Computational Discourse Analysis Across Complexity Levels

Elnaz Davoodi

A thesis

In the Department

of

Computer Science and Software Engineering

Presented in Partial Fulfillment of the Requirements

For the Degree of

Doctor of Philosophy (Computer Science) at

Concordia University

Montréal, Québec, Canada

August, 2017

Concordia University

School of Graduate Studies

This is to certify that the thesis prepared

By: **Elnaz Davoodi**

Entitled: **Computational Discourse Analysis Across Complexity Levels**

and submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy (Computer Science)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the final examining commitee:

Dr. Susan Liscouet-Hanke — Chair

Dr. Robert E. Mercer — External Examiner

Dr. Olga Ormandjieva — Examiner

Dr. Deborah Dysart-Gale — Examiner

Dr. Lata Narayanan — Examiner

Dr. Leila Kosseim — Thesis Supervisor

Approved Dr. Volker Haarslev
Chair of Department or Graduate Program Director

July 3, 2017

Dr. Amir Asif, Dean

Faculty of Engineering and Computer Science

Abstract

Computational Discourse Analysis Across Complexity Levels

Elnaz Davoodi, Ph.D.

Concordia University, 2017

The focus of this thesis is to study computationally the relation between discourse properties and textual complexity. Specifically, we explored three research questions.

The first research question tries to find out *to what degree discourse-level properties can be used to predict the complexity level of a text.* To do so, we considered three types of discourse-level properties: (1) the realization of discourse relations and the representation of discourse relations in terms of (2) the choice of discourse relation and (3) discourse marker. Using datasets from standard corpora in the field of discourse analysis and text simplification, we developed a supervised machine learning model for pairwise text complexity assessment and compared these properties with more linguistic features. Our results show that the use of only discourse features performed statistically as well as using traditional linguistic features. Thus, we can conclude a strong correlation between discourse properties and complexity level.

The second question that we explored is *how exactly does the complexity level of a text influence its discourse-level linguistic choices?* To address this question, we conducted a corpus analysis of the Simple English Wikipedia, the largest annotated corpus based on complexity level. Our analysis used the 16 discourse relations defined in the DLTAG framework and focused on *explicit* relations. Our results show that the distribution of discourse relations is not influenced by a text's complexity level; but *how* these are signalled is.

Finally, given the results of our corpus analysis, our third research question tries to *investigate if we can leverage these differences to mine parallel corpora across complexity levels to automatically discover alternative lexicalizations (AltLexes) of discourse markers.* This work led to the automatic identification of 91 new AltLexes in two corpora: the Simple English Wikipedia and the Newsela corpora.

Overall, this thesis demonstrates that a text's complexity level and discourse level properties are indeed correlated. Discourse properties play an important role in the assessment of a text's complexity level and should be taken into account in the complexity level assessment problem. In addition, we observed that the way that explicit discourse relations are signaled is influenced by textual complexity. Lastly, our thesis shows that the automatic identification of alternative lexializations of discourse markers can benefit from large-scale parallel corpora across complexity levels.

Acknowledgments

Writing this acknowledgment is my last touch on my dissertation. This part of my life was not all about intense learning of science, but also self improvement.

Foremost, I like to express my gratitude to my supervisor, Dr. Leila Kosseim, for her endless support throughout my Ph.D study and research, for her patience, attitude, motivation and enthusiasm. Despite her busy schedule, she always made some time for me.

I also like to thank the chair and the committee members of my defense, Dr. Susan Liscouet-Hanke, Dr. Robert E. Mercer (external), Dr. Olga Ormandjieva, Dr. Deborah Dysart-Gale and Dr. Lata Narayanan. I truly appreciate the time they put in reading my thesis, giving me their feedbacks and attending in my defense sessions. Their valuable comments gave me better insight into more possible future research directions and also improved the final version of this dissertation.

Last but not the least, I like to thank my dear family who support me spiritually throughout my life, and my friends who stood behind me in all ups and downs during my Ph.D study.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

A text is made of words, phrases and sentences that should not taken in isolation. The standard bag of words, bag of phrases or bag of sentences representation may be appropriate for some applications such as information retrieval or text classification, but is inadequate to properly model a discourse. Understanding a text goes beyond understanding its textual units in isolation; the relation between these units must also be understood. For example, consider the following:

Example 1

a. This dissertation is long. It took me 30 minutes to print it.

b. This dissertation is long. It took me 30 minutes to eat lunch.

Although most of the words and phrases are identical, the second passage leaves the reader wondering about the communicative purpose of the second utterance.

According to [WJ12], position, order, adjacency, and context are intrinsic features of a **discourse** that the "bag of" approach does not consider. [Jur00] defines a discourse as *"collocated, structured, coherent groups of sentences"*. Cohesion and

coherence are two fundamental properties of a discourse [Jur00]. **Cohesion** refers to the use of certain linguistic devices to tie together text segments. According to [HH76] these include referential devices, ellipsis, substitution, lexical cohesion as well as conjunction. For example, in Example 1.a, the second pronoun "*it*" refers to "*the dissertation*" and ties the two sentences into a cohesive unit. On the other hand, **coherence** refers to the logical or semantic relations between textual segments which allow the reader to understand the communicative goal of the writer. These relations are referred to as **discourse relations** (also known as **rhetorical relations** and **coherence relations**). For example, in 1.a the second sentence is related to the first by some kind of CAUSE relation. On the other, in 1.b the discourse relation between the two sentences is unclear, hence the reader may not comprehend the communicative goal of the writer; as a consequence 1.b cannot be considered a well-written discourse. This thesis focuses on the coherence properties of well-written discourse.

To create a coherent text, **discourse markers** such as *since*, *but*, etc. are often used to explicitly connect textual units and signal the presence of specific discourse relations such as CONTRAST, CAUSE, etc. These discourse markers are sufficient to signal a relation, but are not necessary. *Explicit* discourse relations are signalled using a discourse marker; while *non-explicit* relations are not signalled by a discourse marker but can still be inferred by the reader. A non-explicit relation can be indicated via an alternative lexicalization to the discourse marker (for example, *the reason for* instead of *because*) or via no lexical marker at all. These non-explicit realizations are referred to as *AltLex* and *implicit* relations respectively. The following examples illustrate these three types of realizations for the discourse relation CAUSE.

Example 2

a. I went to Concordia because I had a class. (*explicit*-CAUSE)

b. I went to Concordia; I had a class. (*implicit*-CAUSE)

c. I went to Concordia; the reason is that I had a class. (*AltLex*-CAUSE)

Previous work (e.g. [Web09, BDK14, PN08]) has shown a correlation between the use of discourse relations and certain textual dimensions, such as genre or level of formality. For example, [Web09] has shown that the distribution of discourse relations in the Penn Discourse Treebank (PDTB) corpus [PDL+08] is influenced by the textual genre; that is, texts from different genres tend to contain certain discourse relations more often than others. Text complexity is an important dimension of a text, that today has enjoyed a renewed interest as it allows online documents to be made more accessible to non-native speakers. However, very little research has investigated the correlation between text complexity and discourse-level properties. In the literature, text complexity is a notion that is very close to text readability [CT14]. [DuB04] offers a comprehensive review of research in readability prediction, which began in the early 20^{th} century. He points out that the complexity of a text is a function of two main factors: (1) the reader (i.e. the reader's literacy level, prior knowledge and interest in the materials), and (2) the text (i.e. text organization, coherence and design). Over the years, many research efforts have considered the influence of the reader and specific reader groups in text readability (e.g. [CMKP13, BT13, WJU+09, DU06, DT98a]). As opposed to this line of work, our thesis focuses on the influence of the text rather than the reader.

More recently, automatic text simplification methods have been developed to reduce linguistic complexities, while still retaining the original information and meaning of a text [Sid14]. In automatic text simplification, identifying the complexity level of a text is typically seen as an initial step. This process, which is referred to as **text complexity assessment**, is beneficial not only in automatic text simplification, but also in many other Natural Language Processing (NLP) tasks. For example, syntactic parsing can benefit from text simplification as syntactic parsers perform better on simpler texts compared to more complex ones [CS97]. Thus, the syntactic parses

of simpler texts are more reliable. In addition to natural language processing applications, human readers can also benefit from text complexity assessment as it can allow them to identify texts that are appropriate to their literacy level [WRO03]. Research in assessing text complexity considers various aspects of a text; however, to our knowledge, the influence of discourse properties on text complexity assessment has been understudied. **The goal of this thesis is to investigate the relation between a text's discourse properties and its complexity level.**

The general methodology followed by our work is data-driven. Similarly to many NLP applications today, the availability of large-scale annotated corpora allowed us to use statistical approaches and conduct reliable data-driven research to study computational discourse analysis across complexity levels. In particular, we used the Simple English Wikipedia [CK11b] corpus and the Newsela [XCBN15] corpus for our work. These corpora constitute the largest publicly available corpora across complexity levels and constitute standard benchmarks in the field. Because the meaning is preserved across parallel corpora at different complexity levels and discourse relations are semantic in nature, we can therefore assume that discourse relations are preserved across complexity levels. However, the realization and representation of discourse relations may change. Using large-scale parallel corpora at different complexity levels, we can therefore study the correlation between discourse-level properties and text complexity levels, and in particular the influence of the complexity level on the realization (i.e. *explicit*, *implicit*, *AltLex*) and representation (e.g. change in the choice of discourse marker, change of discourse marker to an AltLex, etc.) of discourse relations.

1.2 Research Questions

The focus of this thesis is to study the computational discourse processing across complexity levels. Specifically, we addressed three main research questions:

1. *What is the influence of discourse-level properties on text complexity assessment?*

   According to previous work on text complexity assessment, various textual aspects (e.g. lexical, syntactic, etc.) can influence the complexity of a text. In early work in computational complexity assessment, many cognitive theorists and linguists pointed out the importance of coherence in text complexity assessment [DuB04]; however, to our knowledge, this problem has not been studied computationally using a formal discourse theory. This research question is addressed in Chapter 3.

2. *What is the influence of textual complexity on discourse-level linguistic choices?*

   According to previous work on the correlation between discourse structure and other textual dimensions (e.g. [Web09, BDK14, DKB$^+$16]), we suspected that discourse-level linguistic choices might be correlated to textual complexity. To our knowledge, the most relevant previous effort to answer this question was conducted on a small-scale corpus containing only three parallel texts, each parallel text comprising fewer than 1000 words [WRO03]. As opposed to this previous work, we conducted a corpus analysis of the largest publicly available corpus in text simplification, the Simple English Wikipedia corpus [CK11a]. This corpus contains more 60K parallel articles across complexity levels. This research question is addressed in Chapter 4.

3. *Can parallel corpora across complexity levels be used to automatically extract alternative lexicalizations of discourse markers?*

   In parallel corpora across complexity levels, the meaning of sentences is assumed to be preserved across complexity levels. Because discourse relations are semantic in nature, we can also assume that they should be preserved across complexity levels. However, their realization and representation may change. Using these observations, our third research question explored how alternative lexicalization of discourse markers can be automatically extracted from parallel corpora across complexity levels. This research question is addressed in Chapter 5.

1.3 Contributions

The primary focus of this thesis is to study computational discourse processing across complexity levels. As described in Section 1.2, we focused on three main research questions. Each research question brought about contributions to the field (our primary contributions). In addition, through the exploratory nature of our work, we also contributed to the field with secondary contributions.

1.3.1 Primary Contributions

Figure 1 shows the main components of this thesis along with the contributions of each component.

Primary Contribution #1: Text Complexity Assessment

Our first contribution tried to answer research question #1 (see Section 1.2). This work studied the influence of discourse-level properties in the task of text complexity assessment. Our results show the discriminating power of discourse-level properties

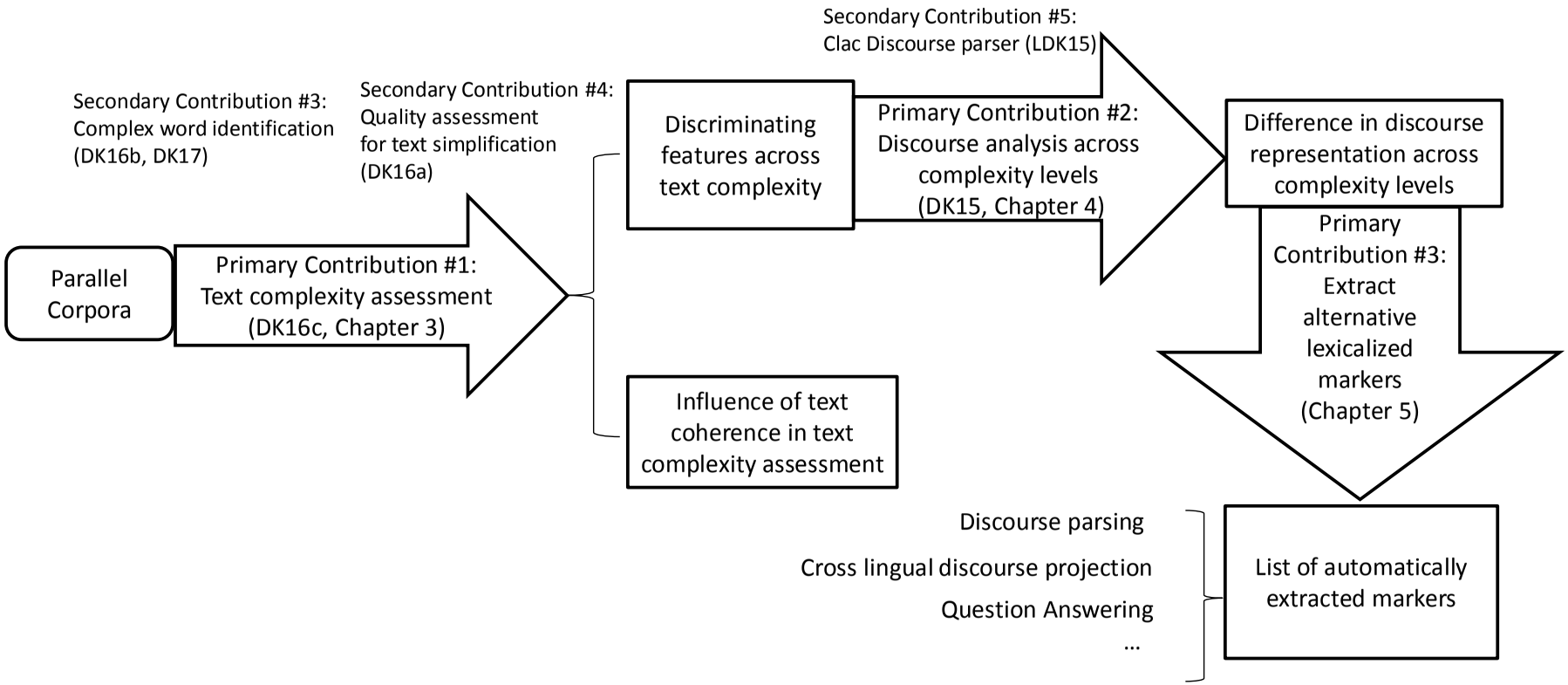


Figure 1: Primary and secondary contributions of this thesis.

(more specifically, coherence properties) for the task of pairwise complexity assessment. This work was published in [DK16c] and is described in Chapter 3.

Primary Contribution #2: Discourse Analysis Across Complexity Levels

This corpus analysis work addressed our second research question. We analysed the influence of the complexity level on discourse-level linguistic choices across the Simple English Wikipedia corpus [CK11a]. Our results show that the way explicit relations are signalled in a text is influenced by its complexity level. This work was published in [DK15] and is described in Chapter 4.

Primary Contribution #3: Automatic Extraction of Alternative Lexicalizations

We have proposed an approach to automatically extract alternative lexicalizations of discourse markers, which can signal a discourse relation. This approach leverages the strength of the End-to-End discourse parser [LNK14] in identifying explicit relations and monolingual corpora used in text simplification. The results show that the use of external resources and monolingual parallel corpora across different complexity levels can be used in the automatic identification of alternative lexicalizations of discourse markers. This work was published at [DK17b] and is described in Chapter 5.

1.3.2 Secondary Contributions

Our secondary contributions are not directly described in this thesis; however, we mention them here as their methods, resources and results indirectly influenced this thesis. Each contribution below has been published in regarded NLP venues.

Secondary Contribution #1: Corpus study of the influence of discourse-level properties across textual genres

A corpus study on the influence of textual genre on discourse structure was performed on various corpora across textual genres: RST Discourse Treebank [CMO03], Maite Taboada's Review Corpus [TAV06, TG04], the Penn Discourse Treebank [PDL+08] and Biomedical Discourse Relation Bank (BioDRB) [PMF+11]. Results show that textual genre influences the distribution of discourse relations. This contribution was published in [BDK14].

Secondary Contribution #2: An analysis of the influence of discourse-level properties in textual genre classification

We experimented with several supervised machine learning models to evaluate the influence of discourse-level properties on textual genre classification. In order to compare the discriminative power of various features in textual genre classification tasks, we used different discourse-level features including discourse markers and discourse relations and the Bag of Word model (BoW). Results show that the distribution of discourse markers are strong indicators of textual genres. This work was published in [DKB+16].

Secondary Contribution #3: Complex Word Identification

To better understand the field of text complexity assessment, we participated in the 2016 NAACL-SemEval international shared task on complex word identification. The results show that context information can be an indicator of complexity level of words. This work was published in [DK16b, DK17a]

Secondary Contribution #4: Quality Assessment for Text Simplification

Another related contribution to text complexity assessment was assessing the quality of text simplification. To do so, we participated in the 2016 LREC-QATS international shared task [DK16a]. Four aspects of textual quality were studied including

grammatical correctness, level of simplification, meaning preservation and the overall quality of the simplification. This work provided us with a goldstandard dataset of automatically simplified texts manually labeled with four aspects of textual quality and allowed us to better understand the notion of meaning preservation in parallel corpora. This contribution was published in [DK16a].

Secondary Contribution #5: Development of the CLaC Discourse Parser

As part of our participation to the 2015 CoNLL international shared task on shallow discourse parsing, we contributed to the development of the CLaC discourse parser based on the Penn Discourse Treebank corpus [PDL+08] to automatically identify discourse relations. This contribution was published in [LDK15].

1.4 Overview of the Thesis

This thesis is organized in six chapters:

- **Chapter 2** discusses background helpful to follow the rest of the thesis. The chapter presents a brief summary of computational discourse processing and an overview of the two main discourse theories: Rhetorical Structure Theory (RST) [MT87] and Discourse Lexicalized Tree Adjoining Grammar (DLTAG) [WJ98, WKSJ99, WSJK03, Web04]. In addition, the two largest corpora annotated based on these two discourse theories are described: the Rhetorical Structure Theory Discourse Treebank (RST-DT) [CMO03] and the Penn Discourse Treebank (PDTB) [PDL+08]. In the second part of the chapter, we explain the notion of text complexity and the efforts towards its assessment as well as the two widely used annotated corpora based on complexity levels: the Simple English Wikipedia corpus [CK11b] and the Newsela corpus [XCBN15]. In Chapter 2, we also describe different aspects of a text that may influence its

complexity level. We discuss the limitations of current approaches to computational text complexity assessment and introduce the notion of pairwise text complexity assessment.

- In **Chapter 3**, our contribution to the area of pairwise text complexity assessment is presented. We used a set of surface and linguistic features to build a supervised model for the task of pairwise text complexity assessment. Our results show the discriminating power of discourse-level properties (more specifically, coherence properties) for the task of pairwise complexity assessment.

- In **Chapter 4**, we investigate the influence of text complexity on discourse-level choices. We use the Simple English Wikipedia corpus [CK11a] to conduct this data-driven analysis. We study three discourse-level linguistic choices: (1) the usage of discourse relations, (2) the usage of discourse markers and (3) the distribution of discourse markers signaling explicit discourse relations. Our results show that the way explicit relations are signalled in a text is influenced by its complexity level.

- In **Chapter 5**, based on our observations of Chapter 4, we propose an approach to identify alternative lexicalizations of discourse markers, using the Simple English Wikipedia corpus [CK11a] and the Newsela corpus [XCBN15]. This approach leads us to the automatic identification of alternative lexicalizations to signal a discourse relation. The results show that the use of external resources and monolingual parallel corpora across different complexity levels can be used in automatic identification of alternative lexicalizations of discourse markers.

- Finally, **Chapter 6** summarizes the thesis and discusses a number of future directions.

Overall, this thesis demonstrates that a text's complexity level and discourse level properties are indeed correlated. Specifically, discourse properties play an important

role in the assessment of a text's complexity level and should be taken into account in the complexity level assessment problem. In addition, we observed that the way that explicit discourse relations are signaled is influenced by textual complexity level. Lastly, our thesis shows that the automatic identification of alternative lexializations of discourse markers can benefit from large-scale parallel corpora across complexity levels.

Chapter 2

Background

In the last few years, computational discourse analysis has received much attention in Natural Language Processing (NLP). Lower-level tasks such as part-of-speech tagging and syntactic parsing have reached performances that have allowed researchers to turn their attention to higher level tasks such as: semantic and discourse analysis. In addition, with the availability of large-scale corpora, the use of robust data-driven approaches to computational discourse processing became possible. Discourse parsing is a clear example of an NLP application that has benefited directly from advances in computational discourse processing (e.g. [LNK14, LDK15, HPAdI10]). However, many other NLP applications, such as Machine Translation (MT) (e.g. [FIK10]), Text Simplification (e.g. [Sid06]) and Text Summarization (e.g. [Mar00]) have also benefited from computational discourse processing.

Text complexity is an important textual dimension, that has enjoyed a renewed interest recently, in particular to make online documents more accessible to non-native speakers (e.g. [Sid14, DEO14, Eva11, BRDS12, YPDNML10]). Computational complexity assessment is another field which has influenced other NLP applications such as Text Simplification [Sid06, Sid14, DT98a, CMC+98a, BBE11, CS97, Kau13] and Syntactic Parsing [CS97]).

As noted in Chapter 1, the focus of our thesis is to study computational discourse analysis across complexity levels. In order to better appreciate the rest of the thesis, this chapter first briefly describes background information on computational discourse processing (see Section 2.1): in Sections 2.1.1 and 2.1.3, we review two of the most widely used discourse theories and in Sections 2.1.2 and 2.1.4, we discuss the two largest-scale annotated corpora based on these discourse theories. Then in Section 2.2 we introduce work in computational text complexity: in Section 2.2.1, an overview of the importance of text complexity assessment is provided, while in Section 2.2.2 we discuss recent efforts to define and evaluate text complexity. Finally, in Section 2.2.3 we describe the two widely used large-scale annotated corpora based on textual complexity levels.

2.1 Discourse Analysis

Considering a text only using lexical, syntactic or semantic information in isolation leads us to treat it as a bag of words, a bag of constituents, or a bag of sentences. Thus, the semantic relations between utterances are ignored. As [WJ12] pointed out, in many NLP applications (e.g. summarization [TVdBPC04], information extraction [PR07, ESR08, MC07], machine translation [FIK10], automatic assessment of students' essays [BC10], biomedical document segmentation [HM14], etc.), this representation is not sufficient. We need to study texts using a higher level representation: the discourse level.

Early work in NLP has taken discourse into account (e.g. [Woo68, Woo78, Win73]), but only ad hoc methods were used then. The need for formal and computational methods to handle discourse information (i.e. cohesion and coherence) led to the introduction of computational discourse processing methods. These computational methods model cohesion (such as centering theory [GS86, BFP87, WJP98, Tet01]),

and also model coherence (i.e. discourse theories [MT87, WJ98, WKSJ99, WSJK03, Web04], see Section 2.1).

In this thesis, we study the coherence of discourse, thus the rest of this chapter will only present background on the coherence aspect of discourse.

Several computational discourse frameworks have been proposed to model the discourse structure of a text formally through a **computational discourse framework**. The most notable efforts include: [HH76] who modeled computationally cohesive devices in discourse; [McK85] who expanded the discourse relations introduced by [Gri75] and [Wil90] to generate coherent texts; [MT87] who proposed Rhetorical Structure Theory (RST), which defines a coherent discourse as a collection of semantically connected text segments; [PS84, SP88] who proposed the Linguistic Discourse Model (LDM) and more recently, [WJ98, WKSJ99, WSJK03, Web04] who developed Discourse Lexicalized Tree-Adjoining Grammar (DLTAG) to model coherence using the LTAG grammar.

In addition, in order to facilitate experiments with these discourse frameworks, **discourse annotated corpora** have been developed. The most notable ones include the RST Discourse Treebank (RST-DT) [CMO03] corpus and the Penn Discourse Treebank (PDTB) [PDL$^+$08] corpus which now constitute standard corpora in the field of computational discourse analysis.

Discourse parsers are computational models that aim to automatically identify the discourse structure of a text. Discourse parsers are built based on a specific discourse theory and discourse annotated corpora. Due to the availability of manually annotated corpora based on RST and DLTAG, a number of discourse parsers have been developed. For example, SPADE (Sentence-level PArsing for DiscoursE) [SM03], which identifies intra-sentence discourse relations, HILDA (HIgh-Level Discourse Analyzer) [HPAdI10] and the RST-style text-level discourse parser of [FH12], are based

on the RST-DT corpus. On the other hand, [LNK14] developed a PDTB-Style End-to-End Discourse Parser, which is based on the PDTB corpus.

In this thesis, we use the PDTB-Style End-to-End Discourse Parser [LNK14], which constitutes the state of the art discourse parser. Since a number of discourse parsers have been developed based on RST, this chapter will review the RST discourse theory in Section 2.1.1 and will present the RST-DT corpus, in Section 2.1.2. We will then review the DLTAG discourse theory in Section 2.1.3 and will describe the PDTB in Section 2.1.4.

2.1.1 Rhetorical Structure Theory

Rhetorical Structure Theory (RST) [MT87] is a framework which models the semantic (rhetorical) relations between text spans. In RST, a text span is defined as a single or multiple clauses that are connected to each other using a rhetorical relation. A rhetorical relation is defined as a semantic relation which relates non-overlapping text spans. In RST, at least one text span is a nucleus and the others can be either nuclei or satellites. A nucleus text span is crucial in a relation as it carries the main meaning of the relation and the meaning of the satellite cannot be understood without the nucleus. On the other hand, the nucleus can still be understood if the satellite is removed. A crucial point in RST is that relations exist between non-overlapping adjacent text spans and these relations can be nested. Hence, a complete discourse annotation of a text using RST forms a hierarchical tree structure over the text. In this tree structure each relation holds between two or more adjacent text spans where each text span is formed by smaller text spans connected to each other using a rhetorical relation. The relations in RST are semantic in nature and are defined based on the judgement of the writer and the reader. In this framework, it is assumed that the judgements of the reader's comprehension of a discourse relation are made on the basis of the text. No assumption is made about the reader's characteristics.

Thus, the relations are defined based on the intention of the writer having a fixed reader in mind. Relations are divided into two groups of:

- **Mononuclear**: Relations that contain only one nucleus span and the other text spans are satellites (also known as nuclear-satellite relation). Figure 2 shows an example of the ELABORATION relation which holds between span 1 and span 2. In this example, text span 1 is the nucleus of the relation[1].

- **Multinuclear**: Relations where the main purpose of the writer is distributed across more than one text segment. In a multinuclear relation no single text segment is more central than another in the relation, but rather there are multiple such text spans. In this case, the removal of any of the nuclei would make the relation and as a result the entire text span, meaningless. Figure 3 shows an example of the multinuclear SEQUENCE relation. In this example, any two consecutive adjacent text spans are related to each other using a SEQUENCE relation[2].

  In principle, the set of rhetorical relations in RST is open; however [MT87] proposed a list of 78 relations consisting of 53 mononuclear and 25 multinuclear relations. These 78 discourse relations are grouped into 16 classes of relations based on the semantic of the relations. Figure 4 shows the 16 classes of RST relations.

2.1.2 RST Discourse Treebank

The RST Discourse Treebank (RST-DT) [CMO03] corpus was built a number of years later to facilitate comparative work with the RST framework. The corpus, consists of 385 Wall Street Journal articles which range over a variety of topics,

[1]This example is taken from [MT87].
[2]This example is taken from [MT87].

Example of the ELABORATION relation in RST

1. One difficulty is with sleeping bags in which down and feather fillers are used as insulation.

2. This insulation has a tendency to slip towards the bottom.

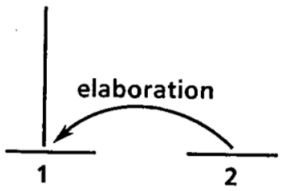


Figure 2: An example of a nucleus-satellite relation in RST. There is an ELABORATION discourse relation between text spans 1 and 2.

Example of the SEQUENCE relation in RST

1. Peel oranges

2. and slice crosswise.

3. Arrange in a bowl

4. and sprinkle with rum and coconut.

5. Chill until ready to serve.

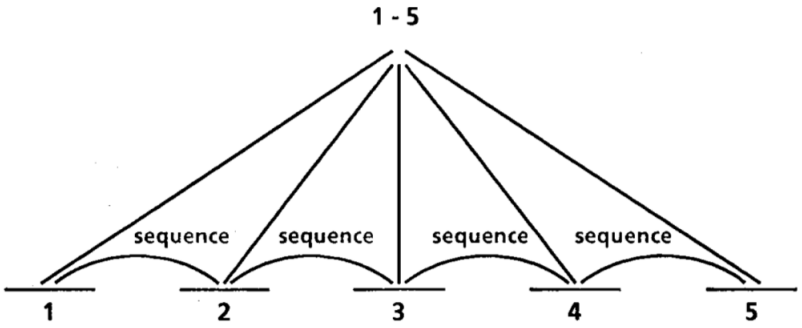


Figure 3: An example of a multinuclear relation in RST. There is SEQUENCE discourse relation between text spans 1-5.

Classes of the RST-DT Discourse Relations	
Attribution	Background
Cause	Comparison
Condition	Contrast
Elaboration	Enablement
Evaluation	Explanation
Joint	Manner-Means
Topic-Comment	Summary
Temporal	Topic-Change

Figure 4: Classes of discourse relations in RST [CMO03].

including financial reports, general interest stories, business-related news, cultural reviews, editorials, letters, etc. from the Penn Treebank [MMS93]. Each article has been manually annotated at the discourse level based on the RST discourse framework. In total, it contains over 176,000 words. The RST-DT is divided into two sets: one for training and one for testing. The training set contains 347 articles while the test set has 38 articles.

2.1.3 Discourse Lexicalized Tree-Adjoining Grammar

Discourse Lexicalized Tree-Adjoining Grammar (DLTAG) [WJ98, WKSJ99, WSJK03, Web04] is a more recent discourse theory based on the framework of Lexicalized Tree-Adjoining Grammar (LTAG) [G$^+$98]. In LTAG, each word is associated with a set of tree structures in which it can appear. Each tree structure shows the minimal syntactic construction (i.e. phrasal structures such as noun phrases, verb phrases, etc.) and the lexicalized grammar associates each structure with a lexical anchor. Two types of tree structures are defined in LTAG:

1. **Initial tree structures** are elementary tree structures which are built from the smallest text units (e.g. noun phrases, verb phrases, etc.). The internal nodes are labeled as non-terminal and the leaves are labeled as terminals.

2. **Auxiliary tree structures** represent recursion where the tree structure can be expanded and/or modified using two specific operation: (1) substitution and (2) adjunction. Substitution sites are shown by $\Downarrow$ and adjunction sites by $*$.

   The substitution involves replacing the node marked as $\Downarrow$ with the tree being substituted. The restriction of this operation is that only initial trees or trees derived from initial trees can be substituted. The root node of the tree being substituted must be the same as the node being replaced.

   The adjunction operation builds a new tree from an auxiliary tree and any other tree (i.e. initial, auxiliary, or derived). The operation can only be applied on non-terminal nodes which are not marked as substitution. For a given auxiliary tree β and another tree γ (γ can be either initial, auxiliary or derived), the root node of β, called n, must have the same label as the node in γ that is being adjoined. The subtree dominated by node n in tree γ is attached to the root node in tree β. As a result, the whole subtree of tree γ dominated by node n, is now dominated by node n in tree β. Then, the root of tree β, which is labeled as n, is adjoined to tree γ. For example, Figure 5.a shows an auxiliary LTAG tree anchored by the discourse connective *often* and Figure 5.b shows a derived tree. The node VP in the auxiliary tree (5.a) can be adjoined to the VP node in tree (5.b). The result of this adjunction operation is the derived tree (5.c). The order of operations in adjoining these trees is shown in Figure 5.

Similarly to LTAG, which is a grammar to build sentences, DLTAG is a grammar to build discourse structures. In DLTAG, the smallest elements are discourse

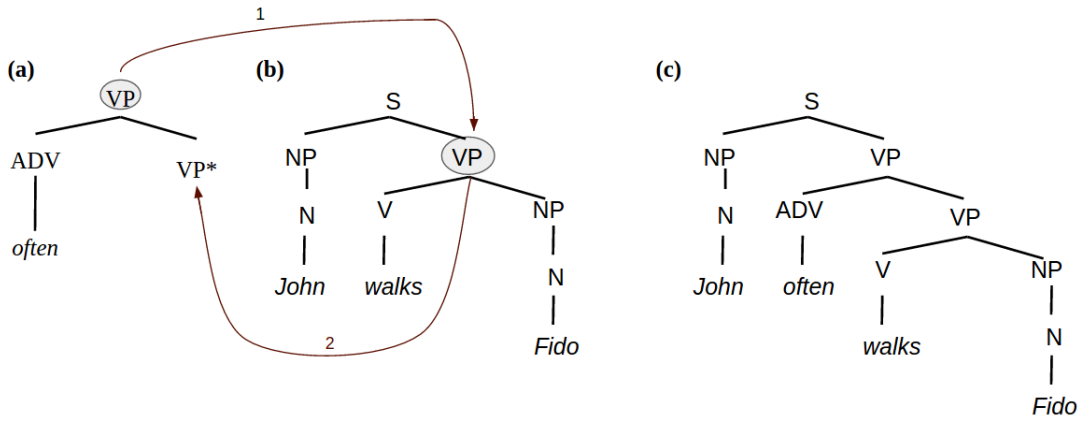


Figure 5: Example of the adjunction operation in the DLTAG framework. (a) is an auxiliary tree, (b) is a derived tree, (c) is the derived tree produced by adjunction of (a) and (b). The example is taken from [BMM05].

segments (aka discourse units equivalent to text spans in RST) and the tree structures are anchored by discourse connectives. These discourse connectives can have a grammatical role such as subordinating conjunctions, coordinating conjunctions, adverbial/propositional phrases or can even be absent (in the case of non-explicit relations). In DLTAG, the initial tree structures can either be parallel or be singleton structures, depending on the discourse connective. In parallel tree structures, the discourse connective is composed of two parts; for example, *on the one hand... on the other hand*. These form two lexical anchors in the initial tree structure of a CONTRAST discourse relation. However, in a singleton tree structure, the discourse connective has only one part, for example *but*; therefore, only one lexical anchor exists. Figure 6 shows the initial tree structure of a parallel contrastive construction[3]. In the figure, D_c, also known as **discourse argument**, stands for discourse clause and $\Downarrow$ indicates the point for a substitution operation. Figure 7 also shows an example of initial tree structure with a singleton construction. In this example, the discourse connective *so* is used between the two discourse arguments. However, a connective could also be

[3]The example is taken from [Web04]

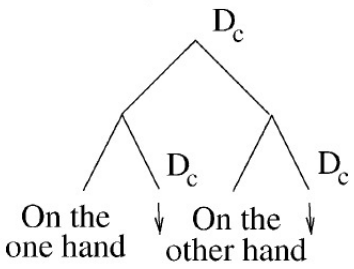


Figure 6: Initial tree structure for a parallel construction representing the construct of a CONTRAST discourse relation signalled by *on the one hand... on the other hand*.

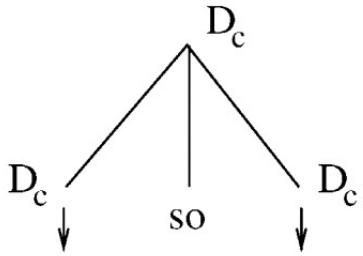


Figure 7: Initial tree structure for a singleton construction used to represent the semantic/syntactic construct of a RESULT discourse relation signalled by *so*.

placed at various other positions such as before the first arguments.

In DLTAG, the meaning of a discourse is represented as a set of predicate-argument relations in which discourse connectives are considered as predicates that take discourse segments as arguments. Unlike the RST framework which explicitly represents the discourse structure of the text as a hierarchical tree structure, in DLTAG the representation of a discourse is linear. This means that the arguments of a discourse predicate can only be text segments, and not nested predicates.

In addition, in DLTAG, discourse relations are either signaled explicitly using discourse connectives or are signalled non-explicitly without using a discourse connective. In this case, the predicate of the discourse relation does not have a lexical realization. The semantics of the discourse should be inferred from the semantic

connection between its discourse arguments.

2.1.4 Penn Discourse TreeBank

The Penn Discourse TreeBank (PDTB) [PDL+08] is much larger than the RST Discourse Treebank corpus [CMO03], with over one million words of articles from the Wall Street Journal [MMS93] manually annotated at the discourse level following the DLTAG discourse framework (see Section 2.1.3). In this corpus, discourse arguments can be either a clause or a sentence. Discourse relations are divided into two main categories: *explicit* and *non-explicit*[4]. Explicit relations are signalled by a discourse connective that links two arguments called Arg1 and Arg2 (Arg2 is the one that is syntactically connected to the discourse connective). Example 3[5] shows an example of an explicit CAUSE relation signalled by the discourse connective *because*.

Example 3

The federal government suspended sales of U.S. savings bonds because **Congress hasn't lifted the ceiling on government debt.** [Explicit CAUSE]

Non-explicit relations hold between adjacent sentences, where there is no discourse connective between them. These relations can be inferred by inserting a discourse connective in Arg2. Example 4[6] shows an implicit CAUSE relation.

Example 4

Several leveraged funds don't want to cut the amount they borrow because it would slash the income they pay shareholders, fund officials said. But

[4]In the examples of this thesis, text segments which are in **bold** represent Arg2; while segment in *italic* refers to Arg1. Discourse connectives in explicit relations and potential discourse connectives in implicit relations are underlined.

[5]The example is taken from the PDTB.

[6]The example is taken from the PDTB.

a few funds have taken other defensive steps. *Some have raised their cash positions to record levels.* Implicit = because **High cash positions help buffer a fund when the market falls.** [Implicit CAUSE]

Apart from these two main types of discourse relations, three other types of discourse relations we're used in the PDTB: AltLex, EntRel and NoRel. AltLex (Alternative Lexicalization) is a type of discourse relation where the insertion of a connective leads to a redundancy. Example 5[7] illustrates an AltLex CAUSE relation signalled by *the most likely reason for this disparity.*

Example 5

I read the exerpts of Wayne Angell's exchange with a Gosbank representative ("Put the Soviet Economy on Golden Rails" editorial page, Oct. 5) with great interest, since the gold standard is one of my areas of research. Mr. Angell is incorrect when he states that the Soviet Union's large gold reserves would give it "great power to establish credibility." *During the latter part of the 19th century, Russia was on a gold standard and had gold reserves representing more than 100% of its outstanding currency, but no one outside Russia used rubles. The Bank of England, on the other hand, had gold reserves that averaged about 30% of its outstanding currency, and Bank of England notes were accepted throughout the world.* AltLex **[The most likely reason for this disparity] is that the Bank of England was a private bank with substantial earning assets, and the common-law rights of creditors to collect claims against the bank were well established in Britain.** [AltLex CAUSE]

AltLexes are inter-sentential relations in the PDTB. If the discourse connective

[7]The example is taken from the PDTB.

is absent and if no implicit or AltLex relation exists between two adjacent sentences, there may exist either an entity transition (i.e. *EntRel* relation) or no relation (*NoRel*). Example 6[8] illustrates an example of an EntRel relation.

Example 6

Hale Milgrim, 41 years old, senior vice president, marketing at Elecktra Entertainment Inc., was named president of Capitol Records Inc., a unit of this entertainment concern. **Mr. Milgrim succeeds David Berman, who resigned last month.** [EntRel]

The PDTB provides a closed list of 100 discourse connectives that can signal an explicit discourse relation. Any connective in this list can either have a discourse usage or not. For example, the connective *"and"* can be used to signal a CONJUNCTION relation as in *"I met my friend and she told me the rumor."*, but could also be used in a non-discourse usage as in *"My friend and I cooked together."*. Discourse parsers rely heavily on discourse connectives to identify explicit relations [PLN09].

Figure 8 shows the inventory of the PDTB discourse relations. As shown in Figure 8, the PDTB introduces a three-level hierarchy of discourse relations. In the first level, four classes of discourse relations are defined: TEMPORAL, CONTINGENCY, COMPARISON and EXPANSION. The relations within this hierarchy are semantically related to each other. The TEMPORAL class contains relations where the two discourse segments (i.e. Arg1 and Arg2) are related to each other temporally. The class CONTINGENCY includes relations that signal a causal relation between discourse segments. A discourse relation belongs to the COMPARISON class when Arg1 and Arg2 are compared and finally when the discourse is expanded using a relation, the relation is classified in the EXPANSION class. At the second level, a total of 16 types are defined (2 types of TEMPORAL relations, 4 types of CONTINGENCY relations, 4 types

[8]The example is taken from the PDTB.

of COMPARISON relations and 6 types of EXPANSION relations). Finally, in the third level, 23 subtypes are defined for most types.

2.2 Text Complexity Assessment

A reader may find a text easy to read, cohesive, coherent, grammatically and lexically sound or on the other hand may find it complex, hard to follow, grammatically heavy or full of uncommon words. To attract and engage readers, a writer should be concerned with the textual content, but also the skillful use of language, the fluent flow of ideas, the lack of grammatical and spelling mistakes, the proper choice of words, etc. Dictionaries can help to choose the most proper words; spelling correction tools can be used to avoid spelling errors; and grammar rules can help avoid grammatical mistakes and build syntactically correct clauses and sentences. However, following writing guidelines, grammar rules and carefully choosing words to deliver a message to a specific reader still allows for a variety of texts that all convey more or less the same content, but with different linguistic styles and nuances. As a result, readers may find some texts more easy to read, concrete, and better-written than others.

The goal of text complexity assessment is to identify if a text is more complex than another [Sid14]. Text complexity is related to the use of linguistic features rather than the reader's readability level or knowledge.

At first glance, text complexity may seem to be an intuitive notion, but it is hard to define precisely. Typical surface features such as sentence length [KFJRC75] and word length [ML69] are often used as indicators of text complexity; however, these features are not accurate enough as they focus on a single textual aspect, the surface level, and ignore other aspects.

In this section we present an overview of previous work toward defining and assessing text complexity. The notion of text complexity is subjective and a clearly

PDTB Discourse Relations	
TEMPORAL	COMPARISON
Synchronous	Contrast
Asynchronous	juxtaposition
precedence	opposition
succession	Pragmatic Contrast
CONTINGENCY	Concession
Cause	exception
reason	contra-exception
result	Pragmatic Concession
Pragmatic Cause	EXCEPTION
justification	Conjunction
Condition	Instantiation
hypothetical	Restatement
general	specification
unreal present	equivalence
unreal past	generalization
factual present	Alternative
factual past	conjunctive
Pragmatic Condition	disjunctive
relevance	chosen alternative
implicit assertion	Exception
	List

Figure 8: Hierarchy of discourse relations in the PDTB [PDL+08] containing semantic classes, types and subtypes.

agreed-upon definition does not seem to exist in the literature. Traditionally, the level of complexity of a text has mostly been correlated with surface features such as word length, number of syllables per word, sentence length, number of tokens per sentence, number of complex words and other surface features (e.g. [KFJRC75, ML69, Gun03, Gun69, DC48]). However, complexity is a characteristic that reflects a collection of linguistic aspects of writing which influence how well a text is written [Sid14]. These aspects influence text complexity at different levels of language:

- Lexical level (e.g. the use of less frequent, uncommon and even obsolete words),
- Syntactic level (e.g. the extortionate or improper use of passive sentences and embedded clauses),
- Discourse level (e.g. vague or weak connections between text segments).

In this thesis, we make two main assumptions:

Assumption #1. Following the steps of [ABE+12, MLB96, TJT02, DH01, STH99, DuB04], we consider that text complexity can be characterized through the study of its linguistic aspects. It is neither influenced by the reader's characteristics, such as their background, education, expertise, level of interest in the material, nor external elements such as typographical features (e.g. text font size, highlights, the use of graphical presentations, etc.). Thus, we assume that the reader's characteristics fit the minimum requirements to understand the text. Matching a reader with a text that can be understood and read easily by the reader is studied in the field of readability prediction which is not the focus of our research [CT14] (for more details see Section 2.2.2).

Assumption #2. Text complexity can be defined as a binary distinction. The availability of relevant data is a major difficulty in any data-driven study. Subject to

the availability of relevant data, text complexity can be assessed as a binary distinction (e.g. *more complex* vs. *less complex* [TITT10], *beginner* vs. *advanced* [CE07]), a multi-level distinction (e.g. *likert scale*, *discrete values* [CE07, FM12], etc.), or even a numeric distinction (e.g. *range of continuous values* [PN08]). Due to the nature of the data used in this thesis (see Section 2.2.3), our work assumes that text complexity is defined as a binary distinction.

Many natural language processing applications can benefit from text complexity assessment. The automatic assessment of text complexity is mostly addressed as a sub-problem of a larger NLP application rather than a standalone problem. For example, in text simplification, text complexity assessment is seen as a first step to identify which textual elements are complex and need to be simplified [Sid06, Sid14, DT98a, CMC$^+$98a, BBE11, CS97, Kau13]. Text complexity assessment can also be used to improve the performance of other natural language processing applications (such as syntactic parsing and machine translation [CS97]).

In the next sections, we provide an overview of the importance and need for text complexity assessment. Then, in Section 2.2.2 we will discuss recent efforts to define and evaluate text complexity. Finally, in Section 2.2.3 we will review the existing annotated corpora across complexity levels.

2.2.1 Motivation for Computational Complexity Assessment

Computational methods for text complexity assessment are useful for human readers and writers as well as for other natural language processing applications.

Following the principle of least effort [Zip49], readers want to have access to information and comprehend it with minimum effort. On the other hand, authors also want to put minimum effort in their writing, while ensuring that their writing is understood. Several factors can influence the comprehension of a text. According to [DuB04] these can be divided into two main categories: 1) *reader-related*, and 2)

text-related factors. *Reader-related* factors depend on the reader characteristics such as prior knowledge, language skills, interest in the topic, etc. Several studies have highlighted the correlation between text comprehension and reader's characteristics (e.g. [Gut81, Pau09, MBB80, LEVDB$^+$00, KSL08, WRO03]). On the other hand, *text-related* factors can themselves be categorized into two groups of features: 2.a) *typographical* and 2.b) *linguistic* factors [DuB04, CT14]. Typographical factors are related to the style and visual features (e.g. text font size, highlights, graphics, etc.). For example, [DC49] showed that the reader's comprehension can be influenced by the existence of supporting graphics and illustrations. Similarly, linguistic factors which focus on the use of language at different levels such as lexical, syntactic and discourse, can influence text complexity.

Several natural language processing applications can benefit from text complexity assessment. This is the case, for example, with syntactic parsing, a core natural language processing application. Syntactic parsing is the process of breaking down a sentence or other sequence of words into its constituents by following a formal grammar, resulting in a parse tree showing the syntactic relations among the constituents [MJ00]. Assessing the complexity of syntactic structures can be useful to improve the performance of syntactic parsers. [CS97], for example, developed a set of corpus-based syntactic transformations to reduce syntactic complexities. As a result, the authors showed an improvement in the performance of parsers and machine translation systems by providing them syntactically simpler sentences as input. Another natural language processing application that can benefit from text complexity assessment is text simplification. Measuring text complexity is a crucial step in automatic text simplification where various aspects of a text need to be simplified in order to make it more accessible [Sid14]. In order to resolve linguistic complexities, a text's complexity level needs to be assessed first. Thus, automatic text complexity assessment can be seen as the first step to automatic text simplification.

The reader's comprehension and ease in reading and following a text is one of the main goals of any writer [DT34, DuB04]. It is inevitable that achieving this goal is influenced by the reader's general knowledge, literacy level and interest in the topic. However, as indicated in Section 2.2, text complexity can be studied computationally if the focus is on the linguistic properties of the text, rather than the reader's characteristics. Adaptive algorithms consider the reader's characteristics to personalize a text for a specific reader, such as [CTC04, HCTCE06, KCTBD12]. Except from these adaptive learning approaches, most other efforts in complexity assessment assume that the target group is fixed and shares similar background knowledge and literacy levels. We followed the same assumption.

It is important to distinguish **text complexity** from **readability analysis**. Readability analysis or readability prediction [CT14] is a field of research that focuses on matching texts to readers. In readability prediction, it is standard to assume that writers do not have specific target readers in mind, but instead that there is a group of readers with different literacy levels and background knowledge (e.g. children, second language learners, professional readers, etc.). The goal of readability analysis is to map the most appropriate texts to target reader groups such that the texts can be well understood by these readers but not as well understood by other reader groups. In contrast, in text complexity assessment, the writers have a specific homogeneous target reader in mind. Indeed, the focus of readability analysis is to take into account the target reader, while the focus of text complexity assessment is to study the linguistic characteristics of the text.

2.2.2 Defining and Assessing Text Complexity

One of the most well-known readability indexes, the Flesch-Kincaid index [KFJRC75], measures a text's complexity level and maps it to an educational level. Traditional complexity measures (e.g. [Cha58, K$^+$63, ZS88, KFJRC75, ML69, Gun03, Gun69,

DC48]) mostly consider a text as a bag of words, a bag of phrases or a bag of sentences and rely on the complexity of a text's building blocks (e.g. words, phrases or sentences). A drawback of this perspective is that is does not take discourse properties into account. Traditional methods do not consider the flow of information in terms of word ordering, phrase adjacency and connections between text segments; all of which can make a text hard to follow, non-coherent and more complex. [WJ12] define discourse using fours aspects: *position of constituents*, *order*, *context* and *adjacency*. Such discourse information plays an important role in text complexity assessment.

More recently, some efforts have been made to improve text complexity assessment by considering richer linguistic features. For example, [SO05] and [CE07] used language models to predict complexity level by using different language models (e.g. a language model for children using children's books, a language model for more advanced readers using scientific papers, etc.). [PN08] also examined a set of cohesion features based on an entity-based approach [BL08] along with other linguistic features. The authors observed that the top five linguistic features which are correlated with textual complexity level are: discourse relations, unigram language model of Simple English Wikipedia, average number of verb phrases, unigram language model of news articles and number of words in a text. However, this corpus-based study was performed on a small corpus where the complexity level of the texts were labeled by human readers. To build such a corpus, the annotators answered five questions on a Likert scale and the average of the grades for each text indicated the complexity level of the text.

The field of text complexity assessment suffers from lack of large scale data. Indeed, to assess text complexity automatically, we need a corpus that is annotated with complexity levels to train and test models. Since there is no consensus on how to measure and label text complexity level, producing such datasets will inevitably be biased to the annotators' understanding of text complexity. Instead, in pairwise text

	Complexity Level	Pairs of Complex and Simple Alignments
1	**Complex**	Griffin Matuszek, who was born without part of his left hand, found his traditional prosthetic hand mostly useless and a bit scary, said his mother, Quinn Cassidy.
		⇓
	Simple	Griffin Matuszek, was born without part of his left hand. He found his artificial hand mostly useless and a bit scary, said his mother, Quinn Cassidy.
2	**Complex**	Cassidy said the hand made Griffin happy and more confident, and didn't break her bank.
		⇓
	Simple	Cassidy said the hand made Griffin happy and more confident. It was not expensive.
3	**Complex**	Printers have been used for other types of prosthetics, but hands were more difficult to develop, designers say.
		⇓
	Simple	Printers have been used for other types of artificial body parts. Hands were more difficult to create, designers say.

Figure 9: Examples of pairs of complex sentences and their simple counterpart from the Newsela corpus [XCBN15].

complexity assessment, there is no need to have an annotated dataset with quantified scores of text complexity. **Pairwise text complexity assessment** is a variation of text complexity assessment where a pair of texts are compared to one another in order to assess if they have the same level of complexity (i.e. both are complex or both are simple) or different levels of complexity (i.e. one is simpler than the other). Generating such datasets is easier and more reliable and can be done both manually and automatically. A few publicly available corpora are manually annotated with pairwise text complexity. *The Simple English Wikipedia* [CK11b] and *Newsela* [XCBN15] are the two major large-scale corpora used in this field (see Section 2.2.3). In addition to these, parallel corpora with different complexity levels can be generated automatically. Producing such monolingual parallel corpora can be viewed as a machine translation problem where the input and output text pairs are in the same language. Consequently, machine translation evaluation metrics are often used in this domain. For example, [ZBG10, WL11, WVDBK12, CK11a] used the BLEU and NIST scores (machine translation metrics that measure word and word sequence overlap between the system output and manual translations) in order to evaluate pairwise text complexity tasks. This way, a parallel corpus for pairwise complexity assessment can be generated automatically by first using text simplification techniques and then preparing a parallel corpus of texts with identical and different complexity levels which are seen as monolingual translation pairs. Manual preparation of parallel corpora at different complexity levels is time consuming, but leads to better quality corpora, on the other hand the automatic generation of such corpora can be done faster, but the quality of such corpora needs to be verified.

The complexity of a text is not characterized by a single aspect (e.g. surface features), but a combination of various linguistic aspects. Figure 9 shows three examples of pairs of texts from the Newsela corpus [XCBN15] which have been manually assessed with their complexity level [XCBN15]. Regardless of the fact that the two texts

Surface feature	More complex	Less complex
Number of sentences	3	6
Average sentence length	19.66	10.00
Average word length	6.08	5.93

Table 1: Comparison of surface features between the more complex and the less complex parallel texts of Figure 9.

Lexical choice in more complex text	$\Rightarrow$	Lexical choice in less complex text
prosthetic	$\rightarrow$	*artificial (body part)*
not break the bank	$\rightarrow$	*not expensive*
develop	$\rightarrow$	*create*

Table 2: Comparison of lexical features between the more complex and the less complex parallel texts of Figure 9.

convey the same information, they have different linguistic characteristics. These differences fall into the following categories:

- **Surface aspect**: According to traditional measures of text complexity assessment (e.g. [KFJRC75, Cha58, K$^+$63, ZS88, ML69, Gun03, Gun69, DC48]), surface aspects of a text are easy-to-compute features which capture a text's characteristics. Such features include the average number of sentences, the average number of words in a sentence, the average length of words, etc. Table 1 shows the differences of the examples of Figure 9 with respect to a few surface features. As can been in Table 1, the three complex sentences are broken into 6 shorter sentences in their less complex counterparts. The average sentence length in the less complex texts is about half of the average sentence length in the more complex texts (10 words versus 19.66 words). In addition, the average word

length is smaller in the less complex texts compared to the more complex ones.

- **Lexical aspect**: Lexical aspects of a text are related to word (lexical) choices. [CD95] pointed out that "It is no accident that vocabulary is also a strong predictor of text difficulty". In the examples of Figure 9, some lexical choices are different across complexity levels; however the meaning is preserved. The lexical variations of the examples in Figure 9 are highlighted in Table 2. Complex words (e.g. *prosthetic*) as well as idioms (e.g. *break the bank*) are simplified.

  Complex word identification which focuses on identifying complex words in context is the first step towards lexical simplification [DT98a, CMC+98a, BBE11]. As noted in Section 1.3, as part of our PhD work, we participated in the 2016 NAACL-SemEval international shared task on complex word identification. This work on complex word identification is published in [DK17a].

- **Syntactic aspect**: In addition to the complexity of individual words and surface features of a text, the complexity of syntactic constituents also influence text complexity [KLP+10]. According to [DEO14], the existence of either subordinated or coordinated embedded clauses increases syntactic complexity. In addition, according to [Sid06], using a passive voice instead of an active voice increases syntactic complexity. The syntactic complexity not only affects human reader's comprehension [JCK+96, LHW+12], but also influences the performance and reliability of natural language processing applications (for example, information extraction [AB92, Eva11], machine translation [GH98] and syntactic parsing [Tom13, MN11]).

  The syntactic transformations used in the examples of Figure 9 are listed as follows:

  - *Dis-embedding relative clauses*:

Griffin Matuszek, who was born without part of his left hand, found his traditional prosthetic hand mostly useless and a bit scary, said his mother, Quinn Cassidy.

⇓

Griffin Matuszek, was born without part of his left hand. He found his artificial hand mostly useless and a bit scary, said his mother, Quinn Cassidy.

– *Separation of subordinating clauses:*

Cassidy said the hand made Griffin happy and more confident, and didn't break her bank.

⇓

Cassidy said the hand made Griffin happy and more confident. It was not expensive.

Printers have been used for other types of prosthetics, but hands were more difficult to develop, designers say.

⇓

Printers have been used for other types of artificial body parts. Hands were more difficult to create, designers say.

- **Discourse aspect**: Compared to the other linguistic aspects, the influence of discourse features on text complexity is understudied [Sid06, Sid03]. Discourse aspects can be divided into two categories: 1) cohesion and 2) coherence aspects. Cohesion focuses on the lexical connection between entities and ideas in the text such as the use of pronouns or other referring expressions [MJ00]. Proper use of referencing influences the ease of following a text and subsequently

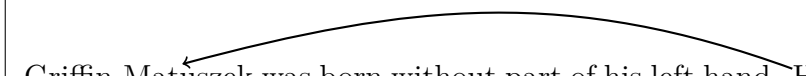
its complexity. On the other hand, coherence focuses on the logical and semantic connections between utterances in a text. Discourse relations are used as linguistic devices to model coherence (see Chapter 2).

In the examples of Figure 9, we observe the following differences at the discourse level between more complex and less complex texts:

– *Co-referencing*:

Griffin Matuszek, who was born without part of his left hand, found his traditional prosthetic hand mostly useless and a bit scary, said his mother, Quinn Cassidy.

⇓



Griffin Matuszek was born without part of his left hand. He found his artificial hand mostly useless and a bit scary, said his mother, Quinn Cassidy.

– *Discourse relation realization (from explicit to implicit)*:

Cassidy said the hand made Griffin happy and more confident, and didn't break her bank. (*explicit*-CONJUNCTION)

⇓

Cassidy said the hand made Griffin happy and more confident. It also was not expensive. (*implicit*-CONJUNCTION)

Printers have been used for other types of prosthetics, but hands were more difficult to develop, designers say. (*explicit*-CONTRAST)

⇓

Printers have been used for other types of artificial body parts. Hands were more difficult to create, designers say. (*implicit*-CONTRAST)

The first phenomenon is related to cohesion; in the more complex text, a longer sentence containing a relative clause is used which in the simpler version the sentence is split into two shorter sentences. To make the resulting text more cohesive the second occurrence of "*Griffin Matuszek*" is therefore substituted with the personal pronoun "He". On the other hand, the last two differences are related to coherence. In the two transformations, the *explicit* realization of the discourse relations are changed to an *implicit* realization. Chapter 5 of this thesis is devoted to the analysis of these last phenomena.

2.2.3 Corpora Across Different Complexity Levels

In order to facilitate the analysis of texts across complexity levels, annotated corpora have been developed. The Simple English Wikipedia corpus [CK11b] and the Newsela corpus [XCBN15] are the two most widely used publicly available corpora across complexity levels.

2.2.3.1 Simple English Wikipedia Corpus

The Simple English Wikipedia (SEW) corpus [CK11b] is a parallel corpus containing regular and simplified versions of Wikipedia articles. The simplified versions of the Wikipedia articles are meant to be more accessible to beginners learning English, such as students, children and adults with learning difficulties. These articles are typically shorter than their regular counterparts, and use simpler words and syntactic structures. The simplified articles were created using their regular counterparts as a basis and following a set of simplification guidelines[9]. In particular, word choices are limited to Basic English[10], a 850-word

[9]https://simple.wikipedia.org/wiki/Wikipedia:How_to_write_Simple_English_pages
[10]https://simple.wikipedia.org/wiki/Wikipedia:Basic_English_ordered_wordlist

	Article Aligned		Sentence Aligned	
	Complex	**Simple**	**Complex**	**Simple**
Average # of sentences per article	42.33	6.41	NA	NA
Average # of words per sentence	25.47	18.57	24.49	18.93

Table 3: Statistics of the Simple English Wikipedia corpus

auxiliary international language, and the VOA Special English Word Book[11], a list of 1580 words. The guidelines are not only limited to lexical choices, but also suggest the use of simpler syntactic structures such as avoiding compound sentences containing embedded conjunctive clauses.

The Simple English Wikipedia (SEW) corpus was first created from Simple Wikipedia articles[12] in 2010. The first version of this corpus contains 137K aligned sentence pairs created from Wikipeda pages downloaded in May 2010. The latest version, released in 2011, contains two parts: a sentence-aligned part containing 167K aligned sentence pairs and 60K aligned articles. As shown in Table 3, as complexity level decreases in both article-aligned and sentence aligned of the Simple English Wikipedia corpus, the sentences tend to become shorter. Also, with decreasing the level of complexity in article-aligned version of this corpus, the average number of sentences per article decreases significantly.

2.2.3.2 Newsela Corpus

The Newsela corpus [XCBN15] contains 1,911 English news articles, which have been manually re-written at most 5 times by professionals, each time with decreasing complexity level. This corpus contains 1,911 original articles, 1,910 articles at complexity levels 1, 2 and 3, 1,847 articles at level 4 and 42 articles

[11]https://simple.wikipedia.org/wiki/Wikipedia:VOA_Special_English_Word_Book

[12]www.simple.wikipedia.org

	Original	**Simp-1**	**Simp-2**	**Simp-3**	**Simp-4**
Average # of sentences per article	49.59	51.27	56.12	56.67	56.78
Average # of words per sentence	23.23	19.44	16.60	14.11	11.91

Table 4: Statistics of the Newsela corpus

at level 5 (easiest). Newsela is meant to help teachers prepare material that match the literacy level required at each grade level. As shown in Table 4, as complexity level decreases in the Newsela corpus, the sentences tend to become shorter; while the average number of sentences per article increases. The Newsela corpus has been used extensively in text simplification and paraphrasing (e.g. [XNP+16, NCBP16]). As both of the Simple English Wikipedia corpus and the Newsela corpus constitute benchmarks in the field, we have used them in our work (see Chapters 3, 4 and 5).

2.3 Conclusion

In this chapter, we have reviewed the most notable and frequently used discourse theories: RST and DLTAG and their associated annotated corpora: the RST-DT corpus and the PDTB corpus. In the second part of this chapter we have reviewed the efforts towards defining text complexity and have introduced the two standard annotated corpora based on complexity levels: the SEW corpus and the Newsela corpus. We have also shown that discourse properties have been traditionally overlooked in text complexity assessment. In the next chapter, we will present our contribution to text complexity assessment using discourse-level properties.

Chapter 3

Contribution of Discourse Features to Text Complexity Assessment

This chapter addresses our research question #1 and presents our contribution to automatic text complexity assessment. The work presented in this chapter was published in [DK16c]. To evaluate the influence of discourse properties for text complexity assessment, we created two data sets based on the Penn Discourse Treebank [PDL+08] (see Section 2.1.4) and the Simple English Wikipedia [CK11a] (see Section 2.2.3) and compared the influence of discourse features with the traditional features used in this task: surface, lexical and syntactic features. Results show that in both data sets coherence features are more correlated to text complexity than the other types of features. In addition, feature selection revealed that with both data sets the most discriminating feature is a coherence feature.

The goal of this work is to address research question #1 (see Section 1.2) by comparing the influence of discourse features to more traditional linguistic features

for text complexity assessment. To do so, we have considered various classes of linguistic features and build a pairwise classification model to compare the complexity of pairs of texts using each class of feature.

3.1 Data Sets

To perform the experiments, we created two different data sets using standard corpora. The first data set was created from the Penn Discourse Treebank (PDTB) [PDL+08]; while, the other was created from the Simple English Wikipedia (SEW) corpus [CK11b]. These two data sets are described below and summarized in Table 5.

3.1.1 The PDTB-based Data Set

Since we aimed to analyze the contribution of different features, we needed a corpus with different complexity levels where features were already annotated or could automatically be tagged. Surface, lexical, syntactic and cohesion features can be easily extracted; however, coherence features are more difficult to extract. Standard resources typically used in computational complexity analysis such as the Simple English Wikipedia [CK11b], the Newsela [XCBN15], Common Core Appendix B[1] and Weebit [VM12] are not annotated with coherence information; hence these features would have to be induced automatically using a discourse parser (e.g. [LNK14], [LDK15]).

In order to have better quality discourse annotations, we used the data set generated by [PN08]. This data set contains 30 articles from the PDTB [PDL+08] (see Section 2.1.4) which are annotated manually with both complexity level

[1]https://www.engageny.org

	PDTB-based Data Set	SEW-based Data Set
Source	Penn Discourse Treebank Corpus	Simple English Wikipedia Corpus
# of pairs of articles	378	1988
# of positive pairs	194	944
# of negative pairs	184	944
Discourse Annotation	Manually Annotated	Extracted using End-to-End parser [LNK14]

Table 5: Summary of the two data sets used in the text complexity assessment experiment.

and discourse information. The complexity level of the articles is indicated on a scale of 1.0 (easy) to 5.0 (difficult). Using this set of articles, we built a data set containing pairs of articles whose complexity levels differed by at least n points. We set $n = 0.7$ which is the standard deviation of the data set (i.e. a pair of articles with a difference in complexity level of 0.7 or more is assumed to have different complexity levels compared to a pair whose complexity scores differ by more than 0.7). As a result, our data set consists of 378 instances with 194 positive instances (i.e. same complexity level where the difference between the complexity scores is less than or equal to 0.7) and 184 negative instances (i.e. different complexity levels where the difference between complexity scores is greater than 0.7). Then, each pair of articles is represented as a feature vector where the value of each feature is the difference between the values of the corresponding feature in each article. For example, for a given pair of articles $< a_1, a_2 >$, the corresponding feature vector will be:

$$V_{a_1,a_2} =< F_1^{a_1} - F_1^{a_2}, F_2^{a_1} - F_2^{a_2}, ..., F_n^{a_1} - F_n^{a_2} >$$

where V_{a_1,a_2} represents the feature vector of a given pair of articles $< a_1, a_2 >$, $F_i^{a_1}$ corresponds to the value of the i^{th} feature for article a_1 and $F_i^{a_2}$ corresponds

Explicit	Implicit
If the light is red, stop because otherwise you will get a ticket. [Explicit, CAUSE]	If the light is red, stop. Otherwise you will get a ticket. [Implicit, CAUSE]

Figure 10: Example of explicit and implicit realizations of a discourse relation.

to the value of the i^{th} feature for article a_2 and n is the total number of features (in our case $n = 16$ (see Section 3.2)).

Because the [PN08] data set is a subset of the PDTB, it is also annotated with discourse information. Recall from Section 2.1.4 that the annotation framework of the PDTB is based on the DLTAG framework [Web04]. In this framework, a set of 100 discourse markers (e.g. *because*, *since*, *although*, etc.) are used as predicates that take two arguments: Arg1 and Arg2, where Arg2 is the argument that contains the discourse marker. The PDTB annotates both explicit and implicit discourse relations. Figure 10, taken from [PDL+08], shows an explicit relation which is changed to an implicit one by removing the discourse marker *because*.

In addition to labeling discourse relation realizations (i.e. explicit or implicit) and discourse markers (e.g. *because*, *since*, etc.), the PDTB also annotates the sense of each relation using three levels of granularity. At the top level, four classes of senses are used: TEMPORAL, CONTINGENCY, COMPARISON and EXPANSION. Each class is expanded into 16 second level senses; themselves subdivided into 23 third-level senses. In our work, we considered the 16 relations at the second-level of the PDTB relation inventory (see Section 2.1.4)[2].

[2]These are: Asynchronous, Synchronous, Cause, Pragmatic Cause, Condition, Pragmatic Condition, Contrast, Pragmatic Contrast, Concession, Pragmatic Concession, Conjunction, Instantiation, Restatement, Alternative, Exception, List.

3.1.2 The SEW-based Data Set

In order to validate our results, we created a larger data set but this time with induced discourse information. To do so, a subset of the Simple English Wikipedia (SEW) corpus [CK11b] (see Section 2.2.3) was randomly chosen to build pairs of articles. Recall that the latest version of SEW corpus contains two sections that are 1) article-aligned and 2) sentence-aligned. We used the article-aligned section which contains around 60K aligned pairs of regular and simple articles. Since this corpus is not manually annotated with discourse information, we used the End-to-End parser [LNK14] to annotate it. In total, we created 1988 pairs of articles consisting of 994 positive and 994 negative instances. Similarly to the PDTB-based data set, each positive instance represents a pair of articles at the same complexity level (i.e. either both complex or both simple). On the other hand, for each negative instance, we chose a pair of aligned articles from the SEW corpus (i.e. a pair of aligned articles containing one article taken from Wikipedia and its simpler version taken from the SEW).

3.2 Features for Predicting Text Complexity

To predict text complexity, we have considered 16 individual features grouped into five classes. These are summarized in Table 6 and described below.

3.2.1 Coherence Features

For a well written text to be coherent, utterances need to be connected logically and semantically using discourse relations. We considered coherence features in order to measure the association between this class of features and text complexity levels. Our coherence features include:

Class of Features	**Index**	**Feature Set**
Coherence features	*F1*	Log_score of <realization-discourse relation>
	F2	Log_score of <discourse relation-discourse marker>
	F3	Log_score of <realization-discourse relation-discourse marker>
	F4	Discourse relation frequency
Cohesion features	*F5*	Average # of pronouns per sentence
	F6	Average # of definite articles per sentence
Surface features	*F7*	Text length
	F8	Average # of characters per word
	F9	Average # of words per sentence
Lexical features	*F10*	Average # of word overlaps per sentence
	F11	Average # of synonyms of words in WordNet
	F12	Average # of frequency of words in Google Ngram corpus
Syntactic features	*F13*	Average # of verb phrases per sentence
	F14	Average # of noun phrases per sentence
	F15	Average # of subordinate clauses per sentence
	F16	Average height of syntactic parse tree

Table 6: List of features used for complexity assessment.

F1. Pairs of <realization, discourse relations> (e.g. <*explicit,* CONTRAST>).

F2. Pairs of <discourse relations, discourse markers>, where applicable (e.g. <CONTRAST, *but*>).

F3. Triplets of <discourse relations, realizations, discourse markers>, where applicable (e.g. <CONTRAST, *explicit, but*>).

F4. Frequency of discourse relations.

Each article was considered as a bag of discourse properties. Then for features *F1*, *F2* and *F3*, the log score of the probability of each article is calculated using Formulas (1) and (2). Considering a particular discourse feature (e.g. pairs of <discourse relations, discourse markers>), each article may contain a combination of n occurrences of this feature with k different feature values. The probability of observing such an article is calculated using the multinomial probability mass function shown in Formula (2). In order to prevent arithmetic underflow and be more computationally efficient, we used the log likelihood of this probability mass function as shown in Formula (1).

$$P = P(n)\frac{n!}{x_1!...x_k!}P_1...P_k \tag{1}$$

$$log\_score(P) = log(P(n)) + log(n!) + \sum_{i=1}^{k}(x_i log(p_i) - log(x_i!)) \tag{2}$$

$P(n)$ is the probability of an article with n instances of the feature we are considering, x_i is the number of times a feature has its i^{th} value and P_i is the probability of a feature to have its i^{th} value based on all the articles of the PDTB. For example, for the feature *F1* (i.e. pair of <realization, discourse relation>), consider an article containing <*explicit,* CONTRAST>, <*implicit,*

CAUSALITY> and <*explicit,* CONTRAST>. In this case, n is the total number of *F1* features we have in the article (i.e. $n = 3$), and $P(n)$ is the probability of an article to have 3 such features across all PDTB articles. In addition, $x_1 = 2$ because we have two <*explicit,* CONTRAST> pairs and P_1 is the probability of observing the pair <*explicit,* CONTRAST> over all possible pairs of <realization, discourse relation>. Similarly, $x_2 = 1$ and P_2 is the probability of observing <*implicit,* CAUSALITY> pair over all possible pairs of <realization, discourse relation>.

3.2.2 Cohesion Features

Cohesion is an important property of well-written texts [GWJ95, BL08]. Addressing an entity for the first time in a text is different from further mentions to the entity. Proper use of referencing influences the ease of following a text and subsequently its complexity. Pronoun resolution can affect text cohesion in the way that it prevents repetition. Also, according to [HH76], definite description is an important characteristic of well-written texts. Thus, in order to measure the influence of cohesion on text complexity, we considered the following cohesive devices.

F5. Average number of pronouns per sentence.

F6. Average number of definite articles per sentence.

3.2.3 Surface Features

Surface features have traditionally been used (e.g. [Cha58, K$^+$63, ZS88, KFJRC75, ML69, Gun03, Gun69, DC48]) to measure readability level. [PN08] showed that the only significant surface feature correlated with text complexity level was the

length of the text. As a consequence, we investigated the influence of surface features by considering the following three surface features:

F7. Text length as measured by the number of words.

F8. Average number of characters per word.

F9. Average number of words per sentence.

3.2.4 Lexical Features

In order to capture the influence of lexical choices across complexity levels, we considered the following three lexical features:

F10. Average number of word overlaps per sentence in pairs of consecutive sentences.

F11. Average number of synonyms of words in WordNet.

F12. Average frequency of words in the Google N-gram (Web1T) corpus.

The lexical complexity of a text can be influenced by the number of words that are used in consecutive sentences. This means that if some words are used repetitively rather than introducing new words in the following sentences, the text should be simpler. This is captured by feature *F10*: "*Average # of word overlaps per sentence*" which calculates the average number of word overlaps in all consecutive sentences.

In addition, the number of synonyms of a word can be correlated to its complexity level. To account for this feature, *F11*: "Average # of synonyms of words in WordNet" is introduced to capture the complexity of the words [Mil95]. Moreover, the frequency of a word can be an indicator of its simplicity. Also, feature

Feature set	No. features	SEW-based Data Set	p-value	Stat. Sign	PDTB-based Data Set	p-value	Stat. Sign
Baseline	N/A	50.00%	N/A	N/A	51.23%	N/A	N/A
All features	16	**94.96%**	N/A	N/A	**69.04%**	N/A	N/A
Coherence only	4	**93.76%**	0.15	=	**64.02%**	0.45	=
Cohesion only	2	66.09%	0.00	⇓	57.93%	0.01	⇓
Surface only	3	83.45%	0.00	⇓	51.32%	0.00	⇓
Lexical only	3	78.20%	0.00	⇓	46.29%	0.00	⇓
Syntactic only	4	79.32%	0.00	⇓	62.16%	0.24	=
All-Coherence	12	**86.70%**	0.00	⇓	**62.43%**	0.08	⇓
All-Cohesion	14	95.32%	0.44	=	68.25%	0.76	=
All-Surface	13	95.10%	0.43	=	68.25%	0.61	=
All-Lexical	13	95.42%	0.38	=	64.81%	0.57	=
All-Syntactic	12	94.30%	0.31	=	66.40%	0.67	=

Table 7: Accuracy of the Random Forest models built using different subset of features.

F12: "*Average frequency of words in Google N-gram corpus*" is used based on the assumption that simpler words are more frequently used. In order to measure the frequency of each word, we used the Google N-gram corpus [MSA+11]. Thus, pairs of articles at the same complexity level tend to have similar lexical features compared to pairs of articles at different complexity levels.

3.2.5 Syntactic Features

According to [KLP+10], syntactic structures seem to affect text complexity level. As [BL08] note, more noun phrases make texts more complex and harder to understand. In addition, [BG01] pointed out that the use of multiple verb phrases in a sentence can make the communicative goal of a text clearer as explicit discourse markers will be used to connect them; however it can also make a text harder to understand for less educated adults or children. The

[SO05] readability assessment model was built based on a trigram language model, syntactic and surface features. Based on these previous works, we used the same syntactic features which included:

F13. Average number of verb phrases per sentence.

F14. Average number of noun phrases per sentence.

F15. Average number of subordinate clauses per sentence.

F16. Average height of syntactic parse tree.

These features were determined using the Stanford parser [TKMS03].

3.3 Results and Analysis

In order to investigate the influence of each class of feature to assess the complexity level of a given pair of articles, we built several Random Forest classifiers and experimented with various subsets of features. Because the data sets are balanced we used accuracy as a measure of performance. Table 7 shows the accuracy of the various classifiers on our data sets using 10-fold cross-validation. In order to test the statistical significance of the results, we conducted a two-sample t-test (with a confidence level of 90%) comparing the models built using each feature set to the model trained using all features. A statistically significant decrease ($\Downarrow$) or no difference ($=$) is specified in the column labeled *Stat. Sign.*

Our baseline is to consider no feature and simply assign the class label of the majority class. As indicated in Table 7, the baseline is about 50% for both data sets. When all features are used, the accuracy of the classifier trained on the

SEW-based data set is 94.96% and the one trained on the PDTB-based data set is 69.04%.

Considering only one class of features, the models trained using coherence features on both data sets outperformed the others (93.76% and 64.02%) and their accuracy are statistically as high as using all features together. However one must also note that there is a significant difference between the number of features (4 for coherence only vs. 16 for all features). Indeed, in both data sets, cohesion features are more useful than lexical features and less than syntactic features.

Furthermore, it is interesting to note that surface features seem to be more discriminating in the SEW articles rather than in PDTB articles; however, the opposite is also true about cohesion features. In addition, the decrease in the accuracy of all classifiers trained on the SEW using only one feature except coherence features is statistically significant. The same is true about the models trained on the PDTB with the only difference being that the one trained using only syntactic features performs as well as the one trained using all features (62.16% vs. 69.04%).

The last section of Table 7 shows the classification results when excluding only one class of features. In this case, removing coherence features leads to a more significant drop in performance compared to the other classes of features. The classifier trained using all features except the coherence features achieves an accuracy of 86.70% and 62.43% on the SEW and PDTB corpus respectively. This decrease in both models is statistically significant; however the changes in the accuracy of other classifiers trained using all features excluding only one class is not statistically significant.

Index	SEW-based Data Set	Index	PDTB-based Data Set
F2	Log_score of <discourse relation-marker>	*F1*	Log_score of <realization-discourse relation>
F9	Average # of words per sentence	*F3*	Log_score of <realization-relation-marker>
F14	Average # of noun phrases per sentence	*F4*	Discourse relation frequency
F7	Text length	*F5*	Average # of pronouns per sentence
F16	Average height of syntactic parse tree	*F9*	Average # of words per sentence
F13	Average # of verb phrases per sentence	*F2*	Log_score of <discourse relation-marker>
F15	Average # of subordinate clauses per sentence	*F7*	Text length
F10	Average # of word overlaps per sentence	*F8*	Average # of characters per word
F8	Average # of characters per word	*F12*	Average frequency of words in Web1T corpus
F4	Discourse relation frequency	*F11*	Average # of synonyms of words in WordNet
F6	Average # of definite articles per sentence	*F6*	Average # of definite articles per sentence
F11	Average # of synonyms of words in WordNet	*F10*	Average # of word overlaps per sentence
F3	Log_score of <realization-relation-marker>	*F15*	Average # of subordinate clauses per sentence
F1	Log_score of <realization-discourse relation>	*F14*	Average # of noun phrases per sentence
F12	Average frequency of words in Web1T corpus	*F13*	Average # of verb phrases per sentence
F5	Average # of pronouns per sentence	*F16*	Average height of syntactic parse tree

Table 8: Features ranked by information gain

3.3.1 Feature Selection

In any classification problem, feature selection is useful to identify the most discriminating features and reduce the dimensionality and model complexity by removing the least discriminating ones. In this classification problem, we built several classifiers using different subsets of features; however, identifying how well a feature can discriminate the classes is helpful in building a more efficient model with fewer features.

Using our pairwise classifier built with all the features, we ranked the features by their information gain. Table 8 shows all the features used in the two models using all the features trained on the PDTB-based data set and the SEW-based data set.

As can be seen in Table 8, coherence features are among the most discriminating features on the PDTB-based data set as they hold the top three positions. Also, the most discriminating feature on the SEW-based data set is a coherence feature. We investigated the power of only the top feature in both data sets by classifying the data using only this single feature and evaluated using 10-fold cross-validation. Using only *F1*: "*log_score of <realization, discourse relation>*" to classify the PDTB-based data set, we achieved an accuracy of 56.34%. This feature on its own outperformed the individual class of surface features and lexical features and performed as well as combining the features of the two classes (four features). It also performed almost as well as the two cohesion features (*F5*, *F6*). In addition, using only the feature *F2*: "*log_score of <discourse relation, discourse marker>*" on the SEW corpus resulted in an accuracy of 77.26% which is much higher than the accuracy of the classifier built using the class of cohesion and almost as good as lexical features.

3.4 Conclusion

In Chapter 2, we introduced the problem of computational text complexity assessment and tried to differentiate from readability prediction in its consideration of target reader. We also discussed the importance of studying text complexity assessment from different perspectives: 1) the readers and 2) other natural language processing applications. Due to the limitations with respect to availability of relevant data, we introduced the problem of pairwise text complexity assessment. In this chapter, we have addressed research question #1 by investigating the influence of discourse features compared to more traditional linguistic features in pairwise text complexity assessment. We experimented with two data sets created from standard corpora and used a combination of

16 features, grouped into five classes (surface, lexical, syntactic, cohesion and coherence features). Although the use of all features resulted in the highest accuracy, the use of only 4 coherence features performed statistically as well on both data sets. In addition, removing only one class of features from the combination of all the features did not affect the accuracy; except for coherence features. Removing the class of coherence features from the combination of all features led to a statistically significant decrease in accuracy. Thus, we can conclude a strong correlation between text coherence and text complexity. More details of this work can be found in [DK16c].

Following the results of this chapter regarding the discriminating power of coherence features for predicting text complexity, in the next chapter we address research question # 2 and study more specifically how textual complexity influences discourse-level choices.

Chapter 4

Influence of Text Complexity on Discourse-Level Linguistic Choices

In Chapter 3, we identified that discourse properties are discriminating to assess textual complexity. In the present chapter, we dig more into this issue to identify exactly how discourse properties change across complexity levels.

Text complexity can be influenced by making different choices at the lexical and syntactic levels (e.g. [DT98b, CMC+98b, BBE11, CS97, Sid06, Kau13]). However, discourse-level choices may also affect a text's complexity. As shown in Chapter 3, discourse-level properties are among the strongest features which can be used to differentiate texts according to their complexity levels. In this chapter, we try to answer research question # 2 (see Section 1.2) and present our contribution on the influence of text complexity on discourse-level linguistic choices. In particular, we investigate the effect of complexity level on (1) the usage of discourse relations, (2) the usage of discourse markers and (3) the distribution of discourse markers signaling explicit discourse relations. A more

condensed version of this chapter was published in [DK15].

4.1 Introduction

A text's discourse-level properties have been shown to be correlated to various dimensions such as their genre, their level of formality, their level of complexity, etc. For example, [Web09] and [BDK14] showed that the textual genre influences the choice of discourse relation. In order to produce texts at various complexity levels, several techniques have been proposed to simplify texts at the lexical level (e.g. [DT98a, YPDNML10]), the syntactic level (e.g. [CS97, SDOCJ$^+$10]) and the discourse level (e.g. [Sid06]). In particular, [WRO03] used "simpler" discourse markers (e.g. *but* instead of *however*) to generate more readable texts for people with a lower level of literacy. In the process of text simplification, the writer's goal is to reformulate a text to make it easier to read and understand; however, its informational content should be preserved. Based on this assumption, we suspected that the simplification process should not change the semantic or logical relations between textual units.

As discussed in Section 2.2.2, one can view texts at different complexity levels as translations of their "regular" counterpart. Using this perspective, we can argue that during the translation, translators may choose to use discourse relations and discourse markers differently in the translated text by adding or removing them or making implicit relations explicit or vice versa; all the while, preserving the meaning of the original text. For example, in the context of machine translation, [MW13] have shown that fewer discourse markers were used in the German or French translations of the Newstest2012 parallel corpus[1] compared to its English counterpart.

[1] http://www.statmt.org/wmt12/

In this chapter, we investigate the influence of the complexity level on (1) the usage of explicit discourse relations, (2) the usage of discourse markers, as well as (3) the distribution of discourse markers. We analyzed the Simple English Wikipedia corpus [CK11b] (see Section 2.2.3). We used the log-likelihood ratio to rank the discourse relations and discourse markers with texts at various levels of complexity.

4.2 Background

As [PDL+08] noted, discourse markers constitute valuable features to identify explicit discourse relations; however, they may be used in a non-discourse context. Several works have already addressed the identification, selection and placement of discourse markers in coherent texts (e.g. [Kno96, MM95, DEMP97, LKN09, PK13, FM13]). However, to our knowledge, no previous work has attempted to investigate the effect of complexity level on the usage of discourse markers and discourse relations using large scale parallel corpora.

Several attempts have been made to enhance the complexity level of texts at different levels (i.e. lexical, syntactic or discourse levels) (e.g. [DT98a, YPDNML10, CS97, SDOCJ+10, Sid06]), or generating texts across different complexity levels for various groups of audiences. For example, Williams' text generation system [WRO03] generates texts at different levels of complexity; however the simplification rules are based on a manual analysis of a small corpus. Three parallel texts (each with an average of 1000 to 2000 words) revealed that some discourse markers like *so* and *but* are preferable to use in simpler texts than other discourse markers such as *therefore* or *hence*. She also reported that a more frequent usage of discourse markers result in more readable texts. This last result seems to contradict our own (see Section 4.4.3) which are based

on a much larger corpus.

Another related work is that of Siddharthan [Sid06] who focused on textual simplification. Although the main focus of this work was on syntactic simplifications, Siddharthan also addressed the use of specific discourse markers in order to increase the textual cohesion of the simplified texts. Once the original sentences were simplified syntactically, he selected specific discourse markers in order to preserve the discourse relation between the resulting conjoined clauses. To do so, he used a set of 13 discourse markers and associated each discourse marker to a single discourse relation. The actual selection of the most appropriate discourse marker was based on [WRO03]'s recommendations. For example, every CONCESSION relation resulted in the use of the discourse marker *but*. Although Siddharthan's main focus was not on discourse-level choices, a number of assumptions were made. In comparison, our work is based on a statistical analysis of a much larger corpus, uses a much larger set of discourse markers (the list of 100 discourse markers from the PDTB [PDL$^+$08]) and does not assume a one-to-one correspondence between discourse markers and discourse relations.

4.3 Data Sets

To investigate the influence of the complexity level on the usage of discourse relations and discourse markers, these were extracted automatically from parallel corpora across different complexity levels.

4.3.1 The Simple English Wikipedia Corpus

As noted in Section 2.2.3, because they are manually annotated with discourse relations, the RST-DT corpus [COM02] and the Penn Discourse Tree Bank

(PDTB) [PDL+08] constitute two of the most widely used corpora for discourse analysis. However, these corpora could not be used in our work because we needed a parallel corpus across different complexity levels. Instead, we again used the Simple English Wikipedia (SEW) corpus [CK11b] (see Section 2.2.3) but had to label it automatically with discourse information.

4.3.2 Labeling the Corpus

Because the Simple English Wikipedia corpus is not discourse-annotated, to label discourse relations and identify discourse markers signalling explicit discourse relations, we have automatically parsed its parallel sentences using the PDTB-style End-to-End discourse parser [LNK14].

Several other publicly available discourse parsers could have been used (eg. [LDK15, HPAdI10, FH12]). We chose the PDTB-style End-to-End discourse parser because we needed local discourse-level information that include the type of discourse relations (i.e. *implicit* or *explicit*), the name of the discourse relation and the discourse marker when applicable. When the work was performed, the PDTB-style End-to-End discourse parser was the best performing parser (with an F-measure between 80.61% and 86.77% depending on the evaluation criteria) providing all these features. Although the parser can identify both explicit and implicit discourse relations, we only considered explicit discourse relations. As the performance of this parser for implicit relations drops significantly and because we are interested in the usage of discourse markers which signal explicit discourse relations, implicit relations were not considered.

The End-to-End parser [LNK14] uses the PDTB inventory of relations [PDL+08] described in Section 2.1.4. Recall that the relations are organized into 3 levels of granularity. Level 1 includes four relations: TEMPORAL, CONTINGENCY,

	Regular version	**Simple version**
# of sentences	167,690	189,572
# of discourse connectives	52,648	48,412
token/sentence ratio	23.36	18.45
discourse connective/token ratio	0.098	0.093
discourse connection/sentence ratio	0.31	0.25

Table 9: Statistics of the Simple English Wikipedia corpus

COMPARISON and EXPANSION. In our experiment, we used the 2^{nd} level that defines 16 relations, but only 12 relations were present in the corpus. In addition, the End-to-End parser uses an inventory of 100 discourse connectives, but only 72 were actually present in the Simple English Wikipedia corpus.

Table 9 provides statistics about the annotation of the regular and simple versions of the Simple English Wikipedia corpus with the End-to-End parser. As shown in Table 9, the regular version of the sentence-aligned part of the corpus contains 167K sentences; however in the simple version, the number of sentences increases to 189K sentences. In the simple version, sentences tend to be shorter (18.45 words versus 23.36) and fewer discourse connectives are used. In addition, the ratio of discourse connective per token is tend to be lower in the simple version compared to the regular version (0.093 vs 0.098).

4.4 Results and Analysis

Once the Simple English Wikipedia was tagged with discourse markers and discourse relations, we analysed: (1) the usage of discourse relations, (2) the usage of discourse markers and (3) the distribution of discourse markers over discourse relations across complexity levels.

4.4.1 Comparing two Corpora using Frequency Profiling

In order to compare the corpora with respect to discourse-level linguistic choices, we used frequency profiling [RG00]. Depending on the discourse-level linguistic choice being studied, we adapted the frequency profiling. If the feature being studied is either (1) the explicit discourse relation, or (2) pairs of <explicit discourse relation, discourse marker>, we created a contingency table containing the frequency of each instance of the discourse-level linguistic choice across complexity levels. For example, if the discourse-level linguistic choice being studied is the usage of explicit discourse relations, each instance is an explicit discourse relation (e.g. CONTRAST, CAUSE, etc.). Table 10 shows a generic contingency table for explicit discourse relations across the complex and simple versions of a corpus. In the table, a_i and b_i correspond to the frequency of each explicit discourse relation in the complex version and simple version of the data set respectively. A and B denote the total frequency of all explicit discourse relations in the complex and simple versions. Considering each discourse relation as a discrete random variable, the expected frequency of each explicit discourse relation in each version of the parallel data set is calculated using Formula 3. In this formula, a_i and b_i are observed values of each random variable in each version of the corpus. The log-likelihood of each explicit discourse relation is calculated using Formula 4.

$$E(i)_{Complex} = \frac{A * (a_i + b_i)}{A + B}$$
$$E(i)_{Simple} = \frac{B * (a_i + b_i)}{A + B} \tag{3}$$

$$LL_i = 2 * ((a_i * ln(\frac{a_i}{E(i)_{Complex}}) + (b_i * ln(\frac{b_i}{E(i)_{simple}})) \tag{4}$$

	Frequency in Regular Version	**Frequency in Simple Version**	**Total**
Relation #1	a_1	b_1	$a_1 + b_1$
Relation #2	a_2	b_2	$a_2 + b_2$
⋮	⋮	⋮	⋮
Total	**A**	**B**	**A+B**

Table 10: Contingency table of explicit discourse relations across complex and simple versions of a corpus.

4.4.2 Effect of Text Complexity on the Usage of Discourse Relations

Once the parallel corpus was parsed with the End-to-End parser [LNK14], we extracted the explicit discourse relations in both the regular and the simple versions. In order to eliminate the effect of corpus size, we considered the relative frequencies of discourse relations, then we performed frequency profiling using the log-likelihood ratio as shown in Section 4.4.1 [RG00]. This measure allows us to compare the frequency of discourse relations across the regular and the simple versions and sort them according to the importance of their relative frequencies. The log-likelihood ratios themselves only provide a measure of which discourse relations are statistically more informative. The results are shown in Table 11 in decreasing order of log-likelihood ratio. The relations at the top of the table are therefore more indicative of the regular version, as compared to the simple versions of the corpus.

According to Table 11, the most differences stem from the relations of CONTRAST, CAUSE and CONCESSION; however in both the regular and the simple

Discourse Relation	**Regular Version**	**Simple Version**	**LL Ratio**
Contrast	18.10%	16.29%	20.76
Cause	7.62%	8.64%	13.82
Concession	2.88%	2.33%	12.49
Restatement	0.31%	0.20%	4.85
Condition	4.06%	4.46%	4.01
Asychronous	14.76%	15.31%	2.22
Synchrony	12.51%	12.75%	0.48
Exception	0.04%	0.05%	0.22
List	0.01%	0.02%	0.17
Conjunction	36.52%	36.72%	0.12
Alternative	1.75%	1.78%	0.06
Instantiation	1.38%	1.39%	0.00

Table 11: Relative frequency of discourse relations across regular and simple versions of the Simple English Wikipedia corpus sorted by log-likelihood ratio

versions, the three most frequent discourse relations are CONJUNCTION, CONTRAST and ASYNCHRONOUS.

In order to verify if these changes are statistically significant, we first performed a normality test using the IBM SPSS software[2] to investigate the characteristics of our data set. According to this test, the relative frequency of discourse relations in the regular and simple versions are not normally distributed. Consequently, we have used the Wilcoxon test [Wil45] of statistical significance to see if the difference across the two corpora are statistically significant. The Wilcoxon test is a non-parametric statistical hypothesis test which is an alternative to the Student's t-test when the population is not normally distributed. According to this test, the differences in the relative frequencies of discourse relations are not statistically significant. As a result, we can conclude that the usage of explicit discourse relations seems to be preserved across different complexity levels of this parallel corpus.

4.4.3 Effect of Text Complexity on the Usage of Discourse Markers

Given that the usage of explicit discourse relations seems to be preserved, we next turned to how they are signalled across complexity levels. Discourse markers can signal more than one discourse relations. For example, *although* can signal both a CONCESSION and a CONTRAST relation. In this experiment, we were interested in investigating the distribution of discourse markers over discourse relations.

Once all the discourse markers and discourse relations were extracted using the End-to-End parser [LNK14], we constructed <discourse marker,discourse

[2]http://www-01.ibm.com/software/analytics/spss/

Discourse Connective/ Discourse Relation pair	**Regular version**	**Simple version**
because/CAUSE	2.24%	3.73%
thus/CAUSE	1.35%	0.74%
although/CONTRAST	1.69%	1.04%
so/CAUSE	1.15%	1.81%
while/CONTRAST	3.55%	2.64%
when/SYNCHRONY	6.20%	7.53%
also/CONJUNCTION	16.97%	18.94%
as/SYNCHRONY	4.71%	3.79%
although/CONCESSION	1.78%	1.28%
but/CONTRAST	6.19%	7.12%

Table 12: Relative frequency of <discourse marker,discourse relation> pairs across regular and simple versions of the Simple English Wikipedia corpus sorted by log-likelihood ratio

relation> pairs in order to disambiguate discourse markers that signal more than one discourse relation. As a result, we created a set of 119 unique <discourse marker,discourse relation> pairs. Then, we again used the log-likelihood ratio to sort the pairs. Hence, a <discourse marker,discourse relation> pair with a higher log-likelihood ratio is more indicative of the regular version, as compared to the simple version of the corpus. Table 12 shows the 10 most discriminating pairs across the regular and simple versions (see also Appendix A).

Using all <discourse marker,discourse relation> pairs extracted automatically, we have again performed a statistical significance test in order to determine if the difference in the relative frequency of <discourse marker,discourse relation> pairs across corpora is statistically significant. Similarly to the first analysis (see

Section 4.4.2), we first performed a normality test using the IBM SPSS software. The results revealed that <discourse marker,discourse relation> pairs are not normally distributed across corpora. The relative frequency of some pairs such as *because*/CAUSE, *so*/CAUSE and *but*/CONTRAST is higher in the simple version, while it is lower for other pairs such as *thus*/CAUSE, *although*/CONTRAST and *while*/CONTRAST. The Wilcoxon statistical significance test showed that the relative frequency of <discourse marker,discourse relation> pairs across different complexity levels is statistically different. More precisely, the Wilcoxon test revealed that in the simple version of the Simple English Wikipedia, discourse markers are used less frequently than in its regular counterpart. This is an interesting finding as it seems to indicate that to make a text more accessible, the use of discourse markers should be reduced; hence not indicating discourse relations explicitly.

4.4.4 Effect of Text Complexity on the Distribution of Discourse Markers over Discourse Relations

Once we determined that there is a difference in how discourse markers are used to signal a discourse relation across corpora, we tried to verify if the distribution of discourse markers to signal different discourse relations is different across complexity levels. For example, the discourse marker *while* can signal a CONTRAST (as in sentence 1 of Example 7); or a SYNCHRONOUS relation as in sentence 2 of Example 7.

Example 7

1. **While** [any form of energy may be conserved], [electricity is the type most commonly referred to in connection with conservation.]/CONTRAST

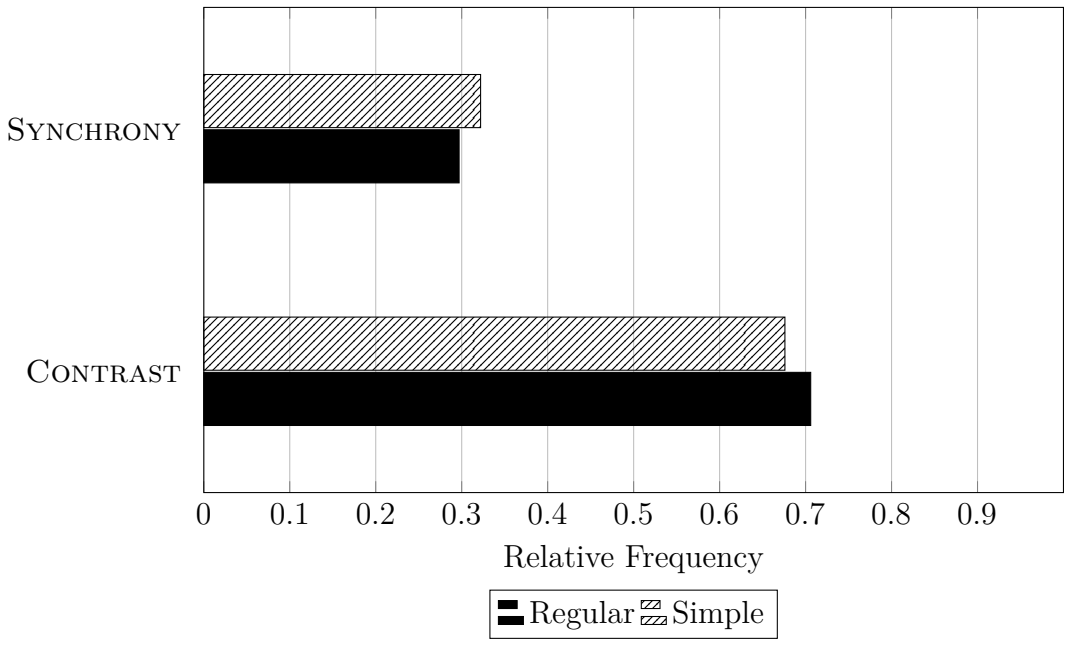


Figure 11: Distribution of the discourse marker *while* with respect to the discourse relation it signals across the Simple English Wikipedia corpus.

2. [He began his career in primary education] **while** [an undergraduate teaching at the Children's Community School]/SYNCHRONOUS

Figure 11 shows the usage of the discourse marker *while* in the Simple English Wikipedia to signal these two discourse relations.

In the regular version of the Simple English Wikipedia corpus, each discourse marker conveys on average 1.68 relations. On the other hand, this number drops to 1.61 in the simple version of the same corpus. As [LK14] noted, in the PDTB corpus, implicit and explict discourse markers combined convey on average 3.05 relations. If we only consider explicit discourse relations, as in our work, this number drops to about 2.6 in the PDTB.

Because of this ambiguity of discourse markers, we wanted to investigate how specific discourse markers are used to signal different discourse relations across different complexity levels. To do so, we identified the set of relations that

each discourse marker conveys, then, the distribution of all discourse markers across regular and simple versions has been computed. We used entropy in order to calculate the information of each distribution; then, used cross entropy to measure the difference between the distributions [DBKMR05, Jay57, CT12]. Formula 5 was used to calculate the entropy of the distribution of each discourse marker (noted as H(x)) across different complexity levels. Each discourse marker is considered as a random variable, x. The range of values that x can take, noted as r_i in Formula 5, are the possible discourse relations that the discourse marker can signal. For example, using the discourse marker *while* of Figure 11, the discourse marker x is *while*, p(r_1) is the probability that the discourse marker *while* is used to signal the CONTRAST relation which is 0.706 in the regular version as opposed to 0.676 in the simple version. Similarly, p(r_2) is the probability that the discourse marker *while* signals a SYNCHRONOUS relation.

$$H(x) = H(p) = -\sum_i p(r_i) log(p(r_i)) \tag{5}$$

Once the entropy of each distribution was computed, we compared the distributions in order to evaluate if there is a significant change in the distribution of discourse markers. To do so, we have used cross entropy. Formula 6 has been used for calculating the cross entropy for a specific discourse marker called x. To compare two distributions using cross entropy, we assume that the first argument (*reg*) is the target probability distribution, and the other one (*simp*) is the estimated distribution that we are trying to compare against. The closer the cross entropy is to the entropy of the target distribution, the less the change in the distribution of the specific discourse marker across complexity levels. In our experiment, *reg* stands for the regular version and $p((r_i)_{reg})$ is the probability that the discourse marker x, signalling the i^{th} relation in the regular version; while *simp* stands for the simple version and $p((x_i))_{simp}$ is the probability that

the discourse marker x, signals the i^{th} relation in the simple version.

$$H(reg, simp) = -\sum_{i} p((x_i)_{reg}) log(p((x_i))_{simp}) \quad (6)$$

The 5 discourse markers that show the most differences in the distribution of discourse relations are *in fact*, *although*, *though*, *while* and *since*. As shown in Figure 12, the discourse marker *in fact* can be used to convey 7 different relations: INSTANTIATION, CONTRAST, CONJUNCTION, CONCESSION, RESTATEMENT, CAUSE and ASYNCHRONOUS. Figure 12 shows the difference in usage of this marker across complexity level. In addition, the distribution of the discourse markers *although* and *though* both signalling CONCESSION and CONTRAST discourse relations are shown in Figures 13 and 14 respectively. As the figures show, both discourse markers are more frequently used to signal a CONCESSION in the simple version and a CONTRAST in the regular version. For example, the discourse marker *although* is used 54.4% of the time to signal a CONCESSION in simple texts as opposed to 50.0% in regular texts. However, both *although* and *though* are more frequently used to signal a contrast in the regular version than in the simpler version. Finally, Figure 15 shows the distribution of the discourse marker *since* to signal ASYNCHRONOUS and CAUSE discourse relations over the corpora. As Figure 15 shows, it is more probable that this discourse marker is used to signal a CAUSE across both versions rather than ASYNCHRONOUS; however, to signal an ASYNCHRONOUS relation, it is more common to use *since* in the simple version than in the regular version.

It is interesting to note that although discourse relations seem to be preserved across complexity levels (see Section 4.4.2), *how* discourse markers are used to signal these relations seems to vary across complexity levels.

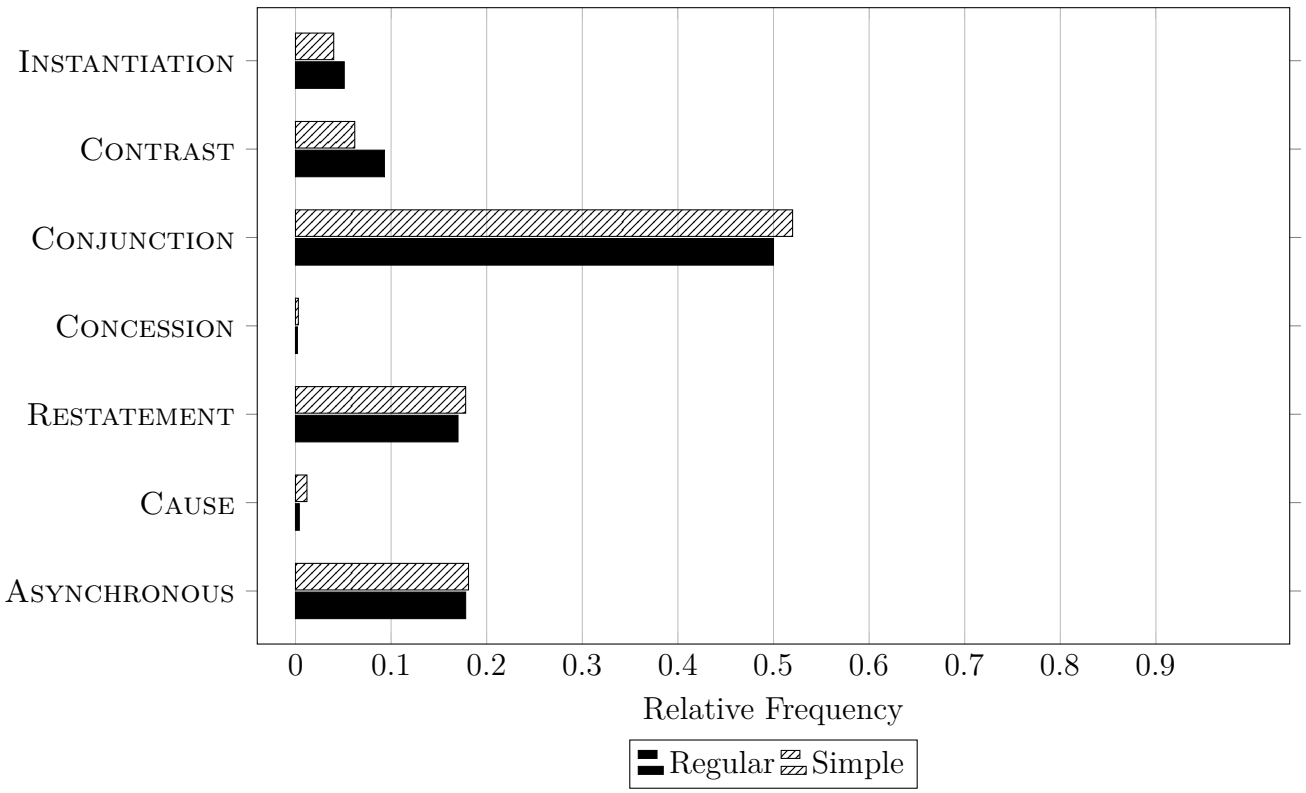


Figure 12: Distribution of the discourse marker *in fact* with respect to the discourse relations it signals across the Simple English Wikipedia corpus.

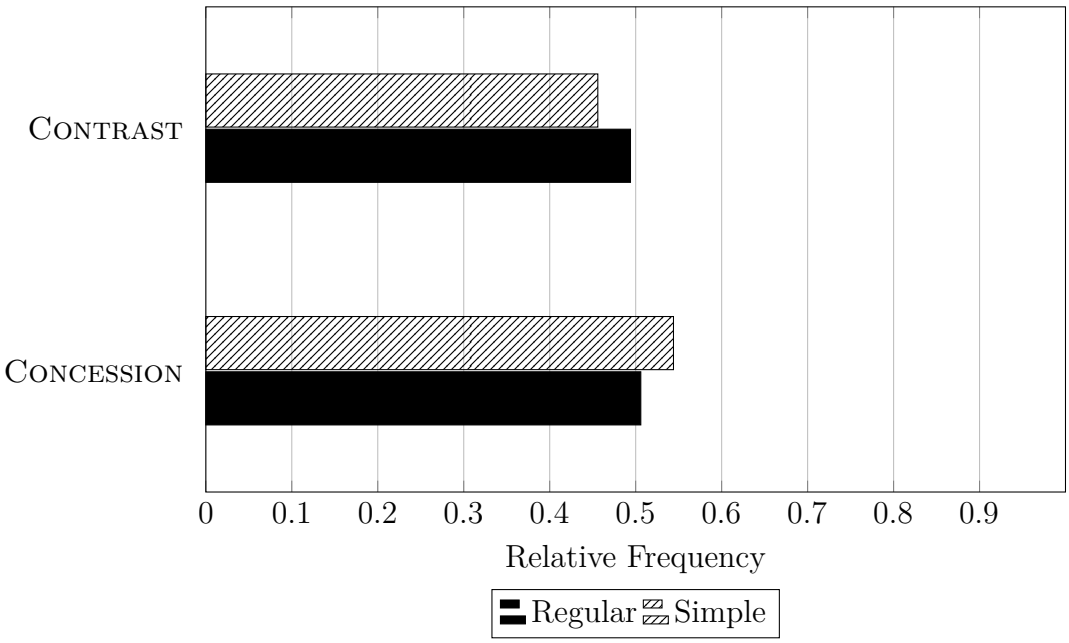


Figure 13: Distribution of the discourse marker *although* with respect to the discourse relations it signals across the Simple English Wikipedia corpus.

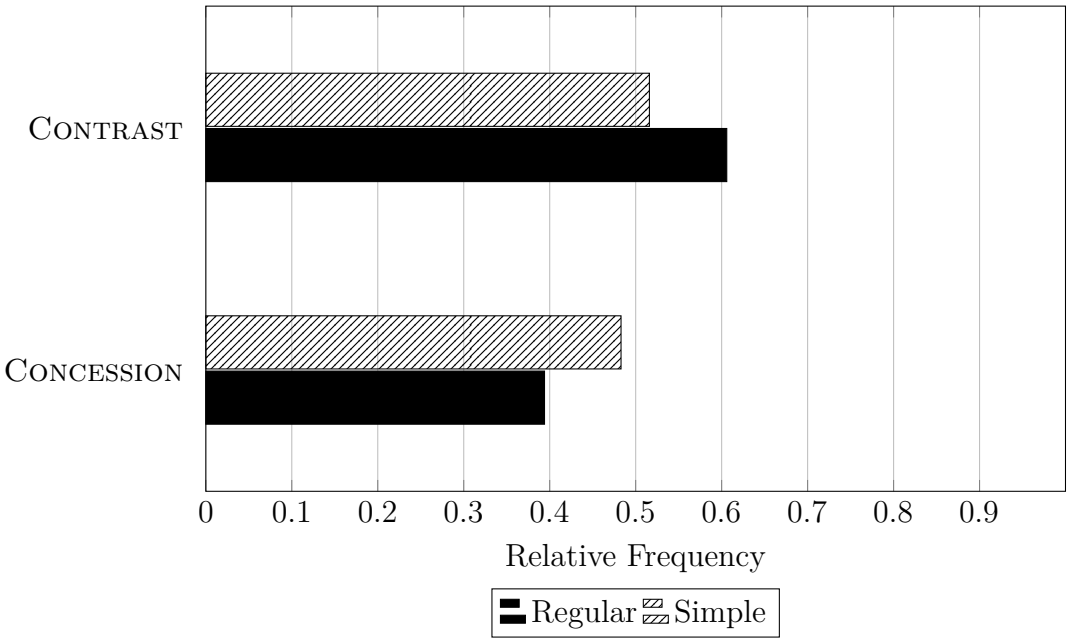


Figure 14: Distribution of the discourse marker *though* with respect to the discourse relations it signals across the Simple English Wikipedia corpus.

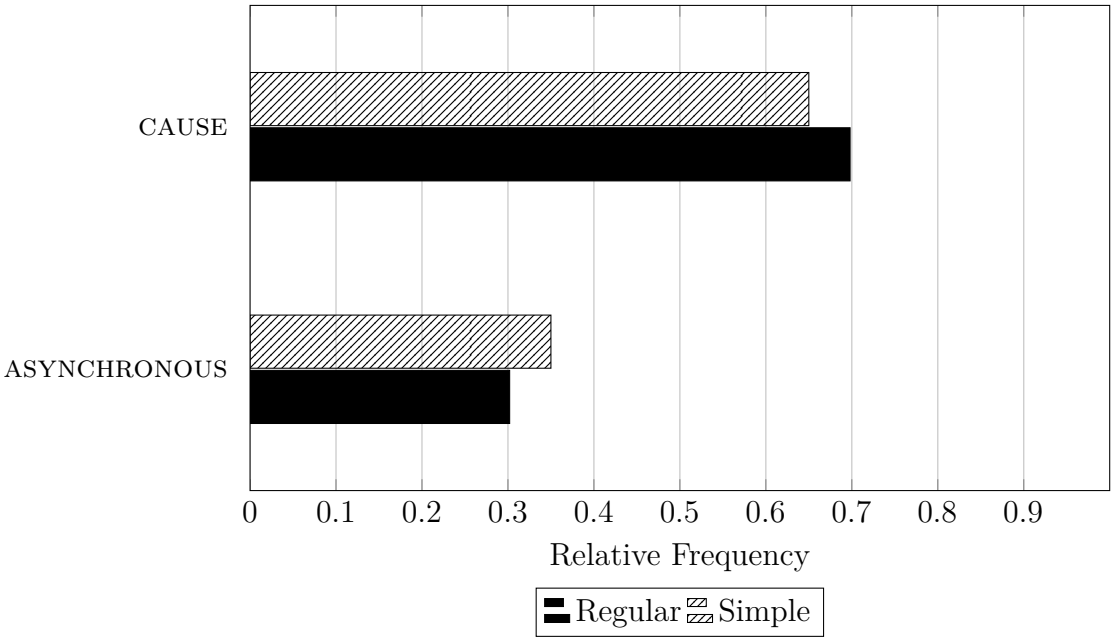


Figure 15: Distribution of the discourse marker *since* with respect to the discourse relations it signals across the Simple English Wikipedia corpus.

4.5 Conclusion

In this chapter, we addressed research question # 2 by performing an analysis of the usage of discourse relations as well as the usage and distribution of discourse markers across different complexity levels. Our analysis of the Simple English Wikipedia corpus shows that the changes in the distribution of explicit discourse relations across complexity levels is not statistically significant. However, the usage of discourse markers is different across the two complexity levels. In particular, we observed that the relative frequency of discourse markers is higher in more complex texts. Additionally, our analysis revealed that the distribution of discourse markers to convey specific relations is different across different complexity levels. These results seem to indicate that although the same logical and semantic information is conveyed in both simple and regular versions; how they are signalled is different.

Based on these results, we want to bring our investigation further to analyse the individual changes. Indeed we are curious that if the usage of discourse markers is different across complexity levels, what are other linguistic devices other than discourse markers can be used to signal a relation? The assumption of meaning preservation in the parallel corpora across complexity level allows us to investigate the changes in different ways a discourse relation can be represented.

In the next chapter we present our work toward the third research question and investigate how texts with different complexity levels can be used to automatically identify other lexicalized forms of discourse markers.

Chapter 5

Automatic Discovery of Alternative Lexicalizations Across Complexity Levels

Discourse markers are often used to signal the presence of a discourse relation. However the absence of a discourse marker does not necessary imply the absence of a discourse relation. In the PDTB, for example, 45.46% of the relations are marked explicitly, while 54.54% are not signaled via a discourse marker. The last two chapters have investigated the relation between text complexity and explicit relations. On the other hand, this chapter focuses on non-explicit relations. The goal of this chapter is to address research question # 3 and investigate how alternative lexicalizations of discourse markers can be automatically identified using parallel corpora at different complexity levels.

In Chapter 4, we noted that the usage of an explicit discourse marker to signal the existence of a discourse relations differs across complexity levels. Based on this observation, we were interested in using the PDTB inventory of discourse markers to explore alternative lexicalizations (AltLexes) which signal a discourse

relation. To do so, we used parallel corpora across complexity levels and the state of the art discourse parser. Our main goal is to propose an approach to identify alternative lexicalizations of discourse markers using monolingual corpora at different complexity levels and discourse parsers in identifying explicit relations. This can be beneficial for many natural language processing applications. For example, the performance of discourse parsers can be improved by identifying implicit and AltLex relations.

5.1 Introduction

Explicit discourse relations are signalled using a discourse marker; while *non-explicit* relations are not signalled by a marker but can still be inferred by the reader. In the Penn Discourse Tree Bank (PDTB) framework [PDL+08] these non-explicit relations include *implicit* and *AltLex* relations. AltLex or alternative lexicalization relations are signalled using an open list of lexical markers that are not part of PDTB inventory of 100 explicit discourse markers. AltLexes are defined as follows in the PDTB ([PDL+08], page 22):

> *"These are cases where a discourse relation is inferred between adjacent sentences but where providing an Implicit connective leads to redundancy in the expression of the relation. This is because the relation is alternatively lexicalized by some "non-connective expression". Such expressions include (1) those which have two parts, one referring to the relation and another anaphorically to Arg1; (2) those which have just one part referring anaphorically to Arg1; (3) those which have just one part referring to the relation."*

Example 8 shows an AltLex relation taken from the PDTB [PDL+08]. In Example 8, the text segment in *italic* represents Arg1, the segment in **bold** refers

Complex	Simple
These works he produced and published himself, whilst his much larger woodcuts were mostly commissioned work. [AltLex, CONTRAST]	He created and published his works himself, **but** his larger works were mostly commissioned work to be sold. [Explicit, CONTRAST]

Figure 16: No explicit relation is detected in the complex sentence (left), while an explicit CONTRAST relation is identified automatically in the simple sentence (right).

to Arg2 and the underlined expression shows the alternative lexicalization.

Example 8

And she further stunned her listeners by revealing her secret garden design method: *Commissioning a friend to spend "five or six thousand dollars...on books that I ultimately cut up."* **After that, the layout had been easy.** [AltLex, TEMPORAL]

According to [PLN09], discourse markers constitute strong clues to detect explicit relations, hence discourse parsers have typically relied on them as valuable features in order to identify explicit discourse relations automatically [LNK14]. Similarly, the presence of alternative lexical markers is a strong indicator of a relation; however since the list of such markers is open, identifying them is a challenge.

Figure 16 shows a pair of sentences that convey the same information; however only one sentence contains a discourse marker from the PDTB inventory[1]. Hence, a discourse parser using the PDTB inventory of markers would easily identify the explicit CONTRAST relation in the first sentence but will likely not tag the second sentence because *whilst* is not part of the PDTB inventory of

[1]The example is taken from the Simple English Wikipedia corpus [CK11a].

discourse markers. However, the writer's intention can be understood using a variety of linguistic and stylistic devices such as an alternative lexical marker (i.e. an AltLex), a change of tense, a structural signal, etc. Thus, discourse parsers can benefit from the automatic identification of AltLexes that can signal discourse relations.

Apart from discourse markers and alternative lexicalizations, other devices can signal the presence of a discourse relation. [TD13] introduced a list of such devices that can signal a relation in the RST discourse framework. Apart from discourse markers, these include:

1. **Entity features**: Links between similar or dissimilar entities can be used to signal a relation. Example 9[2] shows a LIST relation that is signalled through the use of the three similar entities (the company names) which are underlined.

   Example 9

   Earlier this year, Tata Iron Steel Co.'s offer of $355 million of convertible debentures was oversubscribed.
   Essar Gujarat Ltd., a marine construction company, had similar success with a slightly smaller issue.
   Larsen Toubro started accepting applications for its giant issue earlier this month.

2. **Semantic relations**: Semantic relations such as synonymy, antonymy, hypernymy, etc. between words can be used to signal discourse relations. Example 10 shows a CONTRAST relation signalled by the antonomy relation between the two words of *black* and *white*.

[2]This example is taken from [TD13].

Example 10

I wear a black shirt; my friend wears a white one.

3. **Lexical features**: Specific words (e.g. *concede*, *at the same time*, etc.) which are not considered as discourse markers can signal a discourse relation. This category is equivalent to the notion of AltLexes in the PDTB.

4. **Morphological features**: The tense of verbs is often used to signal a temporal relation. Example 11 shows a TEMPORAL relation where the shift in the tense of the verbs signals the relation.

   Example 11

   The children were playing in the yard; it started raining, they are in their rooms now.

5. **Syntactic features**: Some syntactic patterns can be used as a signal of a discourse relation. For example, [TD13] pointed out that subject-verb inversions can be used as a signal of CONDITION. Example 12 shows such a case.

   Example 12

   Should you need further information, please contact us.

6. **Graphical features**: Punctuations are an example of graphical features that can be used as a signal of a discourse relation. In Example 13, the semi-colons are used to signal a LIST relation.

   Example 13

   I like jogging; my friend likes swimming;...

7. **Numerical elements**: These features are often used as a signal of a LIST relation, as in [TD13]. Example 14 shows such a case.

   Example 14

   (a) Remove pizza from all packaging and shrink wrap.

(b) Set oven rack to middle position and preheat oven to 450°F.

(c) Place pizza on middle rack.

(d) Bake for 8-12 minutes.

8. **Genre features**: Genre can be used as a cue to provide some indication of a discourse relation. Other researches also showed that textual genre influences on the distribution of discourse relations [Web09, BDK14, DKB+16].

Our contribution in this chapter focuses on the discovery of new **lexical features** according to [TD13]'s categories of devices.

5.2 Background

Discourse markers [Bla87, KD94, Sch85] are the most informative signals of explicit discourse relations [PLN09]. However, they are not well-defined in linguistics. [Lev83] defined discourse markers as words and phrases such as *after all*, *actually*, *still*, etc. that connect an utterance to the prior discourse. [Zwi85] considered discourse markers as a class of particles, but does not specify what particles are considered as discourse markers. [Sch88] also defined discourse markers as words that connect dependent textual units in a discourse. According to her, discourse markers do not belong to any linguistic class and, except for a few discourse markers such as *oh* and *well*, most carry meaning. [Red91] revised Schiffrin's definition; even though she agreed that discourse markers have meaning by themselves, she argued that they should contribute to the semantic interpretations of the discourse by either linking consecutive sentences or the current sentence to the context. Apart from research efforts aiming at defining discourse markers, another line of research has focused on providing a list of discourse markers in English (e.g. [PDL+08, Abr91, And01, Bla02, Fis00, SSN92, Kno96])

and other languages (e.g. [Pas03, Tra05]). While most of these inventories have been built by hand, some work has attempted to identify them automatically. [LK14], for example, used the Europal parallel corpus and collocation techniques to induce French markers from their English counterparts. Following this work, [HM16] built a parallel corpus of causal and non-causal AltLexes using word alignment with the PDTB discourse markers as initial seeds. Our work is different from these as we used already existing parallel corpora in text simplification and extracted discourse information automatically using a discourse parser. In addition, instead of focusing on a single relation as [HM16] did, we generalize the problem to all PDTB discourse relations. We also used external resources which have been shown to have advantages over word alignment [Ver10] in similar tasks. Lastly, the PDTB AltLexes only capture inter-sentence relations. Our contribution overcomes this limitation by identifying intra-sentence discourse relations.

5.3 Discourse Markers Across Complexity Levels

As noted in Chapter 2, the differences in complexity level may be the result of various linguistic choices: at the lexical level (e.g. using frequent vs. abandoned words), at the syntactic level (e.g. using active vs. passive voice) or even the discourse level (e.g. using an implicit vs. an explicit discourse relation). The main assumption in text simplification is that it is possible to reduce a text's complexity while preserving its meaning as much as possible. We showed in Chapter 4 that the lexical realization of discourse relations (i.e. explicit versus non-explicit) and the choice of a discourse marker (e.g. *but* versus *however*) may change across complexity levels. As noted in Chapter 4, the meaning and

Complex	Simple
When the show was broadcast, Rupert Boneham won the million dollars. [Explicit SYNCHRONY]	Rupert Boneham won the million dollars.

Figure 17: Example of the removal of a discourse argument and consequently the removal of a discourse relation.

consequently the discourse relations can be assumed to be preserved during text simplification. The removal of a discourse relation may happen if the discourse argument is considered non-essential. For example, Figure 17 shows a pair of aligned sentences where the complex version contains an explicit SYNCHRONY relation signalled by *when*; while the discourse argument and consequently the explicit discourse marker has been removed in the simple version. Hence, given a sentence and its simplified version, three phenomena can occur:

1. a discourse marker is replaced by another (e.g. *although* $\Rightarrow$ *though*),
2. a discourse marker is replaced by another lexical device (i.e. word or phrase) which is not considered a discourse marker in the inventory used (e.g. *because* $\Rightarrow$ *the reason for this*), or
3. a discourse marker is removed completely.

In cases (1) and (2) above, the discourse relation is preserved, while in case (3) the discourse relation is either removed or changed to an implicit relation.

To automatically identify AltLexes, we only focused on case (2).

5.4 Data Sets

To discover new lexical markers, we again created two sentence-aligned data sets using the Simple English Wikipedia corpus [CK11a] and the Newsela corpus [XCBN15]. For the Simple English Wikipedia corpus, we used the sentence-aligned section which contains 167,686 pairs of aligned sentences.

In order not to overfit to a specific corpus, in addition to the Simple English Wikipedia corpus, we also used the Newsela (News) corpus [XCBN15]. However, this corpus is not sentence-aligned. Thus, we used the original article and its 4 simplified versions and aligned the corpus at the sentence level. To do so, we used an approach similar to [CK11a] to align sentences automatically. To evaluate the alignments, we then asked two native English speakers to evaluate them manually. The Kappa inter-annotation agreement was 0.898 computed on 100 randomly chosen alignments. The next section will describe this work in more detail.

5.4.1 Sentence Alignment of the Newsela Corpus

Sentence alignment has a long tradition in statistical machine translation (e.g. [BCP+90, ON04, Moo02]). In the context of text simplification, the objective of sentence alignment is to generate a data set containing pairs of sentences that convey the same information but where one is more complex than the other one. This task is also known as monolingual sentence alignment (e.g. [BE03, NS06, QBD04]). The aligned sentences convey the same essential meaning; however, some details may be added or removed. The Newsela corpus is already article-aligned and we observed that the flow of information is the same in the simpler versions of this corpus. Thus, we used this property to develop our sentence alignment algorithm. Our sentence alignment algorithm, inspired by [CK11b], is

based on the TF-IDF (Term Frequency-Inverse Document Frequency) measure. However, matching sentences using TF-IDF ignores the ordering of information in the aligned article. In order to deal with this limitation, we focused on the locality of the alignment. This means that we expect to have the best alignment in relatively the same location in both versions (i.e complex and simple). Locality is considered by restricting the search for the best possible alignment to the same sentence index +/- i sentences in the simpler version. There is a trade-off between the precision of the alignment and the number of alignments generated depending on the value of the locality factor, i. A small locality factor entails that we restrict the algorithm to choose the best alignment (if any) from a small number of candidate solutions. This may lead to sparsity in the solution set. On the other hand, a large locality factor allows the aligned sentence(s) to be located at different relative positions in the aligned articles. As a result, the alignment may not be precise enough. In our experiments we choose the locality factory $i = 3$. This means that, for each sentence or two consecutive sentences in a more complex article, we look for the best alignment (based on the closest cosine similarity) across all of its simpler versions, from the index of three sentences before to three sentences after the index of the sentence in the complex version. In addition, during the simplification process, one sentence may be split into multiple sentences or multiple sentences may be merged together to form a single sentence. Instead of considering the general case of m-to-n alignments, we only considered 1-to-2 and 2-to-1 alignments as these seem to account for the bulk of the corpora. Algorithm 1 shows our sentence alignment algorithm. As shown in Algorithm 1, we first split all the 1,191 original Newsela articles and their corresponding simplified versions into individual sentences (*Step 3*). Each sentence in each original article and its four corresponding simpler versions is considered as a document from which we build a $TF * IDF$ model (*Step 5*).

Then, a score is computed for candidate sentence pairs $< sentence_c, sentence_s >$ such that $sentence_c$ and $sentence_s$ are candidates for sentence alignment taken from a more complex article and a simpler aligned article respectively (*Step 6-9*). The best alignment is then chosen from all 1-to-1, 2-to-1 (i.e. two consecutive complex sentences aligned to one simple sentence) and 1-to-2 (i.e. one sentence aligned to two consecutive simple sentences) alignments which are found based on our locality assumption. For 2-to-1 and 1-to-2 alignments, we considered two cases where the two consecutive sentences are: (1) the current and the next sentence and (2) the current and the previous sentence. The best alignment is chosen according to the cosine similarity measure between the feature vectors. The maximum similarity score for i^{th} sentence in article a_c is denoted as $score(i)$. Since the algorithm forces all sentences in the set of complex articles (a_c) to be aligned, we filtered the resulting alignments to remove those with a low similarity score (below a fixed threshold of 0.6[3]) and the alignments of identical sentences as they do not contain any simplification. Thus, we removed all alignments with a similarity score < 0.6 and 1.0.

To evaluate the quality of the alignments, we randomly choose 100 alignments (i.e. pairs of sentences) and asked two human annotators to label them such that if the meaning is preserved in the alignment, they annotated the pair as *"to be aligned"*, otherwise *"not to be aligned"*. The Kappa inter annotator agreement [Coh60] between the annotators was then calculated using Formula 7. In this formula, p_o is the probability of observed agreement and p_e is the probability of expected agreement.

$$kappa = \frac{p_o - p_e}{1 - p_e} \tag{7}$$

The Kappa value between the two annotators was 0.898.

[3]This value was set experimentally.

	SEW-based DS	News-based DS
# of alignments	167,686	36,768
Average # of words per sentence	21.71	17.05

Table 13: Statistics of the data sets used in automatic discovery of alternative lexicalizations of discourse markers. For the News-based data set, the average number of words per sentence is calculated considering the entire corpus.

In total, after the alignment, the Newsela-based data set contains 36,768 pairs of sentences. Table 13 summarizes statistics of the data sets used.

5.5 External Resources

Laali and Kosseim [LDK15] used statistical methods to discover markers across parallel corpora, however [Ver10] showed that the use of external lexical resources has many advantages for the projection of annotations. Inspired by [Ver10], we used external resources to identify alternative lexicalizations of the PDTB inventory of discourse connectives. The two resources that we used are: (1) the paraphrase database (PPDB) [GVCB13] and (2) WordNet [Mil95] which are described in the following sections.

5.5.1 The Paraphrase Database

The paraphrase database (PPDB) [GVCB13] contains over 220 million paraphrases which consists of 73 million phrasal, 8 million lexical (single word to single word) and 140 million syntactic paraphrases. The paraphrase[4] database comes in six sizes from S to $XXXL$. The smaller versions of the paraphrase

[4]Available at `http://www.cis.upenn.edu/~ccb/ppdb/`

Algorithm 1 Sentence Alignment of the Newsela corpus. The alignments (either 1-to-1, 1-to-2 or 2-to-1) with the maximum similarity score calculated in line 11 are included in our data set.

1: **procedure** SENTENCEALIGNMENT
2: **for** each set of aligned articles a_{all} **do**
3: Split the original and 4 levels of complexity into sentences
4: $sentence_{all} \leftarrow$ *Set of all sentences in Step 3*
5: Build a $TF * IDF$ model for $sentence_{all}$
6: **for** each article $a_c \in a_{all}$ **do**
7: **for** each article $a_s \in a_{all}$ such that complexity of $a_c >$ complexity of a_s **do**
8: $sentence_{total} \leftarrow$ *Total number of sentences in* a_i
9: **for** $sentence_i \in a_c$ such that $i = 0$ *to* $sentence_{total}$ **do**
10: **for** $sentence_j \in a_s$ such that $j = i - 3$ *to* $i + 3$ **do**
11:

$$score(i) = max \begin{cases} sim(sentence_i, sentence_j) \\ sim(sentence_i + sentence_{i+1}, sentence_j) \\ sim(sentence_{i-1} + sentence_i, sentence_j) \\ sim(sentence_i, sentence_{j-1} + sentence_j) \\ sim(sentence_i, sentence_j + sentence_{j+1}) \end{cases}$$

dictionary contain more precise paraphrases, i.e. with higher confidence scores; while the larger versions have more coverage. We choose the version in the middle of this range (the PPDB L version) as we give the same weight to the precision and coverage of paraphrases.

5.5.2 WordNet

WordNet [Mil95] is a lexical database which groups the four classes of words (nouns, verbs, adjectives and adverbs) into sets of cognitive synonyms which are called synsets. It contains 117 thousand synsets which are linked to each other by semantic relations such as *hyperonymy*, *hyponymy*, *meronymy*, *troponyms*, *antonymy*, etc. As opposed to the paraphrase database (PPDB) which is constructed automatically using parallel texts, WordNet was built manually.

5.6 Methodology

According to the PDTB framework, each AltLex can be substituted with at least one discourse marker [PDL$^+$08]. Based on this, to discover AltLexes automatically, we first parsed both sides of the aligned sentences of both data sets to extract discourse information. This was done using the PDTB-style End-to-End parser [LNK14]. Because it uses the PDTB framework, the parser uses the inventory of 100 discourse markers from the PDTB. The result of this tagging was categorized into one of the following cases:

1. *NonExp-NonExp*: a non-explicit[5] discourse relation occurs in both sentences.
2. *Exp-Exp*: the same discourse relation and discourse marker occur in both sentences.

3. *NonExp-Exp*: a non-explicit relation occurs in the complex sentence, but an explicit one is used in the simple sentence.

4. *Exp-NonExp*: an explicit relation occurs in the complex sentence, but no relation is used in the simple sentence.

5. *Other*:

   (a) *Same Relation-Different marker*: the same explicit relation is used but with different discourse markers in both sentences.

   (b) *Different Relation-Different marker*: a different explicit relation and a different discourse marker are used.

   (c) other cases including several explicit relations within a single sentence.

Table 14 shows the frequency of these transformations in the two data sets. As can be seen in Table 14, no relation is identified by the parser in almost 70% of the alignments of the SEW-based data set, while this percentage is 50% in the alignments of the Newsela-based data set. In 11.79% alignments of the SEW-based data set and 7.23% alignments of the Newsela-based data set the same explicit relation is occurred. In 4.69% and 3.07% of the alignments of the SEW-based data set and Newsela-based data set respectively, no explicit relation is identified in the complex part of the alignment; while an explicit relation is identified in the simple part. Similarly in 5.65% and 4.71% of the alignments of the SEW-based data set and Newse-based data set, an explicit relation is identified in the complex part of the alignment and no explicit relation is automatically identified in the simple part.

To discover AltLexes, we only considered cases (3) and (4) where one side of the aligned sentences includes one and only one PDTB discourse marker and the

[5]Recall from Chapter 2 that a non-explicit discourse relation can refer to an implicit or an AltLex discourse relation or to no discourse relation.

Discourse-level Change	SEW-based DS		News-based DS	
(1) NonExp-NonExp	116,852	69.68%	18,384	50.00%
(2) Exp-Exp	19,735	11.76%	2,660	7.23%
(3) NonExp-Exp	**7,868**	**4.69%**	**1,129**	**3.07%**
(4) Exp-NonExp	**9,490**	**5.65%**	**1,733**	**4.71%**
(5) other	13,741	8.22%	12,862	34.99%
Total	167,686	100%	36,768	100%

Table 14: Frequency of the discourse changes across complexity levels in the SEW-based and News-based data sets.

other side includes no marker at all. This gave rise to a total of 20,220 aligned sentences. We then used the two external resources described in Sections 5.5.1 and 5.5.2: the paraphrase database (PPDB) [GVCB13] and WordNet [Mil95]. We took the discourse marker from the explicit side and looked for an alternative lexicalization (a synonym or paraphrase) in either WordNet or PPDB. If any of its alternative lexicalization appeared in the non-explicit side, we considered it as a candidate AltLex to signal the relation. We then replaced this AltLex with the explicit discourse marker from the explicit side and parsed the new sentence with the PDTB-style End-to-End parser again. This process is shown in Example 16. On average, each discourse marker was replaced by 23.2 alternative lexicalizations taken from the PPDB and 12.3 from WordNet.

If the parser detected the same relation (see Example 16), then the potential marker was considered as an AltLex.

Example 15

Complex: It's a very special place **because** this site, this area, has

been tied to the history and life of African-Americans since about the early 1800s. [Explicit CAUSE]

Simple: It has been tied to the history and life of African-Americans *since* [SYNONYM OF BECAUSE] about the early 1800s.

⇓

Simple after substitution: It has been tied to the history and life of African-Americans **because** about the early 1800s.

On the other hand, because the End-to-End discourse parser uses both the discourse marker and syntactic features, if it was not capable of detecting the discourse relation in the replaced sentences, we concluded that either (1) the relation existed, but the parser could not detect it, (2) the AltLex does not signal the discourse relation or (3) the discourse relation does not exist (see Example 15). Because we did not use any syntactic filter, the replacement of the discourse marker may alter the syntax of the sentence such that the parser is unable to detect the relation. This is why, regardless of the reason, if the parser was not able detect the relation, we discarded the AltLex.

Example 16

Complex: Today, the comic arm of the company flourishes *despite* [SYNONYM OF THOUGH] no longer having its own universe of super powered characters.

Simple: Today, the company does very well even **though** they do not have their own universe of super powered characters. [Explicit CONTRAST]

⇓

Complex after substitution: Today, the comic arm of the company flourishes **though** no longer having its own universe of super

powered characters. [Explicit CONTRAST]

5.7 Results and Analysis

Table 15 shows the number of sentence alignments mined and the number of potential AltLexes (i.e. token count) identified in each data set for each level 2 PDTB relation. Overall, by mining 17,358 NonExp-Exp and Exp-NonExp alignments, the SEW-based data set allowed the discovery of 79 AltLexes from the PPDB and 8 from WordNet; whereas, the Newsela-based data set, providing only 2,862 alignments, allowed the discovery of 28 AltLexes from PPDB and 11 from WordNet. Using both corpora and both lexical resources, the method found 91 AltLexes. Appendix B shows the complete list of AltLexes found. Examples 17-23 show alignments containing a discourse marker in one side and an AltLex found in the other side. The discourse marker in the explicit relation is shown in *italic*, while the AltLex found in the other side of the alignment is underlined.

Example 17

Complex: Puerto Rico *or*, officially the Commonwealth of Puerto Rico, Associated Free State of Puerto Rico, is an unincorporated territory of the United States, located in the northeastern Caribbean Sea, east of the Dominican Republic and west of the Virgin Islands. [Explicit, ALTERNATIVE]

Simple: Puerto Rico, also known as the Commonwealth of Puerto Rico, is a territory or colony of the United States in the Caribbean Sea.

Example 18

Complex: Shortly *thereafter*, by chance , Bell came across the violin again and discovered it was about to be sold to a German industrialist to become part of a collection. [Explicit, ASYNCHRONOUS]

Simple: Shortly afterwards, by chance, Bell came across the violin again and discovered it was about to be sold to a wealthy German to become part of a collection.

Example 19

Complex: Archer resigned in October 1986 due to a scandal caused by an article in The News of the World, which led on the story Tory boss Archer pays vice-girl and claimed Archer had paid Monica Coghlan, a prostitute,...

Simple: Archer had to resign *because* of a scandal in October 1986 when the Sunday newspaper The News of the World led on the story Tory boss Archer pays vice-girl. The article claimed that Archer had paid Monica Coghlan, a prostitute,...[Explicit, CAUSE]

Example 20

Complex: Today, the comic arm of the company flourishes despite no longer having its own universe of super powered characters.

Simple: Today, the company does very well even *though* they do not have their own universe of super powered characters. [Explicit, CONCESSION]

Example 21

Complex: Scotland is similarly divided into zones by the A7, A8 and A9 which radiate out from Edinburgh.

Simple: Scotland is *also* divided into zones by the A7, A8 and A9 leave from Edinburgh. [Explicit, CONJUNCTION]

Example 22

Complex: As a result of the development of these two airports, Tempelhof was closed in October 2008, *while* Gatow is now home of the Museum of the German Luftwaffe and a housing development. [Explicit, CONTRAST]

Simple: As a result of these two airports Tempelhof is being close, whilst Gatow no longer serves as an airport and now hosts the Museum of the German Luftwaffe.

Example 23

Complex: Almost *simultaneously*, influential British critic Q. D. Leavis argued in Critical Theory of Jane Austen 's Writing, published in Scrutiny in the early 1940s, that Austen was a professional, not an amateur, writer. [Explicit, SYNCHRONY]

Simple: Almost at the same time, British critic Q. D. Leavis printed Critical Theory of Jane Austen's Writing in Scrutiny in the early 1940s.

It is interesting to note that, overall, the approach did not find any alternate lexicalizations for some relations such as LIST or EXCEPTION and only one for CONDITION. It is not clear if this is because these relations are typically signalled using a rather fixed inventory of discourse markers or because of the low number of such alignments. Indeed, in the PDTB, out of 624 tagged AltLex relations, only 6 are labeled as RESTATEMENT, 1 as EXCEPTION and 2 as CONDITION.

On the other hand, relations such as CONJUNCTION, ASYNCHRONOUS and CAUSE provided a large number of alignments from which we identified a variety of AltLexes. For example, the PPDB identified "*caused by*", "*resulting*", "*causing*", "*this being so*", etc. as AltLexes to signal a CAUSE relation.

In addition, as can be seen in Table 15, the number of potential AltLexes coming from the PPDB is greater than the number of AltLexes coming from WordNet. One reason can be the difference of the coverage of these two resources as WordNet is smaller than the PPDB. Another reason is that each word in the PPDB has a list of paraphrases with various syntactical classes. Thus, if the syntactic class of a discourse marker is changed in the simplification process, it is more probable that the PPDB covers more syntactical variations of the discourse marker compared to WordNet. For example, Example 24 shows an example taken from the Newsela corpus. In this example, the discourse marker *before* that signals an ASYNCHRONOUS relation, in the complex version is tagged as *IN* (i.e. subordinating conjunction in PDTB tag set). In the paraphrase database, *used to* is one of the paraphrases of the discourse marker *before*. In the simple version of this example, the verb *used to* is signalling the same relation (i.e. ASYNCHRONOUS relation) which is captured as an AltLex.

Example 24

Complex: Now they have drones in 15 states, including California and Texas. **Before** they started the business, the two covered fields on foot or in vehicles. [Explicit ASYNCHRONOUS]

Simple: Now they have drones in 15 states, including California and Texas. Fiene used to check farm fields on foot or with vehicles.

5.8 Conclusion

One of the fundamental assumptions in text simplification is meaning preservation. As discussed in Chapter 2, we can assume that discourse relations are preserved across complexity levels, however shown in Chapter 4, the difference

Discourse Relation	SEW-based Data set			News-based Data set			Overall	
		New AltLexes from			New AltLexes from			New markers from
	Alignments	**PPDB**	**WordNet**	**Alignments**	**PPDB**	**WordNet**	**Alignments**	**SEW ∪ News**
Asynchronous	2,561	15	3	327	6	2	2,888	20
Synchrony	1,990	2	1	395	0	0	2,385	3
Cause	1,359	18	1	256	3	0	1,615	19
condition	296	0	0	141	1	0	437	1
contrast	2,568	6	1	667	5	6	3,235	9
concession	393	3	0	64	0	0	457	3
conjunction	7,738	25	1	914	12	3	8,652	27
instantiation	159	3	1	33	0	0	192	3
restatement	63	1	0	13	0	0	76	1
alternative	220	5	0	51	1	0	271	5
exception	8	0	0	0	0	0	8	0
list	3	0	0	1	0	0	4	0
Total	**17,358**	**79**	**8**	**2,862**	**28**	**11**	**20,220**	**91**

Table 15: Number of *Exp-NonExp* and *NonExp-Exp* alignments and newly identified AltLexes.

in the usage of discourse markers is significant across complexity levels. Based on these observations, in this chapter we addressed research question #3 by exploring the use of parallel texts at different complexity levels to automatically identify alternative lexicalizations of discourse markers. Our main contribution in this chapter is to propose an approach for the automatic discovery of lexical features that can signal discourse relations. Since our proposed methodology is corpus-based, applying the method on different parallel corpora should reveal different AltLex expressions and be used to improve the performance of discourse parsers.

Chapter 6

Conclusion

Through this thesis, we have used computational methods to explore the relation between discourse and textual complexity. Specifically, we have first measured the influence of discourse-level properties in text complexity assessment (see Chapter 3). To do so, we designed and developed a supervised machine learning model using traditional linguistic features as well as discourse features for pairwise text complexity assessment. Then, through a corpus study (see Chapter 4), we have investigated the influence of textual complexity on discourse-level linguistic choices. Finally we used parallel corpora at different complexity levels to automatically extract alternative lexicalizations of discourse markers (see Chapter 5).

In this chapter, we summarize the main contributions of our thesis. Then, we conclude the chapter by discussing future research directions.

6.1 Main Contributions

This thesis made three main contributions to the field of Natural Language Processing:

1. Measuring the influence of discourse-level linguistic choices on computational complexity assessment (Chapter 3).
2. Identifying the influence of textual complexity on specific discourse-level linguistic choices (Chapter 4).
3. Developing an approach for the automatic discovery of alternative lexicalizations of discourse markers by leveraging parallel corpora at different complexity levels (Chapter 5).

These contributions involve both theoretical and practical aspects.

6.1.1 Theoretical Contributions

Influence of Discourse-level Properties on Computational Complexity Assessment

In Chapter 2, we introduced the problem of computational complexity assessment and tried to differentiate it from readability prediction. The main difference between these two notions is the consideration of the reader; computational complexity assessment focuses on the linguistic characteristics of the text rather than characteristics of the reader. We have shown that the contribution of discourse features on complexity assessment have not been analysed much in previous research. To address this, in Chapter 3, we developed a supervised model using a combination of 16 features, grouped into five classes: surface, lexical, syntactic, cohesion and coherence. We experimented with two data

sets created from standard corpora. Our results showed that for the task of pairwise complexity assessment, the use of only 4 coherence features performed statistically as well as using all 16 features. In addition, removing the class of coherence features from the combination of all features led to a statistically significant decrease in accuracy. Thus, **our first theoretical contribution is to show empirically the existence of a strong correlation between text coherence and text complexity**. This contribution was published in [DK16c].

Influence of Textual Complexity on Discourse-level Linguistic Choices

In Chapter 4, we studied the influence of text complexity on specific discourse-level linguistic choices. To do so, we performed a corpus analysis of the Simple English Wikipedia corpus to analyze the usage of discourse relations as well as the distribution of discourse markers across complexity levels. **Our second contribution is to show empirically that although discourse relations seem to be preserved across complexity levels, *how* discourse markers are used to signal these relations seems to vary across complexity levels.** In particular, our results seem to contradict those of [WRO03] who, after a manual analysis of three texts, concluded that simpler texts tend to use more explicit relations compared to more complex ones. Our larger scale analysis showed that, statistically, simpler texts use as many explicit relations as more complex ones. In addition, we studied the case of discourse markers that signal multiple discourse relations to investigate if their usage was influenced by the complexity level. We compared the distributions of discourse markers signalling multiple relations and using cross entropy, we measure how consistent their usage is across complexity levels. We identified five discourse markers which have the largest difference in their distribution to signal multiple relations

across different complexity levels.

Automatic Discovery of Alternative Lexicalizations of Discourse Markers

As discussed in Chapter 2, discourse relations are typically preserved across complexity levels, however as shown in Chapter 4, the difference in the usage of discourse markers can differ across complexity levels. Using these observations, in Chapter 5, we used parallel corpora at different complexity levels to automatically identify alternative lexicalizations of discourse markers. **Our third theoretical contribution is the development of an approach to automatically identify alternative lexicalizations of discourse markers using parallel corpora at different complexity level.** When applied to the Simple English Wikipedia corpus and the Newsela corpus, our proposed approach found 91 new AltLexes.

6.1.2 Practical Contributions

In addition to the theoretical contributions of Section 6.1.1, we also developed several practical applications that can be used as standalone NLP tools or can be embedded in larger applications.

Development of a System for Textual Complexity Assessment

To investigate the influence of discourse-level properties for pairwise complexity assessment (see Chapter 3), we developed a system to predict if a pair of texts have the same complexity levels or not. The system, written in Java, is based on Weka[1] and can be used as a standalone application. This system may be

[1] `http://www.cs.waikato.ac.nz/ml/weka/`

useful for readers looking for texts that have a *similar* complexity levels as other texts they know they understand.

Development of a System for Complex Word Identification

As noted in Section 1.3.2, we participated in the 2016 NAACL-SemEval international shared task on complex word identification [BSC+16]. To do so, we developed a system to predict if a given word in a given context is simple or complex. The system described in [DK16b, DK17a] ranked 21^{st} out of 45 at the competition. It can be used to identify complex words in an article, webpage, etc. that need to be simplified. This application can therefore be used as a pre-processing module to a text simplification system (e.g. [Sid06, Sid14, DT98a, CMC+98a, BBE11, CS97, Kau13]) or as a standalone application to highlight difficult words in a text.

Development of a System for Quality Assessment for Text Simplification

As noted in Section 1.3.2, we also participated in the 2016 LREC-QATS (Quality Assessment for Text Simplification) international shared task [ŠPS+16]. To measure the quality of automatic text simplification, we have developed a system with four main components, each focused on measuring a specific textual aspect: grammatical correctness, meaning preservation, simplicity and overall quality. Our system described in [DK16a] can be used to evaluate the output of automatic text simplification tools such as [SDOCJ+10, BT13].

Development of the CLaC Discourse Parser

As noted in Section 1.3.2, we have also participated in the 2015 CoNLL international shared task on shallow discourse parsing [XNP+15]. Our parser described

in [LDK15], ranked as 6^{th} out of 17 parsers. The parser was then improved by [LCK16] and used to participate in the 2016 edition of the task. It is now publicly available at `https://github.com/mjlaali/CLaCDiscourseParser`.

6.2 Future Work

Like everything else which needs a closing, this last section is the closing of this thesis. However, we still do not feel that we can call this thesis a "complete research work"; but I guess we never will. We tried to address specific research questions in the field of computational discourse analysis, but many others still need to be explored. In the rest of this chapter, we discuss some of the future research directions that we would like to investigate some day.

6.2.1 Improvement of Computational Complexity Assessment

In Chapter 3, we have shown that discourse-level properties can be used as an indicator of textual complexity in pairwise complexity assessment. Based on the model presented in Chapter 3, it would be interesting to build another model on top to predict the actual complexity level of a text. Currently, the research community does not have large-scale corpora at different complexity levels, hence pairwise complexity assessment is typically used to address this limitation. If we wish to address true textual complexity assessment, knowing that a pair of texts share the same complexity levels, we could determine only the complexity level of one of them. On the other hand, knowing that a pair of texts have different complexity levels, it would be easier to assign labels as the size of search space will be reduced.

6.2.2 Influence of Text Complexity on Discourse-level Choices-Complimentary Corpus Study

In the work presented in Chapter 4, we used the Simple English Wikipedia corpus to investigate the influence of text complexity on discourse-level choices. Our results challenge the previous work of [WRO03] who used a much smaller corpus to conduct a similar but manual corpus analysis. We used the Simple English Wikipedia corpus because at the time we performed the work it was the only large-scale corpus available. It would be interesting to expand the analysis of Chapter 4 with the Newsela corpus [XCBN15] to validate the results on a different dataset.

6.2.3 Improvements of the Automatic Discovery of Alternative Lexicalizations

Finally, it would be interesting to improve the quality of the automatically identified AltLexes by using a syntactic filter in order to replace potential markers only if they lead to syntactically correct sentences. Currently, we replace potential AltLexes with synonyms and paraphrases of discourse markers; however, the result of this replacement may lead to ungrammatical sentences (see Example 24 in Chapter 5). As a future work, we could use a syntactic filter to ensure that such a replacement is done only if it results in a grammatically correct sentence; otherwise another method should be used to verify if the synonym or paraphrase of a discourse marker is an AltLex.

In the proposed methodology, we extracted the AltLexes using the End-to-End discourse parser of [LNK14]. It would be interesting to try with another discourse parser, such as CLaC parser [LDK15] and compare the results. Finally,

the extraction of the AltLexes could be expanded using a bootstrapping approach. To do so, the PDTB inventory of discourse connectives could be used as initial seeds and in the first round of bootstrap, the methodology explained in Chapter 5 could be used to extract potential AltLexes. Then, the AltLexes that have at least one discourse usage could be added to the seeds of the previous step. This process can be repeated for a number of iterations to expand the list of extracted AltLexes.

Finally, another interesting line of research would be to evaluate to what degree the results of this thesis could be used to improve the performance of discourse parsers.

Bibliography

[AB92] Rajeev Agarwal and Lois Boggess. A simple but useful approach to conjunct identification. In *Proceedings of the 30th annual meeting on Association for Computational Linguistics (ACL)*, pages 15–21, Newark, Delaware, June 1992.

[ABE+12] Jamal Abedi, Robert Bayley, Nancy Ewers, Kimberly Mundhenk, Seth Leon, Jenny Kao, and Joan Herman. Accessible reading assessments for students with disabilities. *International Journal of Disability, Development and Education*, 59(1):81–95, 2012.

[Abr91] Werner Abraham. Discourse particles. *Amsterdam: Benjamins*, 1991.

[And01] Gisle Andersen. *Pragmatic markers and sociolinguistic variation: A relevance-theoretic approach to the language of adolescents*, volume 84. John Benjamins Publishing, 2001.

[BBE11] Or Biran, Samuel Brody, and Noémie Elhadad. Putting it simply: A context-aware approach to lexical simplification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT): short papers-Volume 2*, pages 496–501, Oregon, USA,

2011.

[BC10] Jill Burstein and Martin Chodorow. Progress and new directions in technology for automated essay evaluation. *The Oxford Handbook of Applied Linguistics*, pages 487–497, 2010.

[BCP+90] Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. A statistical approach to machine translation. *Computational Linguistics*, 16(2), 1990.

[BDK14] Félix-Hervé Bachand, Elnaz Davoodi, and Leila Kosseim. An investigation on the influence of genres and textual organisation on the use of discourse relations. In *Proceeding of the 15th International Conference of Computational Linguistics and Intelligent Text Processing (CICLing), LNCS-volume 8404*, pages 454–468. Springer, 2014.

[BE03] Regina Barzilay and Noemie Elhadad. Sentence alignment for monolingual comparable corpora. In *Proceedings of the 2003 conference on Empirical methods in natural language processing (EMNLP)*, pages 25–32, Sapporo, Japan, 2003.

[BFP87] Susan E Brennan, Marilyn W Friedman, and Carl J Pollard. A centering approach to pronouns. In *Proceedings of the 25th annual meeting on Association for Computational Linguistics (ACL)*, pages 155–162, Toronto, Canada, 1987.

[BG01] Alan Bailin and Ann Grafstein. The linguistic assumptions underlying readability formulae: A critique. *Language & Communication*, 21(3):285–301, 2001.

[BL08] Regina Barzilay and Mirella Lapata. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34, 2008.

[Bla87] Diane Blakemore. *Semantic constraints on relevance*. Oxford, 1987.

[Bla02] Diane Blakemore. *Relevance and linguistic meaning: The semantics and pragmatics of discourse markers*, volume 99. Cambridge University Press, 2002.

[BMM05] António Branco, Tony McEnery, and Ruslan Mitkov. *Anaphora processing: linguistic, cognitive and computational modelling*, volume 263. John Benjamins Publishing, 2005.

[BRDS12] Stefan Bott, Luz Rello, Biljana Drndarevic, and Horacio Saggion. Can Spanish be simpler? LexSiS: Lexical simplification for Spanish. In *Proceesing of Coling*, pages 357–374, 2012.

[BSC+16] Steven Bethard, Guergana Savova, Wei-Te Chen, Leon Derczynski, James Pustejovsky, and Marc Verhagen. Semeval-2016 task 12: Clinical tempeval. *Proceedings of SemEval*, pages 1052–1062, 2016.

[BT13] Gianni Barlacchi and Sara Tonelli. Ernesta: A sentence simplification tool for childrenś stories in italian. In *Proceeding of the 14th International Conference of Computational Linguistics and Intelligent Text Processing (CICLing), LNCS-volume 7817*, pages 476–487. 2013.

[CD95] Jeanne Sternlicht Chall and Edgar Dale. *Readability revisited: The new Dale-Chall readability formula*. Brookline Books, 1995.

[CE07] Jamie Callan and Maxine Eskenazi. Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technology (NAACL-HLT)*, pages 460–467, 2007.

[Cha58] Jeanne Sternlicht Chall. *Readability: An appraisal of research and application.* Number 34. Ohio State University Columbus, 1958.

[CK11a] William Coster and David Kauchak. Learning to simplify sentences using wikipedia. In *Proceedings of the Workshop on Monolingual Text-to-Text Generation*, pages 1–9, Portland, Oregon, June 2011.

[CK11b] William Coster and David Kauchak. Simple English Wikipedia: A new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT): Short papers-Volume 2*, pages 665–669, Portland, Oregon, June 2011.

[CMC+98a] John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. Practical simplification of english newspaper text to assist aphasic readers. In *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, pages 7–10, 1998.

[CMC+98b] John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. Practical simplification of English newspaper text

to assist aphasic readers. In *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, page 7–10, Wisconsin, July 1998.

[CMKP13] Jin-Woo Chung, Hye-Jin Min, Joonyeob Kim, and Jong C Park. Enhancing readability of web documents by text augmentation for deaf people. In *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics (WIMS)*, pages 1–10, Madrid, Spain, June 2013.

[CMO03] Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Current and new directions in discourse and dialogue*, pages 85–112. 2003.

[Coh60] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[COM02] Lynn Carlson, Mary Ellen Okurowski, and Daniel Marcu. *RST discourse treebank*. Linguistic Data Consortium, Catalog Number-LDC2002T07, 2002.

[CS97] Raman Chandrasekar and Bangalore Srinivas. Automatic induction of rules for text simplification. *Knowledge-Based Systems*, 10(3):183–190, 1997.

[CT12] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[CT14] Kevyn Collins-Thompson. Computational assessment of text

readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics*, 165(2):97–135, 2014.

[CTC04] Kevyn Collins-Thompson and Jamie Callan. Information retrieval for language tutoring: An overview of the REAP project. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 544–545, Sheffield, UK, July 2004.

[DBKMR05] Pieter-Tjerk De Boer, Dirk P Kroese, Shie Mannor, and Reuven Y Rubinstein. A tutorial on the cross-entropy method. *Annals of operations research*, 134(1):19–67, 2005.

[DC48] Edgar Dale and Jeanne S Chall. A formula for predicting readability: Instructions. *Educational research bulletin*, 27(2):37–54, 1948.

[DC49] Edgar Dale and Jeanne S Chall. The concept of readability. *Elementary English*, 26(1):19–26, 1949.

[DEMP97] Barbara Di Eugenio, Johanna D Moore, and Massimo Paolucci. Learning features that predict cue usage. In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 80–87, Madrid, Spain, 1997.

[DEO14] Iustin Dornescu, Richard Evans, and Constantin Orasan. Relative clause extraction for syntactic simplification. In *Proceedings of the Workshop on Automatic Text Simplification: Methods*

and Applications in the Multilingual Society, International Conference on Computational Linguistics (COLING), pages 1–10, Dublin, Ireland, August 2014.

[DH01] Mary C Dyson and Mark Haselgrove. The influence of reading speed and line length on the effectiveness of reading from screen. *International Journal of Human-Computer Studies*, 54(4):585–612, 2001.

[DK15] Elnaz Davoodi and Leila Kosseim. On the influence of text complexity on discourse-level choices. *International Journal of Computational Linguistics and Applications (IJCLA)*, 6(1):25–44, 2015.

[DK16a] Elnaz Davoodi and Leila Kosseim. CLaC @ QATS: Quality Assessment for Text Simplification. In *Proceeding of the International Workshop on Quality Assessment for Text Simplification at LREC*, pages 53–57. Portoroẑ, Slovenia, 2016.

[DK16b] Elnaz Davoodi and Leila Kosseim. Exploring Linguistic and Psycholinguistic Features for Complex Word Identification. In *Proceeding of International Workshop on Semantic Evaluation at NAACL (SemEval 2016)*, pages 907–911. San Diego, USA, 2016.

[DK16c] Elnaz Davoodi and Leila Kosseim. On the contribution of discourse structure on text complexity assessment. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SigDial)*, pages 166–174, Los Angeles, September 2016.

[DK17a] Elnaz Davoodi and Leila Kosseim. A Context-Aware Approach

for the Identification of Complex Words in Natural Language Texts. In *Proceedings of the IEEE 11th International Conference on Semantic Computing (IEEE-ICSC)*, pages 97–100, San Diego, USA, January 2017.

[DK17b] Elnaz Davoodi and Leila Kosseim. Automatic Identification of AltLexes using Monolingual Parallel Corpora. In *Proceedings of Recent Advances in Natural Language Processing (RANLP 2017)*, page 6 pages. Varna, Bulgaria, 2017.

[DKB+16] Elnaz Davoodi, Leila Kosseim, Félix-Hervé Bachand, Majid Laali, and Emmanuel Argollo. Classification of textual genres using discourse information. In *Proceedings of the 17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing), Lecture Notes in Computer Science*, page 12 pages, Konya, Turkey, April 2016.

[DT34] Edgar Dale and Ralph W Tyler. A study of the factors influencing the difficulty of reading materials for adults of limited reading ability. *The Library Quarterly*, 4(3):384–412, 1934.

[DT98a] Siobhan Devlin and John Tait. The use of a psycholinguistic database in the simplification of text for aphasic readers. *Linguistic databases*, pages 161–173, 1998.

[DT98b] Siobhan Devlin and John Tait. The use of a psycholinguistic database in the simplification of text for aphasic readers. *Linguistic Databases*, page 161–173, 1998.

[DU06] Siobhan Devlin and Gary Unthank. Helping aphasic people process online information. In *Proceedings of the 8th International*

ACM SIGACCESS Conference on Computers and Accessibility, pages 225–226, 2006.

[DuB04] William H DuBay. The principles of readability. *Online Submission*, 2004.

[ESR08] James M Eales, Robert Stevens, and David Robertson. Full-text mining: Linking practice, protocols and articles in biological research. *Proceedings of the BioLink SIG, ISMB*, 2008.

[Eva11] Richard J Evans. Comparing methods for the syntactic simplification of sentences in information extraction. *Literary and Linguistic Computing*, page fqr034, 2011.

[FH12] Vanessa Wei Feng and Graeme Hirst. Text-level Discourse Parsing with Rich Linguistic Features. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL): Long Papers - Volume 1*, pages 60–68, Jeju Island, Korea, 2012.

[FIK10] George Foster, Pierre Isabelle, and Roland Kuhn. Translating structured documents. In *Proceedings of Association for Machine Translation in the Americas (AMTA)*, Denver, Colorado, November 2010.

[Fis00] Kerstin Fischer. *From cognitive semantics to lexical pragmatics: The functional polysemy of discourse particles*. Walter de Gruyter, 2000.

[FM12] Thomas François and Eleni Miltsakaki. Do NLP and machine

learning improve traditional readability formulas? In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*, pages 49–57, 2012.

[FM13] Syeed Ibn Faiz and Robert E Mercer. Identifying explicit discourse connectives in text. In *Proceedings of the Canadian Conference on Artificial Intelligence*, pages 64–76, Sasketchwan, Canada, March 2013.

[G+98] XTAG Research Group et al. A lexicalized tree adjoining grammar for English. *arXiv preprint cs/9809024*, 1998.

[GH98] Laurie Gerber and Eduard Hovy. Improving translation quality by manipulating sentence length. In *Conference of the Association for Machine Translation in the Americas*, pages 448–460, London, UK, 1998.

[Gri75] Joseph Evans Grimes. *The thread of discourse*, volume 207. Walter de Gruyter, 1975.

[GS86] Barbara J Grosz and Candace L Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204, 1986.

[Gun69] Robert Gunning. The Fog index after twenty years. *Journal of Business Communication*, 6(2):3–13, 1969.

[Gun03] Thomas G Gunning. *Building literacy in the content areas*. Allyn & Bacon, 2003.

[Gut81] John T Guthrie. *Comprehension and Teaching: Research Reviews*. ERIC, 1981.

[GVCB13] Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. PPDB: The paraphrase database. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technology (NAACL-HLT)*, pages 758–764, Atlanta, Georgia, June 2013.

[GWJ95] Barbara J Grosz, Scott Weinstein, and Aravind K Joshi. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225, 1995.

[HCTCE06] Michael Heilman, Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskenazi. Classroom success of an intelligent tutoring system for lexical practice and reading comprehension. In *Proceedings of the International Conference on Spoken Language Processing (INTERSPEECH)*, Pennsylvania, USA, September 2006.

[HH76] Michael AK Halliday and Ruqaiya Hasan. *Cohesion in English*. Routledge, 1976.

[HM14] Hospice Houngbo and Robert Mercer. An automated method to build a corpus of rhetorically-classified sentences in biomedical texts. In *Proceedings of the first workshop on argumentation mining*, pages 19–23, Maryland, USA, June 2014.

[HM16] Christopher Hidey and Kathleen McKeown. Identifying causal relations using parallel wikipedia articles. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1424–1433, Berlin, Germany, 2016.

[HPAdI10] Hugo Hernault, Helmut Prendinger, David A. duVerle, and Mitsuru Ishizuka. Hilda: A discourse parser using support vector machine classification. *Dialogue & Discourse*, 1(3):1–33, 2010.

[Jay57] Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

[JCK+96] Marcel Adam Just, Patricia A Carpenter, Timothy A Keller, William F Eddy, and Keith R Thulborn. Brain activation modulated by sentence comprehension. *Science*, 274(5284):114, 1996.

[Jur00] Daniel Jurafsky. Speech and language processing: An introduction to natural language processing. *Computational linguistics, and speech recognition*, 2000.

[K+63] George Roger Klare et al. *Measurement of readability*. Iowa State University Press, 1963.

[Kau13] David Kauchak. Improving text simplification language modeling using unsimplified text data. In *Proceeding of the 51st Annual Meeting of the Association for Computational Linguistics (ACL) Volume 1: Long Papers*, page 1537–1546, Sofia, Bulgaria, August 2013.

[KCTBD12] Jin Young Kim, Kevyn Collins-Thompson, Paul N Bennett, and Susan T Dumais. Characterizing web content, user interests, and search behavior by reading level and topic. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 213–222, 2012.

[KD94] Alistair Knott and Robert Dale. Using linguistic phenomena

to motivate a set of coherence relations. *Discourse Processes*, 18(1):35–62, 1994.

[KFJRC75] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. Derivation of new readability formulas (automated readability index, Fog count and Flesch reading ease formula) for Navy enlisted personnel. Technical report, DTIC Document, 1975.

[KLP+10] Rohit J Kate, Xiaoqiang Luo, Siddharth Patwardhan, Martin Franz, Radu Florian, Raymond J Mooney, Salim Roukos, and Chris Welty. Learning to predict readability using diverse linguistic features. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pages 546–554, Beijing, China, August 2010.

[Kno96] Alistair Knott. A data-driven methodology for motivating a set of coherence relations. *PhD. thesis. University of Edinburgh: College of Science and Engineering: The School of Informatics*, 1996.

[KSL08] Judith Kamalski, Ted Sanders, and Leo Lentz. Coherence marking, prior knowledge, and comprehension of informative and persuasive texts: Sorting things out. *Discourse Processes*, 45(4-5):323–345, 2008.

[LCK16] Majid Laali, Andre Cianflone, and Leila Kosseim. The CLaC Discourse Parser at CoNLL-2016. In *Proceedings of the 20th Conference on Computational Natural Language Learning: Shared Task, (CoNLL)*, pages 92–99, Berlin, Germany, August 2016.

[LDK15] Majid Laali, Elnaz Davoodi, and Leila Kosseim. The CLaC Discourse Parser at CoNLL-2015. In *Proceedings of the 19th Conference on Computational Natural Language Learning: Shared Task, (CoNLL)*, pages 56–60, Beijing, China, 2015.

[Lev83] Stephen C Levinson. *Pragmatics (Cambridge textbooks in linguistics)*. Cambridge University Press, 1983.

[LEVDB+00] Tracy Linderholm, Michelle Gaddy Everson, Paul Van Den Broek, Maureen Mischinski, Alex Crittenden, and Jay Samuels. Effects of causal text revisions on more-and less-skilled readers' comprehension of easy and difficult texts. *Cognition and Instruction*, 18(4):525–556, 2000.

[LHW+12] Joshua Levy, Elizabeth Hoover, Gloria Waters, Swathi Kiran, David Caplan, Alex Berardino, and Chaleece Sandberg. Effects of syntactic complexity, semantic reversibility, and explicitness on discourse comprehension in persons with aphasia and in healthy controls. *American Journal of Speech-Language Pathology*, 21(2):S154–S165, 2012.

[LK14] Majid Laali and Leila Kosseim. Inducing discourse connectives from parallel texts. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*, pages 610–618, Dublin, Ireland, 2014.

[LKN09] Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. Recognizing implicit discourse relations in the Penn discourse treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP): Volume 1*, pages 343–351, Singapore, August 2009.

[LNK14] Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. A PDTB-styled end-to-end discourse parser. *Natural Language Engineering*, 20(2):151–184, 2014.

[Mar00] Daniel Marcu. *The theory and practice of discourse parsing and summarization*. MIT press, 2000.

[MBB80] Bonnie JF Meyer, David M Brandt, and George J Bluth. Use of top-level structure in text: Key for reading comprehension of ninth-grade students. *Reading research quarterly*, pages 72–103, 1980.

[MC07] Mstislav Maslennikov and Tat-Seng Chua. A multi-resolution framework for information extraction from free text. In *Proceeding of the Annual Meeting-Association for Computational Linguistics (ACL)*, pages 592–599, Prague, Czech Republic, June 2007.

[McK85] Kathleen R McKeown. Discourse strategies for generating natural-language text. *Artificial Intelligence*, 27(1):1–41, 1985.

[Mil95] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[MJ00] James H Martin and Daniel Jurafsky. Speech and language processing. *International Edition*, 710, 2000.

[ML69] G Harry Mc Laughlin. Smog grading-a new readability formula. *Journal of reading*, 12(8):639–646, 1969.

[MLB96] J Stephen Mansfield, Gordon E Legge, and Mark C Bane. Psychophysics of reading. xv: Font effects in normal and low vision.

Investigative Ophthalmology & Visual Science, 37(8):1492–1501, 1996.

[MM95] Megan Moser and Johanna D Moore. Investigating cue selection and placement in tutorial discourse. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics (ACL)*, pages 130–135, Cambridge, Massachusetts, USA, 1995.

[MMS93] Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.

[MN11] Ryan McDonald and Joakim Nivre. Analyzing and integrating dependency parsers. *Computational Linguistics*, 37(1):197–230, 2011.

[Moo02] Robert C Moore. Fast and accurate sentence alignment of bilingual corpora. In *Conference of the Association for Machine Translation in the Americas*, pages 135–144, Tampa Florida, USA, August 2002.

[MSA+11] Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, et al. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182, 2011.

[MT87] William C Mann and Sandra A Thompson. Rhetorical structure theory: A framework for the analysis of texts. Technical report, IPRA Papers in Pragmatics 1, 1987.

[MW13] Thomas Meyer and Bonnie Webber. Implicitation of discourse connectives in (machine) translation. In *Proceedings of the 1st DiscoMT Workshop at ACL 2013 (51st Annual Meeting of the Association for Computational Linguistics)*, pages 19–26, Sofia, Bulgaria, 2013.

[NCBP16] Courtney Napoles, Chris Callison-Burch, and Matt Post. Sentential paraphrasing as black-box machine translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 62–66, San Diego, USA, 2016.

[NS06] Rani Nelken and Stuart M Shieber. Towards robust context-sensitive sentence alignment for monolingual corpora. In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 161–168, Trento, Italy, April 2006.

[ON04] Franz Josef Och and Hermann Ney. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449, 2004.

[Pas03] Renate Pasch. *Handbuch der deutschen Konnektoren: linguistische Grundlagen der Beschreibung und syntaktische Merkmale der deutschen Satzverknüpfer (Konjunktionen, Satzadverbien und Partikeln)*, volume 1. Walter de Gruyter, 2003.

[Pau09] Peter V Paul. *Language and deafness*. Jones & Bartlett Learning, 2009.

[PDL+08] Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio

Robaldo, Aravind Joshi, and Bonnie L. Webber. The Penn Discourse TreeBank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, Marrakesh, Morocco, June 2008.

[PK13] Gary Patterson and Andrew Kehler. Predicting the presence of discourse connectives. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 914–923, Seattle, Washington, USA, 2013.

[PLN09] Emily Pitler, Annie Louis, and Ani Nenkova. Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2*, pages 683–691, Suntec, Singapore, August 2009.

[PMF+11] Rashmi Prasad, Susan McRoy, Nadya Frid, Aravind Joshi, and Hong Yu. The Biomedical Discourse Relation Bank. *BMC bioinformatics*, 12(1):188, 2011.

[PN08] Emily Pitler and Ani Nenkova. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 186–195, Honolulu, October 2008.

[PR07] Siddharth Patwardhan and Ellen Riloff. Effective information extraction with semantic affinity patterns and relevant regions. In *Proceedings of the 2007 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 717–727, Prague, Czech Republic, June 2007.

[PS84] Livia Polanyi and Remko Scha. A syntactic approach to discourse semantics. In *Proceedings of the 10th International Conference on Computational Linguistics (COLING)*, pages 413–419, California, USA, July 1984.

[QBD04] Chris Quirk, Chris Brockett, and William B Dolan. Monolingual machine translation for paraphrase generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 142–149, Barcelona, Spain, July 2004.

[Red91] Gisela Redeker. Linguistic markers of discourse structure. *Linguistics*, 29(6):1139–1172, 1991.

[RG00] Paul Rayson and Roger Garside. Comparing corpora using frequency profiling. In *Proceedings of the Workshop on Comparing Corpora*, pages 1–6, Hong Kong, October 2000.

[Sch85] Lawrence Clifford Schourup. *Common discourse particles in English conversation*. Dissertations-G, 1985.

[Sch88] Deborah Schiffrin. *Discourse markers*. Number 5. Cambridge University Press, 1988.

[SDOCJ+10] Carolina Scarton, Matheus De Oliveira, Arnaldo Candido Jr, Caroline Gasperin, and Sandra Maria Aluísio. Simplifica: A tool for authoring simplified texts in brazilian portuguese guided by readability assessments. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL): Demonstrations*, pages 41–44, Los Angeles, CA, USA, June 2010.

[Sid03] Advaith Siddharthan. Preserving discourse structure when simplifying text. In *Proceedings of the European Natural Language Generation Workshop (ENLG) at the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 103–110, Budapest, Hungary, April 2003.

[Sid06] Advaith Siddharthan. Syntactic simplification and text cohesion. *Research on Language & Computation*, 4(1):77–109, 2006.

[Sid14] Advaith Siddharthan. A survey of research on text simplification. *ITL-International Journal of Applied Linguistics*, 165(2):259–298, 2014.

[SM03] Radu Soricut and Daniel Marcu. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL)*, volume 1, pages 149–156, Edmonton, Canada, June 2003.

[SO05] Sarah E Schwarm and Mari Ostendorf. Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL)*, pages 523–530, Ann Arbor, June 2005.

[SP88] Remko Scha and Livia Polanyi. An augmented context free grammar for discourse. In *Proceedings of the 12th Conference on Computational Linguistics (ACL)-Volume 2*, pages 573–577, Budapest, Hungry, August 1988.

[ŠPS+16] S Štajner, M Popovic, H Saggion, L Specia, and M Fishel. Shared task on quality assessment for text classification. In *Proceedings of the LREC Workshop on Quality Assessment for Text Simplification (QATS)*, 2016.

[SSN92] Ted JM Sanders, Wilbert PM Spooren, and Leo GM Noordman. Toward a taxonomy of coherence relations. *Discourse Processes*, 15(1):1–35, 1992.

[STH99] Diane Shorrocks-Taylor and Melanie Hargreaves. Making it clear: A review of language issues in testing with special reference to the national curriculum mathematics tests at key stage 2. *Educational Research*, 41(2):123–136, 1999.

[TAV06] Maite Taboada, Caroline Anthony, and Kimberly Voll. Methods for creating semantic orientation dictionaries. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC)*, pages 427–432, Genoa, Italy, May 2006.

[TD13] Maite Taboada and Debopam Das. Annotation upon annotation: Adding signalling information to a corpus of discourse relations. *Dialogue & Discourse*, 4(2):249–281, 2013.

[Tet01] Joel R Tetreault. A corpus-based evaluation of centering and pronoun resolution. *Computational Linguistics*, 27(4):507–520, 2001.

[TG04] Maite Taboada and Jack Grieve. Analyzing appraisal automatically. In *Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text*, Stanford,CA, March 2004.

[TITT10] Kumiko Tanaka-Ishii, Satoshi Tezuka, and Hiroshi Terada. Sorting texts by readability. *Computational Linguistics*, 36(2):203–227, 2010.

[TJT02] Sandra J Thompson, Christopher J Johnstone, and Martha L Thurlow. Universal design applied to large scale assessments. *Synthesis Report. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.*, 2002.

[TKMS03] Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL-HLT)-Volume 1*, pages 173–180, Edmonton, Canada, May 2003.

[Tom13] Masaru Tomita. *Efficient parsing for natural language: a fast algorithm for practical systems*, volume 8. Springer Science & Business Media, 2013.

[Tra05] Catherine E Travis. *Discourse Markers in Colombian Spanish: A Study in Polysemy*. Walter de Gruyter, 2005.

[TVdBPC04] Gian Lorenzo Thione, Martin Van den Berg, Livia Polanyi, and Chris Culy. Hybrid text summarization: Combining external relevance measures with structural analysis. In *Proceedings of the Annual Meeting-Association for Computational Linguistics, Workshop Text Summarization Branches Out*, pages 51–55, Barcelona, Spain, July 2004.

[Ver10] Yannick Versley. Discovery of ambiguous and unambiguous discourse connectives via annotation projection. In *Proceedings of Workshop on Annotation and Exploitation of Parallel Corpora (AEPC)*, pages 83–82, Tartu, Estonia, December 2010.

[VM12] Sowmya Vajjala and Detmar Meurers. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 163–173, Montreal, Canada, June 2012.

[Web04] Bonnie Webber. D-LTAG: Extending lexicalized tag to discourse. *Cognitive Science*, 28(5):751–779, 2004.

[Web09] Bonnie Webber. Genre distinctions for discourse in the Penn TreeBank. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing (ACL-AFNLP): Volume 2*, pages 674–682, Suntec, Singapore, August 2009.

[Wil90] William Williams. *Composition and Rhetoric by practice*. D. C. Health & Co Publishers, 1890.

[Wil45] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83, 1945.

[Win73] Terry Winograd. A procedural model of language understanding. *Procedural Model of Language Understanding*, pages 152–186, 1973.

[WJ98] Bonnie Lynn Webber and Aravind K Joshi. Anchoring a lexicalized tree-adjoining grammar for discourse. In *Proceedings of the*

Workshop on Discourse Relations and Discourse Markers at the International Conference on Computational Linguistics (COLING), Montreal, Canada, August 1998.

[WJ12] Bonnie Webber and Aravind Joshi. Discourse structure and computation: past, present and future. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 42–54, Jeju, Korea, June 2012.

[WJP98] Marilyn A Walker, Aravind Krishna Joshi, and Ellen Friedman Prince. *Centering theory in discourse*. Oxford University Press, 1998.

[WJU+09] Willian Massami Watanabe, Arnaldo Candido Junior, Vinícius Rodriguez Uzêda, Renata Pontin de Mattos Fortes, Thiago Alexandre Salgueiro Pardo, and Sandra Maria Aluísio. Facilita: Reading assistance for low-literacy readers. In *Proceedings of the 27th ACM International Conference on Design of Communication*, pages 29–36, Indiana, USA, October 2009.

[WKSJ99] Bonnie Webber, Alistair Knott, Matthew Stone, and Aravind Joshi. What are little texts made of? A structural and presuppositional account using lexicalized tag. In *Proceedings of the International Workshop on Levels of Representation in Discourse (LORID'99)*, pages 145–149, 1999.

[WL11] Kristian Woodsend and Mirella Lapata. Learning to simplify sentences with quasi-synchronous grammar and integer programming. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 409–420, Edinburgh, UK, July 2011.

[Woo68] William A Woods. Procedural semantics for a question-answering machine. In *Proceedings of the AFIPS National Computer Conference*, pages 457–471, San Francisco, California, December 1968.

[Woo78] William A Woods. Semantics and quantification in natural language question answering. *Advances in computers*, 17:1–87, 1978.

[WRO03] Sandra Williams, Ehud Reiter, and Liesl Osman. Experiments with discourse-level choices and readability. pages 127–134, Budapest, Hungary, April 2003.

[WSJK03] Bonnie Webber, Matthew Stone, Aravind Joshi, and Alistair Knott. Anaphora and discourse structure. *Computational Linguistics*, 29(4):545–587, 2003.

[WVDBK12] Sander Wubben, Antal Van Den Bosch, and Emiel Krahmer. Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL): Long Papers-Volume 1*, pages 1015–1024, Jeju, Korea, June 2012.

[XCBN15] Wei Xu, Chris Callison-Burch, and Courtney Napoles. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297, 2015.

[XNP+15] Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Rashmi Prasad, Christopher Bryant, and Attapol Rutherford. The conll-2015 shared task on shallow discourse parsing. In *CoNLL Shared Task*, pages 1–16, 2015.

[XNP+16] Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415, 2016.

[YPDNML10] Mark Yatskar, Bo Pang, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. For the sake of simplicity: Unsupervised extraction of lexical simplifications from Wikipedia. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HTL)*, pages 365–368, Los Angeles, California, USA, June 2010.

[ZBG10] Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (ACL)*, pages 1353–1361, Uppsala, Sweden, July 2010.

[Zip49] George Kingsley Zipf. *Human behavior and the principle of least effort: An introduction to human ecology.* Ravenio Books, 1949.

[ZS88] Beverly L Zakaluk and S Jay Samuels. *Readability: Its Past, Present, and Future.* International Reading Association, 1988.

[Zwi85] Arnold M Zwicky. Clitics and particles. *Language*, 61(2):283–305, 1985.

Appendix A

Table 16: Discourse markers and corresponding discourse relations in the Simple English Wikipedia data set.

Discourse Marker	Discourse Relation	Complex Version Relative Frequency	Simple Version Relative Frequency	Log-likelihood Ratio
because	CAUSE	0.0224	0.0373	0.0016
thus	CAUSE	0.0136	0.0074	0.0008
although	CONTRAST	0.0169	0.0104	0.0007
so	CAUSE	0.0115	0.0181	0.0006
while	CONTRAST	0.0355	0.0264	0.0006
when	SYNCHRONY	0.0620	0.0754	0.0006
as	SYNCHRONY	0.0471	0.0379	0.0005
also	CONJUNCTION	0.1697	0.1894	0.0005
but	CONTRAST	0.0619	0.0712	0.0003
although	CONCESSION	0.0178	0.0129	0.0003
though	CONTRAST	0.0118	0.0082	0.0003
whereas	CONTRAST	0.0032	0.0015	0.0003

Discourse Marker	Discourse Relation	Complex Version Relative Frequency	Simple Version Relative Frequency	Log-likelihood Ratio
and	CONJUNCTION	0.1794	0.1667	0.0002
thereafter	ASYNCHRONOUS	0.0021	0.0009	0.0002
however	CONTRAST	0.0458	0.0406	0.0002
ultimately	ASYNCHRONOUS	0.0010	0.0004	0.0002
when	ASYNCHRONOUS	0.0007	0.0017	0.0002
consequently	CAUSE	0.0015	0.0008	0.0002
thereby	CAUSE	0.0012	0.0006	0.0002
then	ASYNCHRONOUS	0.0325	0.0380	0.0001
so that	CAUSE	0.0050	0.0067	0.0001
if	CONDITION	0.0351	0.0392	0.0001
therefore	CAUSE	0.0068	0.0053	0.0001
previously	ASYNCHRONOUS	0.0020	0.0011	0.0001
by	CONTRAST	0.0004	0.0001	0.0001
and	LIST	0.0001	0.0000	0.0001
as a result	CAUSE	0.0032	0.0021	0.0001
in addition	CONJUNCTION	0.0032	0.0021	0.0001
additionally	CONJUNCTION	0.0021	0.0012	0.0001
in particular	INSTANTIATION	0.0004	0.0002	0.0001
until	ASYNCHRONOUS	0.0105	0.0110	0.0001
by comparison	CONTRAST	0.0001	0.0000	0.0001
finally	ASYNCHRONOUS	0.0003	0.0002	0.0001
otherwise	ALTERNATIVE	0.0010	0.0009	0.0001
moreover	CONJUNCTION	0.0009	0.0004	0.0001

Discourse Marker	Discourse Relation	Complex Version Relative Frequency	Simple Version Relative Frequency	Log-likelihood Ratio
rather	RESTATEMENT	0.0004	0.0002	0.0001
nevertheless	CONCESSION	0.0005	0.0003	0.0001
in	CONTRAST	0.0005	0.0003	0.0001
nonetheless	CONCESSION	0.0006	0.0003	0.0001
while	SYNCHRONY	0.0140	0.0124	0.0001
in fact	CONTRAST	0.0003	0.0001	0.0001
besides	CONJUNCTION	0.0002	0.0002	0.0001
still	CONTRAST	0.0002	0.0002	0.0001
in fact	RESTATEMENT	0.0003	0.0002	0.0001
furthermore	CONJUNCTION	0.0014	0.0008	0.0000
indeed	RESTATEMENT	0.0002	0.0000	0.0000
finally	CONJUNCTION	0.0016	0.0011	0.0000
also	LIST	0.0000	0.0002	0.0000
then	CONDITION	0.0034	0.0034	0.0000
for instance	INSTANTIATION	0.0022	0.0016	0.0000
for example	INSTANTIATION	0.0110	0.0119	0.0000
as well	CONJUNCTION	0.0002	0.0001	0.0000
accordingly	CAUSE	0.0003	0.0001	0.0000
but	CONCESSION	0.0001	0.0001	0.0000
since	CAUSE	0.0085	0.0068	0.0000
and	CONTRAST	0.0001	0.0002	0.0000
once	CONDITION	0.0003	0.0003	0.0000
once	ASYNCHRONOUS	0.0033	0.0029	0.0000
in the end	CAUSE	0.0000	0.0001	0.0000

Discourse Marker	Discourse Relation	Complex Version Relative Frequency	Simple Version Relative Frequency	Log-likelihood Ratio
rather	ALTERNATIVE	0.0003	0.0001	0.0000
further	CONJUNCTION	0.0004	0.0002	0.0000
in the end	ASYNCHRONOUS	0.0000	0.0001	0.0000
meantime	SYNCHRONY	0.0001	0.0001	0.0000
since	ASYNCHRONOUS	0.0038	0.0037	0.0000
as soon as	ASYNCHRONOUS	0.0001	0.0001	0.0000
in turn	ASYNCHRONOUS	0.0016	0.0011	0.0000
in other words	RESTATEMENT	0.0006	0.0004	0.0000
in fact	CONJUNCTION	0.0011	0.0012	0.0000
when	ALTERNATIVE	0.0001	0.0001	0.0000
by then	ASYNCHRONOUS	0.0001	0.0001	0.0000
and	CAUSE	0.0003	0.0003	0.0000
specifically	RESTATEMENT	0.0004	0.0002	0.0000
nor	CONJUNCTION	0.0009	0.0006	0.0000
next	ASYNCHRONOUS	0.0002	0.0003	0.0000
overall	RESTATEMENT	0.0005	0.0003	0.0000
or	ALTERNATIVE	0.0058	0.0059	0.0000
till	ASYNCHRONOUS	0.0001	0.0002	0.0000
when	CONCESSION	0.0006	0.0006	0.0000
separately	CONJUNCTION	0.0001	0.0001	0.0000
before and after	ASYNCHRONOUS	0.0000	0.0001	0.0000
simultaneously	SYNCHRONY	0.0004	0.0002	0.0000
similarly	CONJUNCTION	0.0011	0.0009	0.0000

Discourse Marker	Discourse Relation	Complex Version Relative Frequency	Simple Version Relative Frequency	Log-likelihood Ratio
then	CONDITION	0.0001	0.0000	0.0000
earlier	ASYNCHRONOUS	0.0001	0.0001	0.0000
as soon as	SYNCHRONY	0.0002	0.0001	0.0000
on the other hand	CONTRAST	0.0013	0.0011	0.0000
much as	ALTERNATIVE	0.0000	0.0001	0.0000
when	CAUSE	0.0000	0.0001	0.0000
alternatively	ALTERNATIVE	0.0012	0.0007	0.0000
as long as	SYNCHRONY	0.0006	0.0007	0.0000
now that	CAUSE	0.0001	0.0000	0.0000
before	ASYNCHRONOUS	0.0227	0.0236	0.0000
while	CONCESSION	0.0001	0.0001	0.0000
though	CONCESSION	0.0078	0.0077	0.0000
in short	RESTATEMENT	0.0001	0.0001	0.0000
afterward	ASYNCHRONOUS	0.0006	0.0004	0.0000
later	ASYNCHRONOUS	0.0275	0.0282	0.0000
regardless	CONCESSION	0.0001	0.0001	0.0000
in turn	CONJUNCTION	0.0001	0.0001	0.0000
after	ASYNCHRONOUS	0.0385	0.0391	0.0000
yet	CONTRAST	0.0011	0.0008	0.0000
on the contrary	CONTRAST	0.0001	0.0001	0.0000
as though	RESTATEMENT	0.0001	0.0001	0.0000
conversely	CONTRAST	0.0003	0.0002	0.0000
meanwhile	SYNCHRONY	0.0017	0.0015	0.0000

Discourse Marker	Discourse Relation	Complex Version Relative Frequency	Simple Version Relative Frequency	Log-likelihood Ratio
hence	CAUSE	0.0006	0.0004	0.0000
still	CONCESSION	0.0007	0.0007	0.0000
instead	ALTERNATIVE	0.0071	0.0081	0.0000
nevertheless	CONTRAST	0.0013	0.0009	0.0000
in particular	RESTATEMENT	0.0005	0.0004	0.0000
as if	CONCESSION	0.0006	0.0008	0.0000
until	ALTERNATIVE	0.0001	0.0001	0.0000
when	CONDITION	0.0017	0.0017	0.0000
unless	ALTERNATIVE	0.0018	0.0018	0.0000
indeed	CONJUNCTION	0.0013	0.0011	-0.0001
likewise	CONJUNCTION	0.0008	0.0006	-0.0001
as	CAUSE	0.0004	0.0004	-0.0001
except	EXCEPTION	0.0005	0.0006	-0.0001
Total		1.0000	1.0000	

Appendix B

Table 17: Frequency of automatically discovered new AltLexes for each PDTB relation in the Newsela and SEW corpora.

AltLex \ Relation	ALTERNATIVE	ASYNCHRONOUS	CAUSE	CONCESSION	CONDITION	CONJUNCTION	CONTRAST	INSTANTIATION	RESTATEMENT	SYNCHRONY	Total
including						89					89
subsequently		39	1								40
along with						29					29
even				1		19	7				27
well						26					26
included						24					24
eventually		19									19
used		1				18					19
only							18				18
purposes	2		13								15

AltLex \ Relation	ALTERNATIVE	ASYNCHRONOUS	CAUSE	CONCESSION	CONDITION	CONJUNCTION	CONTRAST	INSTANTIATION	RESTATEMENT	SYNCHRONY	Total
added						13					13
the consequence is			13								13
provided						11				1	12
involved						9					9
whilst							4			5	9
and then		8									8
despite				7							7
due			7								7
shared						7					7
afterwards		6									6
indicating		5	1								6
is another						6					6
just							6				6
more than (that)						6					6
much better						6					6
compared							5				5
prior to		5									5
also given						4					4
concerned						4					4
conducted						4					4
enabling		2	2								4

AltLex \ Relation	ALTERNATIVE	ASYNCHRONOUS	CAUSE	CONCESSION	CONDITION	CONJUNCTION	CONTRAST	INSTANTIATION	RESTATEMENT	SYNCHRONY	Total
extended						4					4
in matters						2			2		4
become		1	2								3
consequentially	1		2								3
made			3								3
means			3								3
notwithstanding							3				3
resulting			3								3
similarly						3					3
thereupon		3									3
upon		3									3
ahead		2									2
anyway							2				2
at the same time										2	2
e.g.								2			2
expanded						2					2
immediately following		2									2
known as	2										2
originally		2									2
actually		1									1
already		1									1
and so			1								1

AltLex \ Relation	ALTERNATIVE	ASYNCHRONOUS	CAUSE	CONCESSION	CONDITION	CONJUNCTION	CONTRAST	INSTANTIATION	RESTATEMENT	SYNCHRONY	Total
arises			1								1
as a matter of fact						1					1
combined						1					1
considered			1								1
equally						1					1
exceeded						1					1
exceptions	1										1
following		1									1
formerly		1									1
further		1									1
includes			1								1
independently						1					1
led			1								1
making			1								1
meant			1								1
mostly								1			1
prior		1									1
recommended						1					1
regarding	1										1
remains				1							1
this being the case			1								1

AltLex \ Relation	ALTERNATIVE	ASYNCHRONOUS	CAUSE	CONCESSION	CONDITION	CONJUNCTION	CONTRAST	INSTANTIATION	RESTATEMENT	SYNCHRONY	Total
typically								1			1
unfortunately							1				1
whenever					1						1
withal						1					1
without							1				1
Total	**7**	**104**	**58**	**9**	**1**	**293**	**47**	**4**	**2**	**8**	**533**