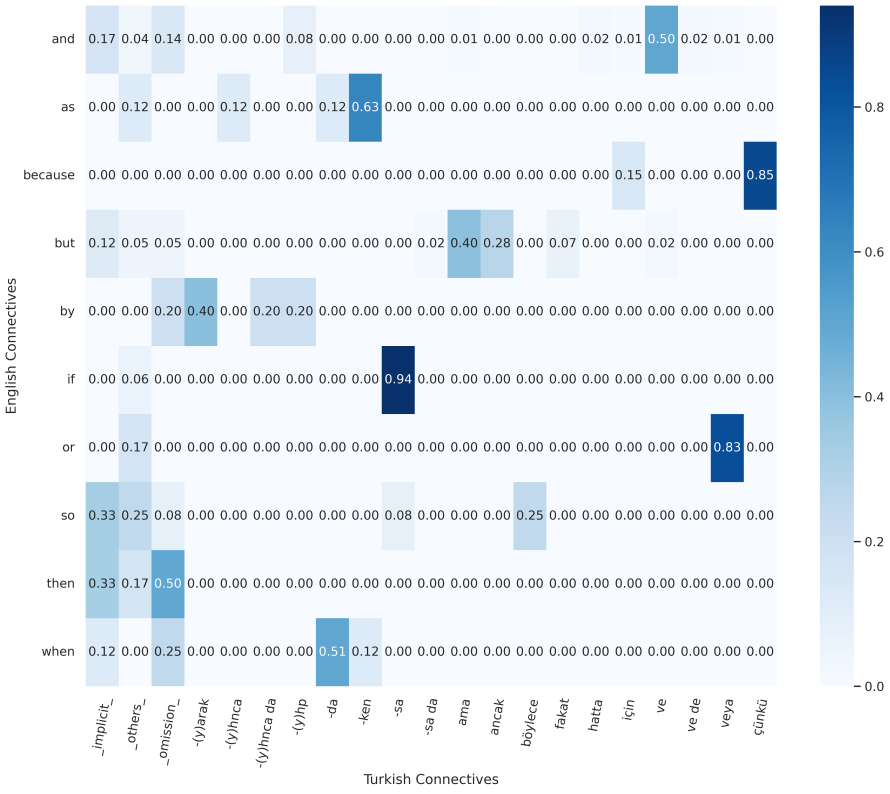
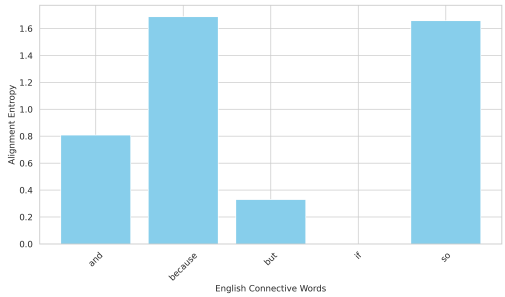
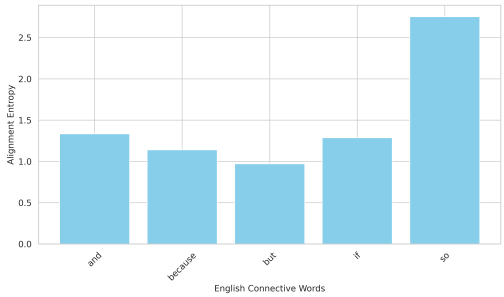


Figure 39: Translation of 10 most frequent English connectives to Turkish Connectives

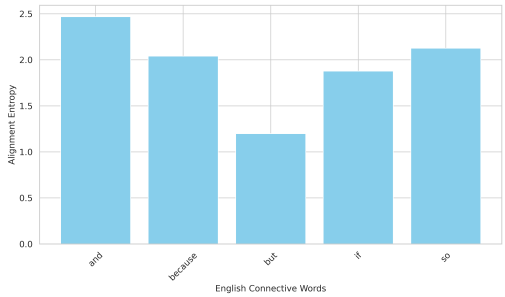
(a) English - German



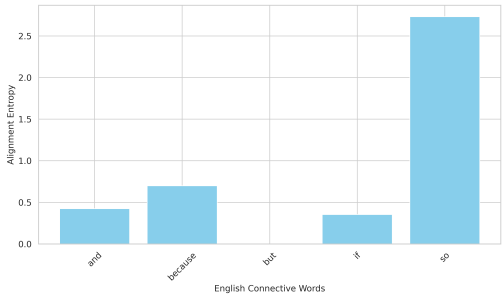
(b) **English - Lithuanian**



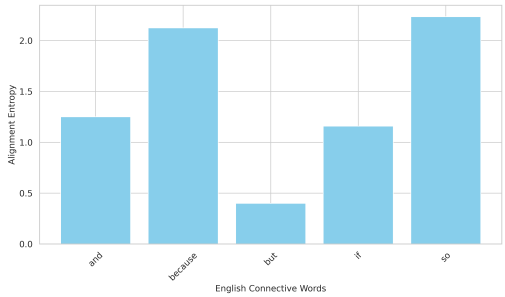
(c) English - Polish



(d) **English - Portuguese**



(e) English - Russian



(f) **English - Turkish**

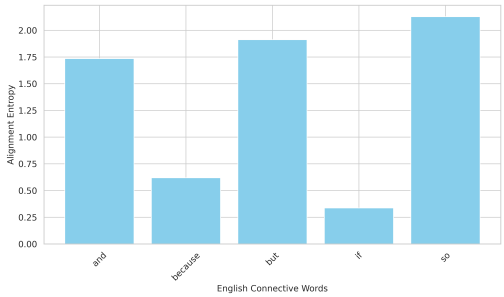


Figure 40: Distribution of English connectives grouped by the entropy of their translation alignments in target languages

CHAPTER 7

CONCLUSION

One of the main inference problems identified in pragmatics is discourse relations [118]. Each discourse relation implies an inference, reflecting a cognitive universal phenomenon. However, given the nuanced nature of natural languages, its manifestation may vary across different languages, often with distinct linguistic forms or even being disregarded altogether, particularly in translated texts. To investigate this phenomenon, each Discourse Relation (DR) must be identified in the Source Language (SL) text and its translation in the Target Language (TL) text. The motivation and contribution of this thesis arise from the need to provide insights into how discourse relations are realized in translated texts. The first objective is to develop a methodology for linking discourse relations in multilingual translational corpora.

This thesis presents a semantic-based framework to align DR annotation labels in parallel corpora, resulting in a set of aligned multilingual parallel DRs. From this data, bilingual discourse connective lexicons were generated. Furthermore, an alternative method for translation spotting is proposed, demonstrating how DR alignments can refine alignments in both SL and TL texts.

Using the aligned discourse relations, a comparative analysis across six languages (German, Lithuanian, European Portuguese, Polish, Turkish, and Russian) was conducted within a multilingual translation corpus. While the realization of discourse relations in both native and source texts has been extensively studied, most existing studies are confined to a small number of languages or focus on specific connectives or connective senses. This study, although based on a one-directional multilingual corpus with English as the source language, offers insights from multiple translated texts across several dimensions: discourse relation types, senses, scope (inter- and intra-sentential), genre effects, and discourse connectives in translation. The findings provide important insights into establishing local coherence, one of the main contributions of this thesis. Despite differences in translation, intrinsic typological properties of the target languages, and variations in annotation, the six languages exhibit similar patterns in translating DRs from the English source texts.

Although the target languages belong to different language families—Turkic, Indo-European (including Germanic, Baltic, and Slavic subfamilies), and Romance—they show similarities in how they realize discourse relations in texts translated from English. Statistical distance measures reveal that these languages exhibit comparable patterns in terms of discourse relation types, Level-I sense distributions, and discourse relation boundaries (inter- and intra-sentential). In a nutshell, while target languages diverge from English in aspects such as implicitation, explicitation, and inter-intra sentential realization, they display minimal differences in terms of discourse relation sense. This suggests that these three dimensions work synergistically to shape discourse in TED talks fundamentally.

The study found notable differences in how English connectives are translated across various languages. In almost all investigated target languages, English connectives are generally translated into multiple possible candidates. In German, Russian, and European Portuguese, English connectives are often translated into those with equal or higher information value, showing a trend of specification. Conversely, in Polish, Lithuanian, and Turkish, English connectives are frequently translated into those with equal or lower information value, showing a pattern of underspecification, with Lithuanian being the most underspecified. The results for German align with the findings of Yung [113].

Several factors, including language-specific characteristics, translational methodology, and annotation strategies, appear to contribute to the implicitation and explicitation of connectives, as well as the individual sense frequency of each connective [111]. Beyond semantic aspects, syntactic factors also play a crucial role. The way discourse relations connect units within the same sentence (intra-sententially) or across different sentences (inter-sententially) influences the degree of implicitness or explicitness. Intra-sentential discourse connectives are more prone to both explicitation and implicitation (with Russian as an exception to that). This is expected given the information load of intra-sentential discourse relations compared to inter-sentential relations, as well as the time and space constraints inherent in the subtitling genre. Additionally, it was observed that in almost all languages, English discourse relations at the beginning of a subtitle line are more frequently implicitated. This investigation of subtitle structure is, to the best of our knowledge, a first in the literature.

While it is not possible at this stage to isolate all the factors presented, the aligned discourse relations, discourse connective lexicons, and preliminary comparative results allow for further examination of the specific strategies employed by each language in encoding discourse structure. The resources and insights provided in this thesis are valuable for linguists studying pragmatic phenomena and for NLP researchers analyzing discursive structures in natural language texts for applications such as machine translation, information extraction, and text summarization.

7.1 Future Prospects

Several future research plans are outlined in our agenda. Our first and foremost aim is related to data linking the data linking methodology introduced in the background chapter aims to align the DR data to the standards of the Linked Language Open Data (LLOD) community and integrate it with the existing linked discourse marker inventory to provide standardized and expanded access for the research community.

Adjustments are planned for the DR alignment algorithm, especially in cases where argument spans and Level-I senses are shared by discourse relations, as illustrated in Example (17). To enhance segment similarity and connective calculations, context-sensitive transformers such as BERT and GPT will be used.

The developed methodology to align DRs is based on a parallel corpora annotated with the PDTB annotation scheme, which follows a lexicalized approach to discourse connectives. It remains a future study whether this methodology can be adapted for other discourse annotation schemes such as RST or SDRT, which are hierarchical. It may be necessary to include structural similarity of higher nodes in addition to the elementary discourse units (EDUs) for aligning all DRs.

Genre and modality differences in translating discourse relation both by human translators or and machine translation systems are pointed out in the literature[119],[120],[121] and [122].Therefore, our methodology tested on TED-MDB, focusing on prepared speech, should also be evaluated on other genres and languages annotated for discourse relations. Additionally, the specific impact of subtitling in the current genre warrants further investigation.

While the current thesis primarily examines the semantic aspect of cross-lingual variations in the realization of DRs across languages, future studies plan to incorporate Part of Speech (POS) information more to explore the interaction between syntax and semantics in DR translations. Also to differentiate between structural and anaphoric realization of discourse relations, further non-adjacent annotations are planned in TED-MDB, at least for Turkish.

Lastly, beyond the comparison with English as a reference language, future research will explore alignments for other language pairs like German-Turkish.

REFERENCES

[1] S. Özer, M. Kurfalı, D. Zeyrek, A. Mendes, and G. Valūnaitė Oleškevičienė, "Linking discourse-level information and the induction of bilingual discourse connective lexicons," *Semantic Web*, vol. 13, no. 6, pp. 1081–1102, 2022.

[2] A. Popescu-Belis, T. Meyer, J. Liyanapathirana, B. Cartoni, and S. Zufferey, "Discourse-level annotation over europarl for machine translation: Connectives and pronouns," in *Proceedings of the eighth international conference on Language Resources and Evaluation (LREC)*, 2012.

[3] M. Stede, S. Afantenos, A. Peldzsus, N. Asher, and J. Perret, "Parallel discourse annotations on a corpus of short texts," in *10th International Conference on Language Resources and Evaluation (LREC 2016)*, pp. 1051–1058, 2016.

[4] D. Samy and A. González-Ledesma, "Pragmatic annotation of discourse markers in a multilingual parallel corpus (arabic-spanish-english).," in *LREC*, Citeseer, 2008.

[5] S. Zufferey and L. Degand, "Annotating the meaning of discourse connectives in multilingual corpora," *Corpus linguistics and linguistic theory*, vol. 13, no. 2, pp. 399–422, 2017.

[6] D. Zeyrek, A. Mendes, Y. Grishina, M. Kurfalı, S. Gibbon, and M. Ogrodniczuk, "Ted multilingual discourse bank (ted-mdb): a parallel corpus annotated in the pdtb style," *Language Resources and Evaluation*, vol. 54, no. 2, pp. 587–613, 2020.

[7] B. Webber, R. Prasad, A. Lee, and A. Joshi, "The penn discourse treebank 3.0 annotation manual," *Philadelphia, University of Pennsylvania*, vol. 35, p. 108, 2019.

[8] M. Merkel, *Understanding and enhancing translation by parallel text processing*. PhD thesis, Linköpings universitet, 1999.

[9] P. Isabelle, "Bi-textual aids for translators," in *Proc. of the Annual Conference of the UW Center for the New OED and Text Research*, pp. 76–89, 1992.

[10] S. Özer and D. Zeyrek, "An automatic discourse relation alignment experiment on ted-mdb.," in *WNLP@ ACL*, pp. 31–34, 2019.

[11] W. A. Gale, K. W. Church, *et al.*, "A program for aligning sentences in bilingual corpora," *Computational linguistics*, vol. 19, no. 1, pp. 75–102, 1994.

[12] M. Kay and M. Roscheisen, "Text-translation alignment," *Computational linguistics*, vol. 19, no. 1, pp. 121–142, 1993.

[13] R. Östling and J. Tiedemann, "Efficient word alignment with markov chain monte carlo," *The Prague Bulletin of Mathematical Linguistics*, vol. 106, no. 1, p. 125, 2016.

[14] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[15] F. J. Och and H. Ney, "A comparison of alignment models for statistical machine translation," in *COLING 2000 Volume 2: The 18th International Conference on Computational Linguistics*, 2000.

[16] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE mobile computing and communications review*, vol. 5, no. 1, pp. 3–55, 2001.

[17] Q. Gao and S. Vogel, "Parallel implementations of word alignment tool," in *Software engineering, testing, and quality assurance for natural language processing*, pp. 49–57, 2008.

[18] C. Dyer, V. Chahuneau, and N. A. Smith, "A simple, fast, and effective reparameterization of IBM model 2," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (L. Vanderwende, H. Daumé III, and K. Kirchhoff, eds.), (Atlanta, Georgia), pp. 644–648, Association for Computational Linguistics, June 2013.

[19] M. Jalili Sabet, P. Dufter, F. Yvon, and H. Schütze, "SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings," in *Findings of the Association for Computational Linguistics: EMNLP 2020* (T. Cohn, Y. He, and Y. Liu, eds.), (Online), pp. 1627–1643, Association for Computational Linguistics, Nov. 2020.

[20] Z. Dou and G. Neubig, "Word alignment by fine-tuning embeddings on parallel corpora," *CoRR*, vol. abs/2101.08231, 2021.

[21] J. Kornfilt, *Turkish*. Routledge, 2013.

[22] J. Tiedemann, "Bitext alignment," 2011.

[23] H. H. Park, K. J. Zhang, C. Haley, K. Steimel, H. Liu, and L. Schwartz, "Morphology matters: A multilingual language modeling analysis," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 261–276, 2021.

[24] R. Cotterell and H. Schütze, "Morphological word embeddings," *arXiv preprint arXiv:1907.02423*, 2019.

[25] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[26] J.-H. Lee, S.-W. Lee, G. Hong, Y.-S. Hwang, S.-B. Kim, and H. C. Rim, "A post-processing approach to statistical word alignment reflecting alignment tendency between part-of-speeches," in *Coling 2010: Posters*, pp. 623–629, 2010.

[27] M. T. Cakmak, S. Acar, and G. Eryigit, "Word alignment for english-turkish language pair.," in *LREC*, pp. 2177–2180, 2012.

[28] I. D. El-Kahlout and K. Oflazer, "Exploiting morphology and local word reordering in english-to-turkish phrase-based statistical machine translation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1313–1322, 2009.

[29] E. Eyigöz, D. Gildea, and K. Oflazer, "Simultaneous word-morpheme alignment for statistical machine translation," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 32–40, 2013.

[30] A. Bisazza and M. Federico, "Morphological pre-processing for turkish to english statistical machine translation," in *Proceedings of the 6th International Workshop on Spoken Language Translation: Papers*, pp. 129–135, 2009.

[31] P. Schulz, W. Aziz, and K. Sima'an, "Word alignment without null words," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 169–174, 2016.

[32] Y. David, N. Grace, W. Richard, *et al.*, "Inducing multilingual text analysis tools via robust projection across aligned corpora," in *Proceedings of the First International Conference on Human Language Technology Research*, pp. 1–8, 2001.

[33] C. Xi and R. Hwa, "A backoff model for bootstrapping resources for non-english languages," in *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pp. 851–858, 2005.

[34] R. Hwa, P. Resnik, A. Weinberg, and O. Kolak, "Evaluating translational correspondence using annotation projection," in *Proceedings of the 40th annual meeting of the association for computational linguistics*, pp. 392–399, 2002.

[35] K. Spreyer and A. Frank, "Projection-based acquisition of a temporal labeller," in *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*, 2008.

[36] Y. Versley, "Discovery of ambiguous and unambiguous discourse connectives via annotation projection," in *Proceedings of Workshop on Annotation and Exploitation of Parallel Corpora (AEPC)*, pp. 83–82, 2010.

[37] M. Laali, *Inducing discourse resources using annotation projection*. PhD thesis, Concordia University, 2017.

[38] Q. Liu, F. Fancellu, and B. Webber, "Negpar: A parallel corpus annotated for negation," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[39] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data-the story so far," in *Linking the World's Information: Essays on Tim Berners-Lee's Invention of the World Wide Web*, pp. 115–143, 2023.

[40] A.-C. N. Ngomo, S. Auer, J. Lehmann, and A. Zaveri, "Introduction to linked data and its lifecycle on the web," *Reasoning Web. Reasoning on the Web in the Big Data Era: 10th International Summer School 2014, Athens, Greece, September 8-13, 2014. Proceedings 10*, pp. 1–99, 2014.

[41] T. Heath and C. Bizer, *Linked data: Evolving the web into a global data space*, vol. 1. Morgan & Claypool Publishers, 2011.

[42] C. Chiarcos, J. McCrae, P. Cimiano, and C. Fellbaum, "Towards open data for linguistics: Linguistic linked data," *New trends of research in ontologies and lexical resources: Ideas, projects, systems*, pp. 7–25, 2013.

[43] A. Gómez-Pérez, D. Vila-Suero, E. Montiel-Ponsoda, J. Gracia, and G. Aguado-de Cea, "Guidelines for multilingual linked data," in *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics*, pp. 1–12, 2013.

[44] C. Chiarcos and M. Ionov, "Linking discourse marker inventories," in *3rd Conference on Language, Data and Knowledge (LDK 2021)*, Schloss-Dagstuhl-Leibniz Zentrum für Informatik, 2021.

[45] D. Gromann, E.-S. Apostol, C. Chiarcos, M. Cremaschi, J. Gracia, K. Gkirtzou, C. Liebeskind, L. Mockiene, M. Rosner, I. Schuurman, *et al.*, "Multilinguality and llod: A survey across linguistic description levels," *Semantic Web*, no. Preprint, pp. 1–44, 2024.

[46] W. C. Mann and S. A. Thompson, "Rhetorical structure theory: Toward a functional theory of text organization," *Text-interdisciplinary Journal for the Study of Discourse*, vol. 8, no. 3, pp. 243–281, 1988.

[47] C. Chiarcos, "Towards interoperable discourse annotation: Discourse features in the ontologies of linguistic annotation," 2014.

[48] P. Cimiano, J. P. McCrae, and P. Buitelaar, "Lexicon model for ontologies: Community report," *W3C Ontology-Lexicon Community Group*, 2016.

[49] E. Pitler, M. Raghupathy, H. Mehta, A. Nenkova, A. Lee, and A. Joshi, "Easily identifiable discourse relations," in *Coling 2008: Companion volume: Posters*, pp. 87–90, 2008.

[50] A. Mendes and D. Zeyrek, "The discourse markers well and so and their equivalents in the portuguese and turkish subparts of the ted-mdb corpus," *Corpora in Translation and Contrastive Research in the Digital Age: Recent advances and explorations*, vol. 158, p. 209, 2021.

[51] M. Cettolo, C. Girardi, and M. Federico, "Wit3: Web inventory of transcribed and translated talks," in *Proceedings of the Conference of European Association for Machine Translation (EAMT)*, pp. 261–268, 2012.

[52] G. V. Oleskeviciene, D. Zeyrek, V. Mazeikiene, and M. Kurfalı, "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, pp. 53–58, 2018.

[53] N. Asher, *Reference to abstract objects in discourse*, vol. 50. Springer Science & Business Media, 2012.

[54] R. Prasad, B. Webber, and A. Joshi, "Reflections on the penn discourse treebank, comparable corpora, and complementary annotation," *Computational Linguistics*, vol. 40, no. 4, pp. 921–950, 2014.

[55] C. Fabricius-Hansen, "Informational density: A problem for translation and translation theory," *Linguistics*, vol. 34, pp. 521–566, 01 1996.

[56] A. Kehler and A. Kehler, "Coherence, reference, and the theory of grammar," 2002.

[57] R. Prasad, A. K. Joshi, and B. L. Webber, "Realization of discourse relations by other means: Alternative lexicalizations," in *COLING 2010, 23rd International Conference on Computational Linguistics, Posters Volume, 23-27 August 2010, Beijing, China*, pp. 1023–1031, 2010.

[58] N. Karamanis, "Supplementing entity coherence with local rhetorical relations for information ordering," *Journal of Logic, Language and Information*, vol. 16, pp. 445–464, 2007.

[59] A. Knott, J. Oberlander, M. O'Donnell, and C. Mellish, "Beyond elaboration: The interaction of relations and focus in coherent text," *Text representation: linguistic and psycholinguistic aspects*, pp. 181–196, 2001.

[60] R. Prasad, E. Miltsakaki, N. Dinesh, A. Lee, A. Joshi, L. Robaldo, and B. Webber, "The penn discourse treebank 2.0 annotation manual," *December*, vol. 17, p. 2007, 2007.

[61] A. Lee, R. Prasad, B. Webber, and A. Joshi, "Annotating discourse relations with the pdtb annotator," in *Proceedings of COLING 2016, The 26th International Conference on Computational Linguistics: System Demonstrations*, pp. 121–125, 2016.

[62] C. Chiarcos and A. Pareja-Lora, "1 open data—linked data—linked open data—linguistic linked open data (llod): A general introduction," *Development of Linguistic Linked Open Data Resources for Collaborative Data-Intensive Research in the Language Sciences*, p. 1, 2019.

[63] M. Baroni and S. Bernardini, "A new approach to the study of translationese: Machine-learning the difference between original and translated text," *Literary and Linguistic Computing*, vol. 21, no. 3, pp. 259–274, 2006.

[64] D. Zeyrek and B. L. Webber, "A discourse resource for turkish: Annotating discourse connectives in the metu corpus," in *Third International Joint Conference on Natural Language Processing, IJCNLP 2008, Hyderabad, India, January 7-12, 2008*, pp. 65–72, 2008.

[65] A. F. Acar, "Discovering the discourse role of converbs in turkish discourse," Master's thesis, Middle East Technical University, 2014.

[66] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

[67] J. Tiedemann, *Recycling translations: Extraction of lexical data from parallel corpora and their application in natural language processing*. PhD thesis, Acta Universitatis Upsaliensis, 2003.

[68] T. Kiss and J. Strunk, "Unsupervised multilingual sentence boundary detection," *Computational linguistics*, vol. 32, no. 4, pp. 485–525, 2006.

[69] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *Proceedings of machine translation summit x: papers*, pp. 79–86, 2005.

[70] M. Artetxe and H. Schwenk, "Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond," *Transactions of the association for computational linguistics*, vol. 7, pp. 597–610, 2019.

[71] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," in *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, MIT Press, 2017.

[72] J. Mírovskỳ, L. Mladová, and Š. Zikánová, "Connective-based measuring of the inter-annotator agreement in the annotation of discourse in pdt," in *Coling 2010: Posters*, pp. 775–781, 2010.

[73] V. Pyatkin, B. Webber, *et al.*, "Discourse relations and conjoined vps: Automated sense recognition," in *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 33–42, Association for Computational Linguistics (ACL), 2017.

[74] P. Christen and K. Goiser, "Quality and complexity measures for data linkage and deduplication," in *Quality measures in data mining*, pp. 127–151, Springer, 2007.

[75] M. Wetzel, S. Zufferey, and P. Gygax, "Second language acquisition and the mastery of discourse connectives: Assessing the factors that hinder l2-learners from mastering french connectives," *Languages*, vol. 5, no. 3, p. 35, 2020.

[76] S. Zufferey, W. Mak, L. Degand, and T. Sanders, "Advanced learners' comprehension of discourse connectives: The role of l1 transfer across on-line and off-line tasks," *Second Language Research*, vol. 31, no. 3, pp. 389–411, 2015.

[77] T. Meyer, A. Popescu-Belis, N. Hajlaoui, and A. Gesmundo, "Machine translation of labeled discourse connectives," in *Proceedings of the 10th Conference of the Association for Machine Translation in the Americas: Research Papers*, 2012.

[78] P. Bourgonje, Y. Grishina, and M. Stede, "Toward a bilingual lexical database on connectives: Exploiting a german/italian parallel corpus," in *Proceedings of the Fourth Italian Conference on Computational Linguistics–CLIC-IT*, pp. 53–58, 2017.

[79] D. Meurers and M. Dickinson, "Evidence and interpretation in language learning research: Opportunities for collaboration with computational linguistics," *Language Learning*, vol. 67, no. S1, pp. 66–95, 2017.

[80] M. Stede and S. Heintze, "Machine-assisted rhetorical structure annotation," in *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pp. 425–431, 2004.

[81] P. Bourgonje and M. Stede, "Exploiting a lexical resource for discourse connective disambiguation in german," in *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 5737–5748, 2020.

[82] M. Kurfalı, S. Ozer, D. Zeyrek, and A. Mendes, "Ted-mdb lexicons: Tr-enconnlex, pt-enconnlex," in *Proceedings of the First Workshop on Computational Approaches to Discourse*, pp. 148–153, 2020.

[83] M. Stede and C. Umbach, "Dimlex: A lexicon of discourse markers for text generation and understanding," in *COLING 1998 Volume 2: The 17th International Conference on Computational Linguistics*, 1998.

[84] C. Roze, L. Danlos, and P. Muller, "Lexconn: a french lexicon of discourse connectives," *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, no. 10, 2012.

[85] A. Feltracco, E. Jezek, B. Magnini, and M. Stede, "Lico: A lexicon of italian connectives," *CLiC it*, p. 141, 2016.

[86] J. Mírovskỳ, P. Synková, M. Rysová, and L. Poláková, "Czedlex-a lexicon of czech discourse connectives," *The Prague Bulletin of Mathematical Linguistics*, vol. 109, no. 1, p. 61, 2017.

[87] A. Mendes, I. D. R. Gayo, M. Stede, and F. Dombek, "A lexicon of discourse markers for portuguese–ldm-pt," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[88] D. Zeyrek and K. Başıbüyük, "TCL - a lexicon of Turkish discourse connectives," in *Proceedings of the First International Workshop on Designing Meaning Representations* (N. Xue, W. Croft, J. Hajic, C.-R. Huang, S. Oepen, M. Palmer, and J. Pustejovksy, eds.), (Florence, Italy), pp. 73–81, Association for Computational Linguistics, Aug. 2019.

[89] D. Das, M. Stede, S. S. Ghosh, and L. Chatterjee, "Dimlex-bangla: A lexicon of bangla discourse connectives," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 1097–1102, 2020.

[90] M. Stede, T. Scheffler, and A. Mendes, "Connective-lex: A web-based multilingual lexical resource for connectives," *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, no. 24, 2019.

[91] L. Poláková, K. Rysová, M. Rysová, and J. Mírovskỳ, "Geczlex: Lexicon of czech and german anaphoric connectives," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 1089–1096, 2020.

[92] C. Chiarcos, "Inducing discourse marker inventories from lexical knowledge graphs," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 2401–2412, 2022.

[93] A. Mendes and I. del Río, "Using a discourse bank and a lexicon for the automatic identification of discourse connectives," in *Computational Processing of the Portuguese Language: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings 13*, pp. 211–221, Springer, 2018.

[94] D. Das, T. Scheffler, P. Bourgonje, and M. Stede, "Constructing a lexicon of english discourse connectives," in *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pp. 360–365, 2018.

[95] M. Skrabal and M. Vavrin, "The translation equivalents database (treq) as a lexicographer's aid," in *Electronic lexicography in the 21st century: Proceedings of eLex 2017 conference*, pp. 124–137, Lexical Computing, 2017.

[96] B. Cartoni, S. Zufferey, T. Meyer, and A. Popescu-Belis, "How comparable are parallel corpora? measuring the distribution of general vocabulary and connectives," in *BUCC'11: Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pp. 78–86, 2011.

[97] B. Webber, M. Stone, A. Joshi, and A. Knott, "Anaphora and discourse structure," *Computational linguistics*, vol. 29, no. 4, pp. 545–587, 2003.

[98] K. Forbes-Riley, B. Webber, and A. Joshi, "Computing discourse semantics: The predicate-argument semantics of discourse connectives in d-ltag," *Journal of Semantics*, vol. 23, no. 1, pp. 55–106, 2006.

[99] D. Zeyrek, Ü. D. Turan, I. Demirşahin, and R. Cakıcı, "Differential properties of three discourse connectives in turkish," *Constraints in Discourse 3: Representing and Inferring Discourse Structure*, vol. 223, p. 183, 2012.

[100] R. A. Van der Sandt, "Presupposition projection as anaphora resolution," *Journal of semantics*, vol. 9, no. 4, pp. 333–377, 1992.

[101] M. Straka and J. Straková, "Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe," in *Proceedings of the CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies*, pp. 88–99, 2017.

[102] M. Honnibal and I. Montani, "spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. github," 2017.

[103] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A python natural language processing toolkit for many human languages," *arXiv preprint arXiv:2003.07082*, 2020.

[104] D. Altinok, "A diverse set of freely available linguistic resources for Turkish," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Toronto, Canada), pp. 13739–13750, Association for Computational Linguistics, July 2023.

[105] H. Gottlieb, "Subtitling-a new university discipline," in *Teaching translation and interpreting*, p. 161, John Benjamins, 1992.

[106] A. A. Rosa, "Features of oral and written communication in subtitling," in *(Multi) Media Translation*, p. 213, John Benjamins, 2001.

[107] D. Zeyrek, A. Mendes, G. V. Oleškevičienė, and S. Özer, "An exploratory analysis of ted talks in english and lithuanian, portuguese and turkish translations: Results from the analysis of an annotated multilingual corpus," *Contrastive Pragmatics*, vol. 3, no. 3, pp. 452–479, 2022.

[108] D. Zeyrek and M. Kurfalı, "TDB 1.1: Extensions on Turkish discourse bank," in *Proceedings of the 11th Linguistic Annotation Workshop*, (Valencia, Spain), pp. 76–81, Association for Computational Linguistics, Apr. 2017.

[109] Y. Ji and J. Eisenstein, "One vector is not enough: Entity-augmented distributed semantics for discourse relations," *Transactions of the Association for Computational Linguistics*, vol. 3 , pp. 329–344, 2015.

[110] J. Park and C. Cardie, "Improving implicit discourse relation recognition through feature set optimization," in *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, (Seoul, South Korea), pp. 108–112, Association for Computational Linguistics, July 2012.

[111] L. Poláková, J. Mírovskỳ, Š. Zikánová, and E. Hajičová, "Discourse relations and connectives in higher text structure," *Dialogue & Discourse*, vol. 12, no. 2, pp. 1–37, 2021.

[112] L. Crible, Á. Abuczki, N. Burkšaitienė, P. Furkó, A. Nedoluzhko, S. Rackevičienė, G. V. Oleškevičienė, and Š. Zikánová, "Functions and translations of discourse markers in ted talks: A parallel corpus study of underspecification in five languages," *Journal of Pragmatics*, vol. 142, pp. 139–155, 2019.

[113] F. Yung, M. Scholman, E. Lapshinova-Koltunski, C. Pollkläsener, and V. Demberg, "Investigating explicitation of discourse connectives in translation using automatic annotations," in *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 21–30, 2023.

[114] L. Poláková and J. Mírovskỳ, "Anaphoric connectives and long-distance discourse relations in czech," *Computación y Sistemas*, vol. 23, no. 3, pp. 711–717, 2019.

[115] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, 1948.

[116] S. Zufferey and B. Cartoni, "A multifactorial analysis of explicitation in translation," *Target. International Journal of Translation Studies*, vol. 26, no. 3, pp. 361–384, 2014.

[117] K. Klaudy, "The asymmetry hypothesis in translation research," *Translators and their readers. In Homage to Eugene A. Nida. Brussels: Les Editions du Hazard*, vol. 283, p. 303, 2009.

[118] D. Jurafsky, "Pragmatics and computational linguistics," *The handbook of pragmatics*, pp. 578–604, 2006.

[119] M. Taboada and M. Gómez-González, "Discourse markers and coherence relations: Comparison across markers, languages and modalities," *Linguistics and the Human Sciences*, vol. 6, no. 1-3, pp. 17–41, 2012.

[120] E. Lapshinova-Koltunski, "Exploration of inter-and intralingual variation of discourse phenomena," in *Proceedings of the Second Workshop on Discourse in Machine Translation*, pp. 158–167, 2015.

[121] L. Crible and V. Demberg, "When do we leave discourse relations underspecified? the effect of formality and relation type," *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, no. 26, 2020.

[122] M. Scholman, T. Dong, F. Yung, and V. Demberg, "Comparison of methods for explicit discourse connective identification across various domains," in *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pp. 95–106, 2021.

Appendix A

LIST OF DISCOURSE RELATIONS

Table 21: Discourse Relations Realized Across More Than One Sentences

Lang.	File	Connective	Segment	Level-I Sense	POS Tag	Position
pt	1927	Mas	*O investimento cauteloso e a teoria financeira não estão subordinados à sustentabilidade . São compatíveis . Portanto , não estou a falar apenas de compensações* Mas **os investidores institucionais são a peça chave para a sustentabilidade**	Comparison	CCONJ	Start
pt	1927	Assim	*Cerca de 80 % de_os administradores veem a sustentabilidade como a base para o crescimento na inovação e para alcançar vantagens competitivas nos seus setores industriais. Mas 93 % consideram o ASG como o futuro, ou igualmente importante para o futuro de_os seus negócios* Assim **a visão de_os administradores é clara**	Contingency	ADV	Start
pt	1971	Resumindo	*Usei imagens de ressonância magnética para captar a real anatomia de_o paciente, depois usei modelagem por elementos finitos para prever melhor as tensões e os esforços internos sobre as forças normais, para depois criar um encaixe de prótese para fabrico. Usamos uma impressora 3D para criar um encaixe feito de vários materiais que alivia a pressão onde é necessário, de acordo com a anatomia de_o paciente* Resumindo **estamos a usar dados para fazer novos encaixes, rapidamente e com baixo custo**	Expansion	VERB	Start

Table 21 cont.

Lang.	File	Connective	Segment	Level-I Sense	POS Tag	Position
pt	1976	E	*toda essa luz destrói a sombra. É uma sombra terrível* E **não conseguimos ver planetas**	Expansion	CCONJ	Start
pt	2150	assim	*Têm na mesma o grupo de pessoas envolvidas no governo, na imprensa, os políticos, os colunistas. TEDxRio está em baixo à direita, a_ o pé de_ os bloguistas e escritores. E também temos esta grande diversidade de pessoas interessadas em diferentes tipos de música. Até há fãs de Justin Bieber aqui representados. Outras boy bands, cantores country, música gospel, funk e rap e artistas de comédia stand-up. Há mesmo toda uma secção em torno de drogas e anedotas. Não é fixe? A equipa de futebol Flamengo também está representada aqui* **Temos** assim **o mesmo tipo de difusão no desporto, na sociedade civil, na arte e na música**	Contingency	ADV	Within
tr	1927	Ancak	*Sonsuz sosyal sorumlulukları yoktur ve mantıklı yatırım ile finans teorisi sürdürülebilirlikten aşağı değildir. Birbiriyle uyumludurlar. O yüzden burada birinden ödün vermekten bahsetmiyorum* Ancak **kurumsal yatırımcılar sürdürülebilirlikte x faktörüdür**	Comparison	CCONJ	Start
tr	1927	Ve	*Mavide, MSCI Worldü görüyoruz. Bu, dünyadaki gelişmiş pazarlardaki büyük şirketlerin indeksi* Ve **altın rengi olarak, ÇSY performansı en iyi olarak değerlendirilmiş şirketlerin bir alt kümesini görüyoruz**	Expansion	CCONJ	Start
tr	1971	Kısacası	*Hastanın anatomisinin gerçek biçimini yakalamak için manyetik rezonans görüntülemesini kullandım ve normal kuvvetlerde iç gerilme ve deformasyon noktalarını daha iyi tahmin edebilmek için sonlu elemanlar modellemesini kullandım, sonra da üretilmek üzere bir protez soketi yaptım. Hastanın anatomisinde gerekli yerlerdeki baskıyı azaltan çok parçalı bir protez soketi yapmak için üç boyutlu yazıcı kullandık* Kısacası **çabuk ve ucuz yeni soketler yapmak için verileri kullanıyoruz**	Expansion	ADV	Start
tr	2150	Fakat	*Ağın diğer ucunda Hip-hop müzik gibi şeylerden hoşlanan topluluğu görebilirsiniz. hatta onlar DC/Maryland/Virgina bölgesinde, doğru tanımıyla Baltimore şehrinin üst tarafında yaşamakla özdeşleşirler* Fakat **ortada iki toplumu birleştiren bir şey görürsünüz bu spordur**	Comparison	CCONJ	Start

Table 21 cont.

Lang.	File	Connective	Segment	Level-I Sense	POS Tag	Position
tr	2150	Ancak	*Şehirlerin haritalarını oluşturmayı düşündüğümüzde yollar, sokaklar, caddeler, binalar ve şehirlerin oluşumuna yol açan yerleşim hikayeleri aklımıza gelir. Ya da bir kentsel tasarımcının cesur vizyonunu düşünebilirsiniz* Ancak **şehirlerin haritalarını oluşturmayı düşünmenin ve yapmanın başka yolları da var**	Comparison	CCONJ	Start
tr	2150	da	*LGBT toplumunun anti-sosyal topluluk ile çok da iyi geçinmediğini görebilirsiniz. Keza, sanat topluluğu, müzik topluluğu ile de* **Bu** da **bu tür şeylere sebebiyet veriyor**	Expansion	CCONJ	Within
tr	2150	da	*LGBT toplumunun anti-sosyal topluluk ile çok da iyi geçinmediğini görebilirsiniz. Keza, sanat topluluğu, müzik topluluğu ile de* **Bu** da **bu tür şeylere sebebiyet veriyor**	Expansion	CCONJ	Within
tr	2150	Böylece	*Rio, Baltimore ve San Francisco'nun tersine tam bir heterojen şehir. Hala hükümetle, gazetelerle, politikayla, köşe yazarlarıyla ilintili bir grup insan var. TEDx Rio, sağ aşağıda, blogcuların ve yazarların tam yanındadır. Fakat, değişik müzik tarzlarıyla ilgilenen büyük miktarda insan çeşitliliğine de sahipsiniz ayrıca. Justin Bieber hayranları bile burada gösterilmektedir. Diğer müzik grupları, country şarkıcıları, dini müzik, funk, rap ve stand up komedi, uyuşturucu ve esprileri içeren tüm bir bölge bile var. Güzel, değil mi? Devam edersek, Flamengo futbol takımı da burada temsil edilmektedir* Böylece **spor, müzik, sivil, sanat ve müziğin aynı dağılımına sahipsiniz**	Contingency	ADV	Start
tr	2150	Peki	*bütün verilere sahibiz. Şehirler hakkında inanılmaz zenginlikte bilgiye sahibiz şimdi. Belki de şimdiye kadar sahip olduklarımızdan en zengini* Peki **bununla ne yapabiliriz**	Expansion	ADV	Start
en	1927	But	*They don't have indefinite social obligations, and prudent investing and finance theory aren't subordinate to sustainability. They're compatible. So I'm not talking about tradeoffs here* But **institutional investors are the x-factor in sustainability**	Comparison	CCONJ	Start
en	1927	But	*Good, you like it. I like it too. (Laughter) I like it because it pokes fun at both sides of the climate change issue* But **what I really like about it is that it reminds me of something Mark Twain said, which is, Plan for the future, because that's where you're going to spend the rest of your life**	Expansion	CCONJ	Start

Table 21 cont.

Lang.	File	Connective	Segment	Level-I Sense	POS Tag	Position
en	1927	So	*About 80 percent of global CEOs see sustainability as the root to growth in innovation and leading to competitive advantage in their industries. But 93 percent see ESG as the future, or as important to the future of their business* So **the views of CEOs are clear**	Contingency	ADV	Start
en	1927	And	*In blue, we see the MSCI World. It's an index of large companies from developed markets across the world* And **in gold, we see a subset of companies rated as having the best ESG performance**	Comparison	CCONJ	Start
en	1971	In short	*I used magnetic resonance imaging to capture the actual shape of the patient's anatomy, then use finite element modeling to better predict the internal stresses and strains on the normal forces, and then create a prosthetic socket for manufacture. We use a 3D printer to create a multi-material prosthetic socket which relieves pressure where needed on the anatomy of the patient* In short **we're using data to make novel sockets quickly and cheaply**	Expansion	ADP ADJ	Start
en	1976	That's why	*All we're seeing is the big beaming image of the star that's ten billion times brighter than the planet, which should be in that little red circle. That's what we want to see* That's why **it's hard**	Contingency	ADP ADV	Start
en	2150	But	*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. (Laughter) Now, on the other side of the network, you tend to have primarily African-American and Latino folks who are really concerned about somewhat different things than the geeks are, but just to give some sense, the green part of the network we call Smalltimore, for those of us that inhabit it, because it seems as though we're living in a very small town. We see the same people over and over again, but that's because we're not really exploring the full depth and breadth of the city. On the other end of the network, you have folks who are interested in things like hip-hop music and they even identify with living in the DC/Maryland/Virginia area over, say, the Baltimore city designation proper* But **in the middle, you see that there's something that connects the two communities together, and that's sports**	Comparison	CCONJ	Start

Table 21 cont.

Lang.	File	Connective	Segment	Level-I Sense	POS Tag	Position
en	2150	So	*It's a very heterogeneous city in a way that Baltimore or San Francisco is not. You still have the lobe of people involved with government, newspapers, politics, columnists. TEDxRio is down in the lower right, right next to bloggers and writers. But then you also have this tremendous diversity of people that are interested in different kinds of music. Even Justin Bieber fans are represented here. Other boy bands, country singers, gospel music, funk and rap and stand-up comedy, and there's even a whole section around drugs and jokes. How cool is that? And then the Flamengo football team is also represented here* So **you have that same kind of spread of sports and civics and the arts and music, but it's represented in a very different way, and I think that maybe fits with our understanding of Rio as being a very multicultural, musically diverse city**	Contingency	ADV	Start
en	2150	So	*we have all this data. It's an incredibly rich set of data that we have about cities now, maybe even richer than any data set that we've ever had before* So **what can we do with it**	Contingency	ADV	Start
de	1976	Deshalb	*Alles, was wir sehen, ist das große strahlende Bild des Sterns, der 10-Mrd.-mal heller leuchtet als der Planet, der in dem kleinen roten Kreis sein sollte. Das wollen wir sehen* Deshalb **ist es so schwer**	Contingency	ADV	Start
de	1978	allerdings	*Erfolg haben, bedeutet also, den Ring Nr. 10 zu treffen, aber Meisterschaft lässt erkennen, dass das nichts bedeutet, wenn man es nicht immer wieder wiederholen kann* **Meisterschaft ist** allerdings **nicht gleich Spitzenleistung**	Comparison	ADV	Within

Table 21 cont.

Lang.	File	Connective	Segment	Level-I Sense	POS Tag	Position
de	1978	mit anderen Worten	*Elizabeth Murray hat mich mit der Ansicht über ihre früheren Gemälde überrascht. Der Maler Paul Cézanne hat so oft gedacht, seine Arbeiten wären unvollständig, dass er sie unbeachtet links liegen ließ mit der Absicht, sie später wieder hervorzuholen, aber am Ende seines Lebens hatte er als Ergebnis nur zehn Prozent seiner Gemälde signiert. Sein Lieblingsroman war "Das unbekannte Meisterwerk" von Honoré de Balzac und er fühlte sich selbst als Hauptdarsteller. Franz Kafka sah Unvollständigkeit, wenn andere seine Werke nur loben konnten, so sehr, dass er all seine Tagebücher, seine Manuskripte, Briefe und sogar Skizzen nach seinem Tod verbrannt haben wollte. Sein Freund lehnte diese Bitte ab, weshalb wir heute all die Werke von Kafka kennen: "Amerika" ("Der Verschollene"), "Der Prozess" und "Das Schloss"; eine Arbeit so unvollständig, dass sie sogar mitten im Satz aufhört* **Die Verfolgung der Meisterschaft** mit anderen Worten **ist fast so etwas wie ein ewiges Vorwärtsstreben**	Expansion	ADP DET NOUN	Within
lt	1927	Nes	*Jie artėja prie 100 procentų tvaraus investavimo, sistemingai integravę ASV visose fondo veiklose. Kodėl?* Nes **jie mano, kad tai lemia geriausią ilgalaikę grąžą, ne mažiau**	Contingency	SCONJ	Start
lt	1927	Taigi	*Aplikosauga apima energijos vartojimą, prieigą prie vandens, atliekų tvarkymą ir taršą ir ekonomišką išteklių naudojimą. Socialinė pusė – žmogiškasis kapitalas, įdarbinimo klausimai ir gebėjimas imtis inovacijų, taip pat tiekimo grandinės valdymas ir darbuotojų teisės bei žmogaus teisės. O valdymas – tai įmonių priežiūra vykdoma valdybų ir investuotojų* Taigi **kaip ir sakiau, tai išties patrauklūs dalykai**	Contingency	PART	Start
lt	1976	Todėl	*Tematome tik šviečiantį žvaigždės vaizdą, kuris dešimt milijardų kartų ryškesnis už planetą, kuri turėtų būti raudoname apskritime. Tai ir norime pamatyti* Todėl **tai ir sunku**	Contingency	ADV	Start
lt	2150	Bet	*Vis dar turime sluoksnį žmonių dirbančių valdžioje, laikraščiuose, politikoje, rašytojų. TEDxRio yra apačioje dešinėje, šalia blogerių ir rašytojų* Bet **taip pat yra ši stebinanti žmonių įvairovė, žmonių, kurie domisi įvairiausia muzika**	Comparison	CCONJ	Start

Appendix B

LIST OF EXPLICIT DISCOURSE CONNECTIVES

Table 22: Connective Alignment List

Source Language					Target Language					Alignment Frequency
Language	Connective Word	Level-I Sense	Relative Entropy	Frequency	Language	Connective Word	Level-I Sense	Relative Entropy	Frequency	
en	also	Expansion	0.39	6	de	aber	Expansion	0.097	37	1
en	as	Temporal	0.326	8	de	aber	Comparison	0.097	37	1
en	but	Comparison	0.4	46	de	aber	Comparison	0.097	37	30
en	on the one hand..but	Comparison	0	1	de	aber	Comparison	0.097	37	1
en	however	Comparison	0	1	de	allerdings	Comparison	0	2	1
en	though	Expansion	0.6	2	de	allerdings	Comparison	0	2	1
en	as	Temporal	0.326	8	de	als	Temporal	0	7	2
en	when	Temporal	0.326	8	de	als	Temporal	0	7	4
en	so	Contingency	0	19	de	also	Contingency	0.321	7	5
en	also	Expansion	0.39	6	de	auch	Expansion	0	7	3
en	and	Expansion	0.112	124	de	auch	Expansion	0	7	1

Table 22 cont.

Source Language					Target Language					Alignment Frequency
Language	Connective Word	Level-I Sense	Relative Entropy	Frequency	Language	Connective Word	Level-I Sense	Relative Entropy	Frequency	
en	because	Contingency	0	15	de	da	Contingency	0	2	1
en	since	Contingency	0	1	de	da	Contingency	0	2	1
en	and	Expansion	0.112	124	de	dadurch	Expansion	0	1	1
en	then	Temporal	0	7	de	danach	Expansion	0	1	1
en	then	Temporal	0	7	de	dann	Temporal	0	1	1
en	because	Contingency	0	15	de	darum..weil	Contingency	0	1	1
en	because	Contingency	0	15	de	denn	Contingency	0	4	3
en	so	Contingency	0	19	de	deshalb	Contingency	0	4	2
en	thats why	Contingency	0	1	de	deshalb	Contingency	0	4	1
en	by	Expansion	0	5	de	indem	Expansion	0	2	1
en	through	Expansion	0	2	de	indem	Expansion	0	2	1
en	especially when	Expansion	0	1	de	insbesondere wenn	Expansion	0	1	1
en	in short	Expansion	0	1	de	kurz gesagt	Expansion	0	1	1
en	but indeed	Expansion	0	1	de	nicht nur..sondern	Expansion	0	1	1
en	or	Expansion	0	6	de	oder	Expansion	0	6	5
en	lastly	Expansion	0	1	de	schlussendlich	Expansion	0	1	1
en	so	Contingency	0	19	de	so dass	Contingency	0	1	1
en	and	Expansion	0.112	124	de	sodass	Contingency	0	1	1
en	and	Expansion	0.112	124	de	sogar	Expansion	0	2	1
en	but	Expansion	0.4	46	de	sondern	Comparison	0	2	2
en	in order	Contingency	0	2	de	um	Contingency	0	12	2
en	so that	Contingency	0	1	de	um	Contingency	0	12	1
en	also	Comparison	0.39	6	de	um..zu	Contingency	0	1	1
en	and	Comparison	0.112	124	de	und	Expansion	0	75	1

Table 22 cont.

Source Language					Target Language					Alignment Frequency
Language	Connective Word	Level-I Sense	Relative Entropy	Frequency	Language	Connective Word	Level-I Sense	Relative Entropy	Frequency	
en	and	Expansion	0.112	124	de	und	Expansion	0	75	58
en	and	Expansion	0.112	124	de	und gleichzeitig	Expansion	0	1	1
en	and	Expansion	0.112	124	de	und sogar	Expansion	0	1	1
en	in fact	Expansion	0	2	de	vielmehr	Expansion	0	1	1
en	as	Temporal	0.326	8	de	während	Temporal	0	2	1
en	while	Temporal	0	1	de	während	Temporal	0	2	1
en	because	Contingency	0	15	de	weil	Contingency	0	13	9
en	for	Contingency	0	1	de	weil	Contingency	0	13	1
en	if	Contingency	0	17	de	wenn	Contingency	0.24	28	13
en	if..if	Contingency	0	1	de	wenn	Contingency	0.24	28	1
en	if..if..if	Contingency	0	1	de	wenn	Contingency	0.24	28	1
en	so	Contingency	0	19	de	wenn	Contingency	0.24	28	1
en	when	Temporal	0.326	8	de	wenn	Contingency	0.24	28	2
en	especially when	Expansion	0	1	lt	ar	Contingency	0.246	12	1
en	or	Expansion	0	6	lt	ar	Expansion	0.246	12	3
en	or	Expansion	0	6	lt	arba	Expansion	0	3	3
en	so	Contingency	0	19	lt	argi	Contingency	0	1	1
en	but	Comparison	0.4	46	lt	bet	Comparison	0.48	42	29
en	but	Comparison	0.4	46	lt	bet	Expansion	0.48	42	3
en	but	Expansion	0.4	46	lt	bet	Comparison	0.48	42	1
en	but	Expansion	0.4	46	lt	bet	Expansion	0.48	42	4
en	on the one hand..but	Comparison	0	1	lt	bet	Comparison	0.48	42	1
en	however	Comparison	0	1	lt	deja	Comparison	0	1	1
en	and	Expansion	0.112	124	lt	dėl kurio	Expansion	0	1	1

Table 22 cont.

Source Language					Target Language					Alignment Frequency
Language	Connective Word	Level-I Sense	Relative Entropy	Frequency	Language	Connective Word	Level-I Sense	Relative Entropy	Frequency	
en	so	Contingency	0	19	lt	dėl to	Contingency	0	1	1
en	because	Contingency	0	15	lt	dėl to kad	Contingency	0	1	1
en	also	Expansion	0.39	6	lt	ir	Expansion	0.119	77	3
en	and	Expansion	0.112	124	lt	ir	Comparison	0.119	77	1
en	and	Expansion	0.112	124	lt	ir	Expansion	0.119	77	56
en	but	Expansion	0.4	46	lt	ir	Expansion	0.119	77	1
en	rather..than	Expansion	0	1	lt	ir	Expansion	0.119	77	1
en	so	Contingency	0	19	lt	ir	Expansion	0.119	77	1
en	lastly	Expansion	0	1	lt	ir pabaigai	Expansion	0	1	1
en	if	Contingency	0	17	lt	jei	Contingency	0	15	7
en	if..if	Contingency	0	1	lt	jei	Contingency	0	15	1
en	if..if..if	Contingency	0	1	lt	jei	Contingency	0	15	1
en	if	Contingency	0	17	lt	jei tik	Contingency	0	1	1
en	if	Contingency	0	17	lt	jeigu	Contingency	0	9	7
en	with	Expansion	0	1	lt	ką	Expansion	0	6	1
en	and	Expansion	0.112	124	lt	kad	Contingency	0.499	82	2
en	and	Expansion	0.112	124	lt	kad	Expansion	0.499	82	3
en	but indeed	Expansion	0	1	lt	kad	Expansion	0.499	82	1
en	in order	Contingency	0	2	lt	kad	Contingency	0.499	82	1
en	in order	Contingency	0	2	lt	kad	Expansion	0.499	82	1
en	so	Contingency	0	19	lt	kad	Expansion	0.499	82	1
en	so that	Contingency	0	1	lt	kad	Contingency	0.499	82	1
en	though	Expansion	0.6	2	lt	kad	Expansion	0.499	82	1
en	when	Temporal	0.326	8	lt	kad	Expansion	0.499	82	1

Table 22 cont.

Source Language					Target Language					Alignment Frequency
Language	Connective Word	Level-I Sense	Relative Entropy	Frequency	Language	Connective Word	Level-I Sense	Relative Entropy	Frequency	
en	since	Contingency	0	1	lt	kadangi	Contingency	0	1	1
en	as	Temporal	0.326	8	lt	kai	Temporal	0.184	18	1
en	through	Expansion	0	2	lt	kai	Temporal	0.184	18	1
en	when	Temporal	0.326	8	lt	kai	Temporal	0.184	18	3
en	while	Temporal	0	1	lt	kai	Temporal	0.184	18	1
en	and	Expansion	0.112	124	lt	kaip	Expansion	0.246	12	1
en	as	Temporal	0.326	8	lt	kaip	Temporal	0.246	12	1
en	though	Comparison	0.6	2	lt	kaip	Expansion	0.246	12	1
en	as	Comparison	0.326	8	lt	kaip kad	Comparison	0	1	1
en	in short	Expansion	0	1	lt	kitaip tariant	Expansion	0	1	1
en	until	Contingency	0	1	lt	kol..tol	Contingency	0	1	1
en	as	Temporal	0.326	8	lt	kuomet	Temporal	0.595	2	1
en	in	Expansion	0	1	lt	negu	Comparison	0	2	1
en	because	Contingency	0	15	lt	nes	Contingency	0	18	12
en	for	Contingency	0	1	lt	nes	Contingency	0	18	1
en	but	Comparison	0.4	46	lt	nors	Comparison	0	1	1
en	and	Comparison	0.112	124	lt	o	Comparison	0.828	35	1
en	and	Expansion	0.112	124	lt	o	Comparison	0.828	35	5
en	and	Expansion	0.112	124	lt	o	Contingency	0.828	35	1
en	and	Expansion	0.112	124	lt	o	Expansion	0.828	35	13
en	but	Comparison	0.4	46	lt	o	Comparison	0.828	35	2
en	then	Temporal	0	7	lt	po to	Temporal	0	1	1
en	but	Comparison	0.4	46	lt	tačiau	Comparison	0	2	2
en	and	Expansion	0.112	124	lt	tad	Contingency	0	7	1

Table 22 cont.

Source Language					Target Language					Alignment Frequency
Language	Connective Word	Level-I Sense	Relative Entropy	Frequency	Language	Connective Word	Level-I Sense	Relative Entropy	Frequency	
en	so	Contingency	0	19	lt	tad	Contingency	0	7	2
en	then	Temporal	0	7	lt	tada	Temporal	0	4	3
en	so	Contingency	0	19	lt	taigi	Contingency	0	6	4
en	also	Expansion	0.39	6	lt	taip pat	Expansion	0	2	2
en	then	Temporal	0	7	lt	tik tada	Temporal	0	1	1
en	so	Contingency	0	19	lt	todėl	Contingency	0	9	3
en	thats why	Contingency	0	1	lt	todėl	Contingency	0	9	1
en	because	Contingency	0	15	lt	todėl kad	Contingency	0	2	2
en	at the same time	Temporal	0	1	lt	tuo pačiu	Temporal	0	1	1
en	and	Expansion	0.112	124	pl	a	Comparison	0.628	23	7
en	and	Expansion	0.112	124	pl	a	Expansion	0.628	23	8
en	as	Temporal	0.326	8	pl	a	Temporal	0.628	23	1
en	but	Comparison	0.4	46	pl	a	Comparison	0.628	23	2
en	and	Expansion	0.112	124	pl	a nie	Comparison	0	1	1
en	then	Temporal	0	7	pl	a po tym	Temporal	0	1	1
en	then	Temporal	0	7	pl	a potem	Temporal	0	1	1
en	also	Expansion	0.39	6	pl	a także	Expansion	0	1	1
en	in fact	Expansion	0	2	pl	aby	Contingency	0	8	1
en	so	Contingency	0	19	pl	aby	Contingency	0	8	1
en	so that	Contingency	0	1	pl	aby	Contingency	0	8	1
en	or	Expansion	0	6	pl	albo	Expansion	0	2	2
en	but	Comparison	0.4	46	pl	ale	Comparison	0.236	32	21
en	but	Comparison	0.4	46	pl	ale	Expansion	0.236	32	1
en	but	Expansion	0.4	46	pl	ale	Comparison	0.236	32	1

Table 22 cont.

Source Language					Target Language					Alignment Frequency
Language	Connective Word	Level-I Sense	Relative Entropy	Frequency	Language	Connective Word	Level-I Sense	Relative Entropy	Frequency	
en	but	Expansion	0.4	46	pl	ale	Expansion	0.236	32	2
en	so	Contingency	0	19	pl	ale	Comparison	0.236	32	1
en	but indeed	Expansion	0	1	pl	ale i	Temporal	0	1	1
en	and	Expansion	0.112	124	pl	bo	Contingency	0	11	1
en	because	Contingency	0	15	pl	bo	Contingency	0	11	8
en	and	Expansion	0.112	124	pl	by	Contingency	0	5	1
en	so	Contingency	0	19	pl	czyli	Expansion	0	4	1
en	thats why	Contingency	0	1	pl	dlatego	Contingency	0	3	1
en	until	Contingency	0	1	pl	dopóki	Temporal	0	1	1
en	because	Contingency	0	15	pl	gdyż	Contingency	0	2	2
en	and	Contingency	0.112	124	pl	i	Temporal	0.707	31	1
en	and	Expansion	0.112	124	pl	i	Expansion	0.707	31	18
en	and	Expansion	0.112	124	pl	i	Temporal	0.707	31	3
en	as	Temporal	0.326	8	pl	i	Temporal	0.707	31	1
en	so	Contingency	0	19	pl	i	Temporal	0.707	31	1
en	where	Comparison	0	1	pl	i	Temporal	0.707	31	1
en	then	Temporal	0	7	pl	i potem	Temporal	0	1	1
en	as	Comparison	0.326	8	pl	jak	Comparison	0.484	6	1
en	but	Comparison	0.4	46	pl	jak	Comparison	0.484	6	1
en	for	Contingency	0	1	pl	jak	Expansion	0.484	6	1
en	and	Expansion	0.112	124	pl	jakby	Comparison	0	1	1
en	and	Expansion	0.112	124	pl	jednak	Comparison	0.312	7	1
en	but	Comparison	0.4	46	pl	jednak	Comparison	0.312	7	3
en	but	Comparison	0.4	46	pl	jednak	Expansion	0.312	7	1

Table 22 cont.

Source Language					Target Language					Alignment Frequency
Language	Connective Word	Level-I Sense	Relative Entropy	Frequency	Language	Connective Word	Level-I Sense	Relative Entropy	Frequency	
en	however	Comparison	0	1	pl	jednak	Comparison	0.312	7	1
en	on the one hand..but	Comparison	0	1	pl	jednak	Comparison	0.312	7	1
en	if	Contingency	0	17	pl	jeśli	Contingency	0	9	6
en	if..if	Contingency	0	1	pl	jeśli	Contingency	0	9	1
en	if..if..if	Contingency	0	1	pl	jeśli	Contingency	0	9	1
en	if	Contingency	0	17	pl	jeśli..to	Contingency	0	3	3
en	if	Contingency	0	17	pl	jeżeli	Contingency	0	3	3
en	as	Temporal	0.326	8	pl	kiedy	Temporal	0.715	10	1
en	when	Comparison	0.326	8	pl	kiedy	Comparison	0.715	10	1
en	when	Temporal	0.326	8	pl	kiedy	Expansion	0.715	10	1
en	when	Temporal	0.326	8	pl	kiedy	Temporal	0.715	10	4
en	while	Temporal	0	1	pl	kiedy	Temporal	0.715	10	1
en	and	Expansion	0.112	124	pl	który	Contingency	0	1	1
en	because	Contingency	0	15	pl	nie dlatego..że	Contingency	0	1	1
en	but	Expansion	0.4	46	pl	nie tyle..ile	Expansion	0	1	1
en	and	Expansion	0.112	124	pl	oraz	Expansion	0	1	1
en	and	Expansion	0.112	124	pl	podobnie jak	Comparison	0	1	1
en	because	Contingency	0	15	pl	ponieważ	Contingency	0	3	2
en	since	Contingency	0	1	pl	ponieważ	Contingency	0	3	1
en	then	Temporal	0	7	pl	potem	Temporal	0	2	2
en	if	Contingency	0	17	pl	skoro	Contingency	0	2	2
en	so	Contingency	0	19	pl	tak że	Contingency	0	1	1
en	and	Expansion	0.112	124	pl	to	Contingency	0.527	2	1
en	because	Contingency	0	15	pl	to że	Contingency	0	1	1

Table 22 cont.

Source Language					Target Language					Alignment Frequency
Language	Connective Word	Level-I Sense	Relative Entropy	Frequency	Language	Connective Word	Level-I Sense	Relative Entropy	Frequency	
en	and	Expansion	0.112	124	pl	w porównaniu z	Comparison	0	1	1
en	in short	Expansion	0	1	pl	w skrócie	Expansion	0	1	1
en	and	Expansion	0.112	124	pl	więc	Contingency	0.206	13	1
en	by	Expansion	0	5	pl	więc	Expansion	0.206	13	1
en	so	Contingency	0	19	pl	więc	Contingency	0.206	13	6
en	as	Temporal	0.326	8	pl	wtedy	Temporal	0	2	1
en	and	Expansion	0.112	124	pl	z kolei	Comparison	0	1	1
en	and	Expansion	0.112	124	pl	żeby	Contingency	0	5	1
en	in order	Contingency	0	2	pl	żeby	Contingency	0	5	2
en	as	Temporal	0.326	8	pt	a o	Temporal	0.494	3	1
en	by	Expansion	0	5	pt	a o	Temporal	0.494	3	1
en	at the same time	Temporal	0	1	pt	a o mesmo tempo	Temporal	0	1	1
en	and	Expansion	0.112	124	pt	ainda por cima	Expansion	0	1	1
en	so	Contingency	0	19	pt	assim	Contingency	0	3	2
en	until	Contingency	0	1	pt	até	Temporal	0	1	1
en	through	Expansion	0	2	pt	através de	Expansion	0	2	1
en	since	Contingency	0	1	pt	como	Contingency	0	1	1
en	so	Contingency	0	19	pt	de modo a	Contingency	0	1	1
en	rather..than	Expansion	0	1	pt	de o que	Expansion	0	1	1
en	then	Temporal	0	7	pt	depois	Temporal	0	5	5
en	also	Expansion	0.39	6	pt	e	Expansion	0.083	90	1
en	and	Comparison	0.112	124	pt	e	Comparison	0.083	90	1
en	and	Contingency	0.112	124	pt	e	Expansion	0.083	90	1
en	and	Expansion	0.112	124	pt	e	Comparison	0.083	90	1

Table 22 cont.

Source Language					Target Language					Alignment Frequency
Language	Connective Word	Level-I Sense	Relative Entropy	Frequency	Language	Connective Word	Level-I Sense	Relative Entropy	Frequency	
en	and	Expansion	0.112	124	pt	e	Expansion	0.083	90	69
en	because	Contingency	0	15	pt	e	Expansion	0.083	90	1
en	lastly	Expansion	0	1	pt	e	Expansion	0.083	90	1
en	as	Temporal	0.326	8	pt	enquanto	Temporal	0.388	5	4
en	while	Temporal	0	1	pt	enquanto	Comparison	0.388	5	1
en	so	Contingency	0	19	pt	então	Contingency	0.538	4	2
en	then	Temporal	0	7	pt	então	Temporal	0.538	4	1
en	especially when	Expansion	0	1	pt	especialmente	Expansion	0	1	1
en	but	Comparison	0.4	46	pt	mas	Comparison	0.279	43	34
en	but	Comparison	0.4	46	pt	mas	Expansion	0.279	43	2
en	but	Expansion	0.4	46	pt	mas	Comparison	0.279	43	1
en	but	Expansion	0.4	46	pt	mas	Expansion	0.279	43	2
en	so	Contingency	0	19	pt	mas	Comparison	0.279	43	1
en	though	Expansion	0.6	2	pt	mas	Comparison	0.279	43	1
en	in fact	Expansion	0	2	pt	na verdade	Expansion	0	2	2
en	but indeed	Expansion	0	1	pt	não só..mas..também	Expansion	0	1	1
en	however	Comparison	0	1	pt	no entanto	Comparison	0	1	1
en	or	Expansion	0	6	pt	ou	Expansion	0	5	5
en	and	Expansion	0.112	124	pt	para	Contingency	0	24	1
en	by	Expansion	0	5	pt	para	Contingency	0	24	1
en	in order	Contingency	0	2	pt	para	Contingency	0	24	2
en	so that	Contingency	0	1	pt	para que	Contingency	0	1	1
en	because	Contingency	0	15	pt	por	Contingency	0	4	1
en	for	Contingency	0	1	pt	por	Contingency	0	4	1

Table 22 cont.

Source Language					Target Language					Alignment Frequency
Language	Connective Word	Level-I Sense	Relative Entropy	Frequency	Language	Connective Word	Level-I Sense	Relative Entropy	Frequency	
en	so	Contingency	0	19	pt	por isso	Contingency	0	2	2
en	because	Contingency	0	15	pt	porque	Contingency	0	16	13
en	so	Contingency	0	19	pt	portanto	Contingency	0	2	2
en	as	Temporal	0.326	8	pt	quando	Temporal	0	16	2
en	when	Temporal	0.326	8	pt	quando	Temporal	0	16	6
en	in short	Expansion	0	1	pt	resumindo	Expansion	0	1	1
en	if	Contingency	0	17	pt	se	Contingency	0	20	14
en	if..if..if	Contingency	0	1	pt	se	Contingency	0	20	1
en	so	Contingency	0	19	pt	se	Contingency	0	20	1
en	if	Contingency	0	17	pt	se..então	Contingency	0	1	1
en	also	Expansion	0.39	6	pt	também	Expansion	0	3	1
en	and	Expansion	0.112	124	pt	também	Expansion	0	3	2
en	by	Expansion	0	5	tr	-(y)arak	Expansion	0.481	12	2
en	through	Expansion	0	2	tr	-(y)arak	Expansion	0.481	12	2
en	as	Temporal	0.326	8	tr	-(y)hnca	Temporal	0	1	1
en	by	Expansion	0	5	tr	-(y)hnca da	Temporal	0	1	1
en	and	Expansion	0.112	124	tr	-(y)hp	Expansion	0.163	21	10
en	by	Expansion	0	5	tr	-(y)hp	Expansion	0.163	21	1
en	as	Temporal	0.326	8	tr	-da	Temporal	0.541	9	1
en	when	Temporal	0.326	8	tr	-da	Temporal	0.541	9	4
en	as	Temporal	0.326	8	tr	-ken	Temporal	0.425	10	5
en	when	Comparison	0.326	8	tr	-ken	Comparison	0.425	10	1
en	while	Temporal	0	1	tr	-ken	Comparison	0.425	10	1
en	if	Contingency	0	17	tr	-sa	Contingency	0.276	30	14

Table 22 cont.

Source Language					Target Language					Alignment Frequency
Language	Connective Word	Level-I Sense	Relative Entropy	Frequency	Language	Connective Word	Level-I Sense	Relative Entropy	Frequency	
en	if	Contingency	0	17	tr	-sa	Expansion	0.276	30	1
en	if..if	Contingency	0	1	tr	-sa	Contingency	0.276	30	1
en	if..if..if	Contingency	0	1	tr	-sa	Contingency	0.276	30	1
en	rather..than	Expansion	0	1	tr	-sa	Expansion	0.276	30	1
en	so	Contingency	0	19	tr	-sa	Contingency	0.276	30	1
en	but	Comparison	0.4	46	tr	-sa da	Comparison	0	1	1
en	if	Contingency	0	17	tr	-sa..-sa	Contingency	0	1	1
en	but	Expansion	0.4	46	tr	aksine	Expansion	0	1	1
en	and	Expansion	0.112	124	tr	ama	Comparison	0.169	20	1
en	but	Comparison	0.4	46	tr	ama	Comparison	0.169	20	16
en	but	Expansion	0.4	46	tr	ama	Expansion	0.169	20	1
en	but	Comparison	0.4	46	tr	ancak	Comparison	0.348	14	10
en	but	Comparison	0.4	46	tr	ancak	Expansion	0.348	14	1
en	but	Expansion	0.4	46	tr	ancak	Expansion	0.348	14	1
en	however	Comparison	0	1	tr	ancak	Comparison	0.348	14	1
en	though	Expansion	0.6	2	tr	ancak	Comparison	0.348	14	1
en	in fact	Expansion	0	2	tr	aslında	Expansion	0	3	2
en	also	Expansion	0.39	6	tr	ayrıca	Expansion	0	1	1
en	on the one hand..but	Comparison	0	1	tr	bir tarafta..bir tarafta da	Comparison	0	1	1
en	so	Contingency	0	19	tr	böylece	Contingency	0	6	3
en	because	Contingency	0	15	tr	çünkü	Contingency	0	11	11
en	also	Expansion	0.39	6	tr	da	Expansion	0	2	1
en	and	Expansion	0.112	124	tr	da	Expansion	0	2	1
en	also	Expansion	0.39	6	tr	de	Expansion	0	2	1

Table 22 cont.

Source Language					Target Language					Alignment Frequency
Language	Connective Word	Level-I Sense	Relative Entropy	Frequency	Language	Connective Word	Level-I Sense	Relative Entropy	Frequency	
en	and	Expansion	0.112	124	tr	de	Expansion	0	2	1
en	so	Contingency	0	19	tr	dolayısıyla	Contingency	0	1	1
en	but	Comparison	0.4	46	tr	fakat	Comparison	0	3	3
en	and	Expansion	0.112	124	tr	hatta	Expansion	0	2	2
en	and	Expansion	0.112	124	tr	hem de	Expansion	0	1	1
en	and	Expansion	0.112	124	tr	için	Contingency	0.108	36	1
en	because	Contingency	0	15	tr	için	Contingency	0.108	36	2
en	in order	Contingency	0	2	tr	için	Contingency	0.108	36	2
en	since	Contingency	0	1	tr	için	Contingency	0.108	36	1
en	and	Expansion	0.112	124	tr	ise	Expansion	0.572	5	1
en	as	Comparison	0.326	8	tr	kadar	Comparison	0.589	2	1
en	until	Contingency	0	1	tr	kadar	Contingency	0.589	2	1
en	in short	Expansion	0	1	tr	kısacası	Expansion	0	1	1
en	especially when	Expansion	0	1	tr	özellikle de..gelince	Expansion	0	1	1
en	so	Contingency	0	19	tr	peki	Expansion	0	1	1
en	but	Expansion	0.4	46	tr	sadece	Expansion	0	1	1
en	and	Expansion	0.112	124	tr	sanki	Comparison	0	1	1
en	then	Temporal	0	7	tr	sonra da	Temporal	0	1	1
en	and	Comparison	0.112	124	tr	ve	Expansion	0.119	76	2
en	and	Contingency	0.112	124	tr	ve	Expansion	0.119	76	1
en	and	Expansion	0.112	124	tr	ve	Contingency	0.119	76	1
en	and	Expansion	0.112	124	tr	ve	Expansion	0.119	76	58
en	and	Expansion	0.112	124	tr	ve	Temporal	0.119	76	1
en	but	Comparison	0.4	46	tr	ve	Expansion	0.119	76	1

Table 22 cont.

Source Language					Target Language					Alignment Frequency
Language	Connective Word	Level-I Sense	Relative Entropy	Frequency	Language	Connective Word	Level-I Sense	Relative Entropy	Frequency	
en	and	Expansion	0.112	124	tr	ve de	Expansion	0	2	2
en	and	Expansion	0.112	124	tr	veya	Expansion	0	8	1
en	or	Expansion	0	6	tr	veya	Expansion	0	8	5
en	or	Expansion	0	6	tr	ya da	Expansion	0	1	1
en	so	Contingency	0	19	tr	yani	Contingency	0.478	4	1

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: ÖZER, Sibel **Nationality:** Turkish (TC)

EDUCATION

Degree	Institution	Year of Graduation
M.S.	METU, Informatics Institute Cognitive Science	2010
B.S.	Ege Üniversity, Computer Engineering	2002
High School	Burhaniye High School	1998

PROFESSIONAL EXPERIENCE

Year	Place	Enrollment
2002-Present	Havelsan A.Ş.	Software Engineer
2000-2 months	Aselsan A.Ş.	Intern

PUBLICATIONS

International Conference Publications

S. Özer and D. Zeyrek, "An automatic discourse relation alignment experiment on TED-MDB", in *Proceedings of the Workshop on Natural Language Processing at ACL*, 2019, pp. 31-34.

M. Kurfalı, S. Özer, D. Zeyrek, and A. Mendes, "TED-MDB Lexicons: Tr-EnConnLex, Pt-EnConnLex," in *Proceedings of the First Workshop on Computational Approaches to Discourse*, 2020, pp. 148-153.

S. Özer, M. Kurfalı, D. Zeyrek, A. Mendes, and G. V. Oleškevičienė, "Linking discourse-level information and the induction of bilingual discourse connective lexicons," *Semantic Web*, vol. 13, no. 6, pp. 1081-1102, 2022, published by IOS Press.

D. Zeyrek, A. Mendes, G. V. Oleškevičienė, and S. Özer, "An Exploratory Analysis of TED Talks in English and Lithuanian, Portuguese and Turkish Translations: Results from the Analysis of an Annotated Multilingual Corpus," *Contrastive Pragmatics*, vol. 3, no. 3, pp. 452-479, 2022, published by Brill.