Inducing Discourse Resources Using Annotation Projection

Majid Laali

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

November 2017

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Majid Laali**

Entitled: **Inducing Discourse Resources Using Annotation Projection**

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 Committee:

______________________ Chair
Dr. Luis Amador

______________________ External Examiner
Dr. Eduard Hovy

______________________ Examiner
Dr. Rachida Dssouli

______________________ Examiner
Dr. Gregory Butler

______________________ Examiner
Dr. Sabine Bergler

______________________ Supervisor
Dr. Leila Kosseim

Approved by ______________________
Dr. Sudhir Mudur, Chair
Department of Computer Science and Software Engineering

February 8, 2018

____
Dr. Amir Asif, Dean
Faculty of Engineering and Computer Science

Abstract

Inducing Discourse Resources Using Annotation Projection

Majid Laali, Ph.D.

Concordia University, 2017

An important aspect of natural language understanding and generation involves the recognition and processing of discourse relations. Building applications such as text summarization, question answering and natural language generation needs human language technology beyond the level of the sentence. To address this need, large scale discourse annotated corpora such as the Penn Discourse Treebank (PDTB; Prasad et al., 2008a) have been developed.

Manually constructing discourse resources (e.g. discourse annotated corpora) is expensive, both in terms of time and expertise. As a consequence, such resources are only available for a few languages. In this thesis, we propose an approach that automatically creates two types of discourse resources from parallel texts: 1) PDTB-style discourse annotated corpora and 2) lexicons of discourse connectives. Our approach is based on annotation projection where linguistic annotations are projected from a source language to a target language in parallel texts.

Our work has made several theoretical contributions as well as practical contributions to the field of discourse analysis. From a theoretical perspective, we have proposed a method to refine the naive method of discourse annotation projection by filtering annotations that are not supported by parallel texts. Our approach is based on the intersection between statistical word-alignment models and can automatically identify 65% of unsupported projected annotations. We have also proposed a novel approach for annotation projection that is independent of statistical word-alignment models. This approach is more robust to longer discourse connectives than approaches based on statistical word-alignment models.

From a practical perspective, we have automatically created the Europarl ConcoDisco corpora

from English-French parallel texts of the Europarl corpus (Koehn, 2009). In the Europarl ConcoDisco corpora, around 1 million occurrences of French discourse connectives are automatically aligned to their translation. From the French side of Europarl ConcoDisco, we have extracted our first significant resource, the FrConcoDisco corpora. To our knowledge, the FrConcoDisco corpora are the first PDTB-style discourse annotated corpora for French where French discourse connectives are annotated with the discourse relations that they signalled. The FrConcoDisco corpora are significant in size as they contain more than 25 times more annotations than the PDTB. To evaluate the FrConcoDisco corpora, we showed how they can be used to train a classifier for the disambiguation of French discourse connectives with a high performance. The second significant resource that we automatically extracted from parallel texts is *ConcoLeDisCo*. *ConcoLeDisCo* is a lexicon of French discourse connectives mapped to PDTB discourse relations. While *ConcoLeDisCo* is useful by itself, as we showed in this thesis, it can be used to improve the coverage of manually constructed lexicons of discourse connectives such as LEXCONN (Roze et al., 2012).

Acknowledgments

First and foremost, I would like to thank my mother and my father, Zahra, and Javad. I have to say, this is not just my Ph.D. that I owed to them. I cannot imagine my life without them. I cannot find the appropriate words that could properly describe my appreciation for their sacrifice, devotion, and support.

I am lucky that I have had Nasrin, my beloved wife by my side during my Ph.D. She has been my strength, my courage and my power in difficult situations.

This thesis could not be finished without the guidance and support of my supervisor, Dr. Leila Kosseim. Leila was not just my supervisor, but also my first Canadian friend. She was also my teacher. She taught me various skills, such as how to listen and constructively criticize and how to be patient when expressing an idea.

Finally, I would like to thank the chair and the committee members of my defense: Dr. Sabine Bergler, Dr. Eduard Hovy, Dr. Rachida Dssouli, Dr. Gregory Butler and Dr. Luis Amador. I really appreciate their valuable feedback and also the time and effort they put into my thesis.

Contents

List of Figures

List of Tables

List of Algorithms

Glossary

CLaC DC Disambiguator An automatic tool build in this thesis which disambiguates the discourse-usage and the discourse relations of English discourse connectives. Note that the French version of the CLaC DC Disambiguator only disambiguates the discourse-usage of French discourse connectives. 7, 8, 10–12, 36, 39, 58, 62–65, 67, 71, 73, 75, 90, 103–105, 107–109

ConcoLeDisCo A resource that we created in this thesis. ConcoLeDisCo is a lexicon of French discourse connectives mapped to PDTB discourse relations. iv, xiii, 8–10, 12, 79, 81, 83, 84, 86, 104, 105, 108, 109

ANNODIS This corpus includes the annotation of 3,355 French discourse relations using the SDRT framework. See (Afantenos et al., 2012). 4, 18, 26, 27, 34

CoNLL In 2015 and 2016, the Conference on Computational Natural Language Learning (CoNLL) organized a shared-task on shallow discourse parsing. See (Xue et al., 2015) and (Xue et al., 2016). xii, 8, 36, 39, 45, 46, 50–52

D-LTAG Discourse Lexicalized Tree Adjoining Grammar (D-LTAG) is a model based on tree-adjoining grammar to describe discourse structures. D-LTAG is the framework behind the Penn Discourse Treebank. See (Webber et al., 2003). 18, 19

EDU In RST, Elementary Discourse Units (EDUs) relate two different abstract objects in a discourse such as events, states or propositions. Their closest equivalent in the PDTB are called *discourse arguments*. See (Mann and Thompson, 1987). 24

Europarl ConcoDisco A set of resources that we have created in this thesis. The Europarl ConcoDisco corpora are based on the Europarl corpus and has English and French discourse connectives aligned to each other. Europarl ConcoDisco includes four specific corpora that differ in the word-alignment model used: ConcoDisco-Naive-Grow-diag corpus, ConcoDisco-Grow-diag corpus, ConcoDisco-Intersection corpus, ConcoDisco-Direct, FrConcoDisco-Inverse. The Europarl ConcoDisco corpora were used to build the French equivalent FrConcoDisco corpora. iii, iv, x, xiii, 8–10, 12, 34, 35, 58, 62, 67, 68, 79–82, 94, 104–106, 108, 109

Europarl parallel corpus One of the largest available parallel corpora. It contains sentence-aligned texts extracted from the proceeding of the European parliament. The English-French part of the corpus contains around 50 millions words and 2 million sentences. See (Koehn, 2005). 9, 62, 67, 90, 91, 95

FDTB The French Discourse Treebank (FDTB) contains more than 10,000 instances of LEXCONN's French discourse connectives annotated with discourse-usage. However, the discourse connectives are not annotated with discourse relations. See (Danlos et al., 2015). x, xii, xiii, 10, 27, 28, 34, 41, 45–48, 50, 54, 62, 68, 74, 75

FrConcoDisco A resource that we created in this thesis. The FrConcoDisco corpora constitute the first French PDTB-style corpora annotated with discourse connectives and the discourse relations that they convey. FrConcoDisco includes four specific corpora that differ in the word-alignment model used: FrConcoDisco-Naive-Grow-diag corpus, FrConcoDisco-Grow-diag corpus, FrConcoDisco-Intersection corpus, FrConcoDisco-Direct, FrConcoDisco-Inverse. iv, xiii, 8–11, 35, 58, 59, 62, 68, 69, 73–77, 104, 105, 109

LEXCONN A manually built lexicon of French discourse connectives associated with their discourse relations. LEXCONN contains 371 discourse connectives where 343 are mapped to an average of 1.3 discourse relations taken from various sources including RST, SDRT, and PDTB. See (Roze et al., 2012). xiii, xiv, 26, 27, 34, 62, 64, 79–86, 88, 90, 92, 95–98, 101, 102

NDU Usage of a connective that does not signal a discourse relation. xiii, 64, 66, 67, 69, 71, 73–78

NLP Natural Language Processing. 7, 29

PDTB The Penn Discourse TreeBank (PDTB) is the largest annotated corpus of discourse information. The corpus is based on the Penn TreeBank, includes all articles in the *Wall Street Journal corpus* (2159 articles) and contains 1 million words. The PDTB includes low-level discourse structures and relations between two text spans are tagged. The PDTB does not represent a full discourse structure of texts. Instead, textual discourse structure are flat and may not be fully connected. See (Pardo et al., 2008) . 3, 10, 12

POS Part of Speech. xiv, 39, 90, 95

RST-DT Rhetorical Structure Theory Discourse Treebank (RST-DT) is one of the first and largest RST-based discourse annotated corpora. This corpus contains the annotations of 385 texts from the *Wall Street Journal*. See (Carlson et al., 2001) . 5, 15, 17

SDRT Segmented Discourse Representation Theory (SDRT) is a recent discourse theory which focuses on extending existing theories of sentence semantics to the discourse level. SDRT uses a graph-based representation. See (Asher and Lascarides, 2003). xii, 17, 18

SMT Statistical Machine Translation. 7, 107

Chapter 1

Introduction

To compose a text, a writer (or speaker) semantically or rhetorically connects text spans (e.g. sentences and clauses) together. For example, in (Ex. 1), the second sentence is an *Expansion* of what is claimed in the first sentence.

(Ex. 1) Failure is an option here. If things are not failing, you are not innovating enough. (Elon Musk, Feburary 2005)

In addition, the second sentence consists of two clauses where the first clause *'If things are not failing'* is a *Condition* of the second clause *'you are not innovating enough'*. Here, *Expansion* and *Condition* are discourse relations that semantically or rhetorically connect the text spans of (Ex. 1).

Theories of discourse coherence study the rules that govern how clauses and sentences are combined with each other to construct a coherent text (Mann and Thompson, 1987; Asher, 1993). While syntax theories focus on the internal structure of sentences, discourse theories investigate the structure of texts beyond sentences. The building blocks of discourse theories are sentences and clauses which are referred to as *discourse arguments* (Mann and Thompson, 1987; Asher and Lascarides, 2003; Prasad et al., 2008a). The semantic content of discourse arguments is referred to as an *abstract object* (Asher, 1993). An abstract object is a proposition, a fact, an event, a situation or a belief. For example, (Ex. 2) is a discourse containing a sequence of sentences and clauses each explaining a fact and/or an event.

(Ex. 2) Men have a tragic genetic flaw. As a result, they cannot see dirt until there is enough of it to support agriculture.[1]

It is important to recognize that within a discourse, the whole conveys more than the sum of its parts (Webber and Joshi, 2012). For example, while each sentence in (Ex. 3) asserts a single event, the second sentence is meant to provide a *Reason* for the first event (i.e. 'not worrying').

(Ex. 3) Don't worry about the world coming to an end today. It is already tomorrow in Australia.[1]

1.1 Annotating Text at the Discourse Level

Identifying discourse relations allows the reader (or hearer) to better understand the communicative goal of the writer (or speaker). Therefore, to interpret the meaning of a discourse, it is essential to recognize its discourse structure: the semantic and/or rhetorical relations between its abstract objects (e.g. a *Reason* relation between the two sentences in (Ex. 3)). These relations are referred to as *discourse relations* or *rhetorical relations*.

To provide a test bed for discourse theories and promote the development of computational approaches, the field of corpus linguistics has developed different projects aiming at the development of discourse annotated corpora (e.g. the RST Discourse Treebank (Carlson et al., 2001), the DISCOR corpus (Reese et al., 2007), the Penn Discourse Treebank (Prasad et al., 2008a)). Discourse annotated corpora consist of texts (from a few hundred to a few thousand articles) annotated with discourse information.

However, annotating discourse structures within a text is difficult, time-consuming and requires expert human annotators. For example, to build the RST Discourse Treebank (Carlson et al., 2001), professional language analysts with prior experience in data annotation were hired. Moreover, these annotators underwent extensive hands-on training during roughly one year. Even with these resources, Carlson et al. (2001) were only able to annotate 385 out of the 2159 newspaper articles of the *Wall Street Journal* corpus (Mitchell et al., 1995).

To avoid the heavy cost of expert manual discourse annotations, Prasad et al. (2008a) chose a different approach and only annotated surface discourse relations when creating the Penn Discourse

[1]The example was taken from (Webber and Joshi, 2012).

Treebank (PDTB). In the PDTB, discourse relations are assumed to be binary relations between two *discourse arguments* and discourse relations are associated to lexical elements, so-called *discourse connectives*. More specifically, discourse relations between two discourse arguments are triggered by either lexical elements (or *explicit discourse connectives*) such as *however* or *because*, or without any lexical element and are inferred by the reader. If a discourse relation is not explicitly signalled, annotators of the PDTB inserted an inferred discourse connective (or *implicit discourse connective*) between the text spans which conveys the same discourse relation.

For example, (Ex. 4) and (Ex. 5) show the PDTB annotations for an explicit discourse relation and an implicit discourse relation respectively.

(Ex. 4) Men have a tragic genetic flaw. As a result they cannot see dirt until there is enough of it to support agriculture. (*CONTINGENCY:Cause:result*)

(Ex. 5) Don't worry about the world coming to an end today. Implicit = BECAUSE It is already tomorrow in Australia. (*CONTINGENCY:Cause:reason*)

In (Ex. 4), a *CONTINGENCY:Cause:result* discourse relation[2] is explicitly signaled by the discourse connective *as a result*. On the other hand, in (Ex. 5), the *CONTINGENCY:Cause:reason* relation is implicit between the first and the second sentences. In this example, the discourse connective *because* has been inferred by the reader and inserted between the two discourse arguments.

As a result of its annotation schema, the PDTB heavily relies on discourse connectives to annotate discourse relations. The PDTB used an inventory of 100 English discourse connectives: all instances of this pre-defined list of connectives have first been marked, then manually annotated by experts. Given this approach, a lexicon of English discourse connectives mapped to their potential discourse relations is very useful to build PDTB-style discourse annotated corpora. For example, Table 1.1 shows a few entries of a lexicon of discourse connectives extracted from the PDTB. As this table shows, a relation can be signed by different connectives, and the same connective can be used to signal different relations.

Although, the PDTB approach to annotated discourse relations does suffer from limitations compared to other approaches (especially in the annotation of implicit discourse relations), its less

[2]The inventory of the PDTB discourse relations is discussed in Chapter 2.

English Discourse Connective	Relation
because	*CONTINGENCY:Cause:result*
but	*COMPARISON:Contrast*
for example	*EXPANSION:Instantiation*
while	*TEMPORAL:Synchronous*
while	*COMPARISON:Contrast*

Table 1.1: A few entries of a lexicon of English discourse connectives.

comprehensive and less costly approach allowed Prasad et al. (2008a) to annotate all 2159 articles of the *Wall Street Journal* corpus (Marcus et al., 1993). As a result, the PDTB is today the largest discourse annotated corpus for English as it contains the annotations of 40,600 discourse relations.

Because of its significant size, the PDTB has been used to develop several discourse related applications, in particular discourse parsers, classifiers that automatically identify discourse relations with a usable accuracy[3] (Faiz and Mercer, 2013; Lin et al., 2014; Xue et al., 2015, 2016; Versley, 2010; Lin et al., 2014; Xue et al., 2016).

The trade-off between the simple discourse annotations and the size of the PDTB makes this framework interesting for developing discourse annotated corpora. As a result, the methodology used in the PDTB has been adopted to create corpora for other languages (e.g. Spanish (Da Cunha et al., 2011), German (Stede, 2004), Czech (Mladová et al., 2008), Turkish (Zeyrek et al., 2010), Arabic (Al-Saif and Markert, 2010), Chinese (Zhou et al., 2012) and French (Afantenos et al., 2012; Danlos et al., 2015)). Nevertheless, the PDTB project still took six years to be developed and required human expert annotators.

1.2 Research Objectives

To date, many languages suffer from a lack of discourse annotated corpora. If such resources do exist, their size is often restrictive. For example, ANNODIS (Afantenos et al., 2012), a corpus for French, contains only 3355 annotations of discourse relations within 86 documents. Given the importance of annotated corpora and the lack of such resources in many languages, **the goal of this thesis is to develop an approach to automatically build**:

[3] See Chapter 3 for more details.

(1) **a PDTB-style discourse annotated corpus for French, and**

(2) **a lexicon of discourse connectives for French mapped to PDTB relations.**

We chose the PDTB framework to annotate discourse relations because: (1) the large size of the PDTB allowed us to build a more reliable discourse parser, (2) the PDTB has been widely adopted in various projects and languages which allows us to evaluate and compare our work.

In our thesis, we used French as the target language because of our access to bilingual English-French speakers. However, we make no assumption about the target language except the availability of a parallel corpus with English; hence the approach should be easy to expand to other similar languages.

To achieve our objectives, we attempted to answer to following research questions:

(Q. 1) Can English discourse connectives be automatically annotated? (see Chapter 3)

(Q. 2) How can annotations of discourse connectives be automatically projected withing parallel texts in order to induce PDTB-style discourse annotated corpora? (see Chapter 4)

(Q. 3) How can lexicons of discourse connectives for the target language be induced from parallel texts? (see Chapter 5 and Chapter 6)

1.3 Scope and Limitations

In this thesis, we focused on the case of explicit discourse relations. According (Prasad et al., 2008b), explicit discourse relations account for 45% of the discourse relations annotated in the PDTB, and according to Stede and Grishina (2016), they account for 37% of the RST relations annotated in the Potsdam Commentary Corpus (Stede and Neumann, 2014)[4]. Moreover, we chose to focus on explicit discourse relations because they form a common denominator of different discourse theories. For example, any phrase that starts with a discourse connective is always considered to be connected to other phrases with a discourse relation in RST-DT too (Carlson et al., 2001).

[4]See Appendix A for a more detailed discussion on mapping explicit PDTB discourse relations to RST relations.

Moreover, automatic identification of explicit discourse relation is more robustness and efficient. This makes them an attractive linguistic phenomena, specifically for studying different aspect of discourse relations (Meyer and Poláková, 2013; Taboada and de los Ángeles Gómez-González, 2012; Zufferey and Degand, 2014; Zufferey and Gygax, 2015; Hoek and Zufferey, 2015) (see Section 2.2 for details).

The underlying assumption of our work is that using available resources, we can annotate French texts based on their English translation. More specifically, we made the following three main assumptions:

Assumption 1: Parallel texts can be built more reliably than discourse resources, hence they are available for more languages. Parallel texts can be extracted from various resources such as bilingual websites, subtitles of movies and translated books. Currently, parallel texts are available for many languages[5] (Tiedemann, 2009, 2012).

Assumption 2: Explicit discourse connectives and the relations that they signal can be automatically identified in the English side of parallel texts with a high accuracy. This assumption is confirmed by research on the development of discourse parsers (e.g. (Versley, 2010; Lin et al., 2014; Xue et al., 2015, 2016)).

Assumption 3: Discourse relations are typically preserved during the translation process, and therefore, French discourse connectives can be labeled using their translation. For example, in the parallel sentences shown in Figure 1.1, the French discourse connective *car* has been translated by the English discourse connective *since*, therefore, we can infer that they both signal the same discourse relation. This assumption has been made in many other previous work (e.g. (Prasad et al., 2010; Versley, 2010; Meyer, 2011; Popescu-Belis et al., 2012; Cartoni et al., 2013; Laali and Kosseim, 2014; Hidey and McKeown, 2016)).

[5]See `http://opus.lingfil.uu.se` for a list of publicly available parallel corpora.

1.4 Motivation

A method to automatically build discourse annotated corpora and lexicons of discourse connectives in different languages has both practical and theoretical motivations:

(1) **Practical Motivations**: Such a method would allow us to quickly build initial discourse resources (i.e. discourse annotated corpora and lexicons of discourse connectives) for resource-poor languages and reduce the gap between resource-rich and resource-poor languages. Not only are the resulting discourse annotated resources useful in themselves, but they can also be used to improve the coverage of manually constructed discourse resources. Moreover, these extended resources can themselves be used to develop or improve discourse-related applications such as discourse parsers.

(2) **Theoretical Motivations**: Automatically building discourse annotated corpora from parallel texts would provide more resources and evidence to discourse studies in a cross-linguistic perspective. In addition, parallel discourse annotated corpora can provide insight on how explicit discourse relations are affected by the translation process. Modeling such differences is useful in many NLP applications that model the translation process such as Statistical Machine Translation (SMT) (Meyer and Webber, 2013; Meyer and Poláková, 2013).

1.5 Overall Methodology

Figure 1.1 shows an overview of our methodology to project discourse annotations from English onto French. The input to our approach consists of two parallel sentences such as those in Figure 1.1a. As Figure 1.1 shows, we automatically label English discourse connectives with the discourse relations that they signal. To do so, we developed a pipeline of two classifiers called the *CLaC DC Disambiguator* based on the PDTB (see Chapter 3). Figure 1.1b shows the output of the classifier after annotating the discourse connective *since* which signals a *CONTINGENCY.Cause.reason* relation in the English sentence.

Then, we project the discourse annotations from the English discourse connectives onto their French counterparts. For example, as shown in Figure 1.1c, the projection would annotate *car* with

EN: I would ask that they reconsider, since this is not the case.

FR: Je demande que cette décision soit reconsidérée car ce n'est pas le cas.

(a) Sample input parallel sentences from Europarl (≈ 2 millions parallel sentences).

CONTINGENCY.Cause.reason

EN: I would ask that they reconsider, since this is not the case.

FR: Je demande que cette décision soit reconsidérée car ce n'est pas le cas.

(b) Sample of discourse annotation of the English side of Europarl.

In step 1, we automatically tag the 100 English discourse connectives listed in the PDTB with discourse relations. This is done using the *CLaC DC Disambiguator* that we developed for the CoNLL Shared Tasks (see Chapter 3).

CONTINGENCY.Cause.reason

EN: I would ask that they reconsider, since this is not the case.

Annotation Projection

FR: Je demande que cette décision soit reconsidérée car ce n'est pas le cas.

CONTINGENCY.Cause.reason

(c) Sample of the Europarl ConcoDisco corpora

In step 2, we project the discourse annotation of the English discourse connectives onto the French discourse connectives. By varying the word-alignment model used, we create a set of parallel and annotated corpora that we call the Europarl ConcoDisco corpora. From the French side of the Europarl ConcoDisco corpora, we create a PDTB-style discourse annotated corpus for French that we call the FrConcoDisco corpora (see Chapter 4).

Discourse Connective (DC)	Relation
si	*CONTINGENCY.Condition*
si	*COMPARISON.Concession*
lorsque	*CONTINGENCY.Condition*
néanmoins	*COMPARISON.Concession*
...	...

(d) Sample of *ConcoLeDisCo*

In step 3, we use the French discourse connectives listed in LEXCONN and the FrConcoDisco corpora, to map discourse relations to French discourse connectives. We call this lexicon, *ConcoLeDisCo* (see Chapter 5). To remove the dependency to LEXCONN, we propose a new approach, that is independent of statistical word-alignment, to automatically induce a list of French discourse connectives from parallel texts (see Chapter 6).

Figure 1.1: Overall methodology followed in the thesis.

the discourse relation *CONTINGENCY.Cause.reason*. Finding the French connectives onto which the annotations should be projected is based on the alignment between French words and their best English translation within parallel sentences. We used statistical word-alignment models (Brown et al., 1993) to automatically identify these alignments and identify the best translation of French discourse connectives. By varying the word-alignment model used, we created a set of parallel and annotated corpora that we call the Europarl ConcoDisco corpora (our first main resource). From the French side of the Europarl ConcoDisco corpora, we created a PDTB-style discourse annotated corpus for French that we call the FrConcoDisco corpora (see Chapter 4).

Finally, to build lexicons of French discourse connectives (our second main resource), we mined the parallel texts after the projection of discourse annotations. For example, as shown in Figure 1.1d, we identify two discourse relations for the French discourse connective *si*: *CONTINGENCY.Condition* and *COMPARISON.Concession*. We used the FrConcoDisco corpora and the French discourse connectives listed in LEXCONN (Roze et al., 2012; Danlos et al., 2015), to map discourse relations to French discourse connectives. We call this lexicon, *ConcoLeDisCo* (see Chapter 5). Finally, to remove the dependency to LEXCONN, we proposed a new approach, that is independent of statistical word-alignment, to automatically induce a list of French discourse connectives from parallel texts (see Chapter 6).

To evaluate the FrConcoDisco corpora, we proceeded with two methods: 1) an intrinsic evaluation of the discourse annotated corpora using crowdsourcing and 2) an extrinsic evaluation of the discourse annotated corpora using the task of the disambiguation of the usage of French discourse connectives (see Chapter 4). To evaluate *ConcoLeDisCo*, we compared it with LEXCONN, and we manually analyzed a random sample of *ConcoLeDisCo* entries.

1.6 Contributions

Our work has made several practical contributions as well as theoretical contributions to the field of discourse analysis. On the practical side, we have automatically induced two discourse resources for French from the English-French portion of the Europarl parallel corpus (Koehn, 2005); namely:

(1) **The Europarl ConcoDisco corpora**: As shown in Figure 1.1c, the Europarl ConcoDisco

corpora are English-French parallel corpora where the English translation of around 1 million French discourse connectives have been automatically marked. In these corpora, English discourse connectives and French discourse connectives have been automatically annotated with the PDTB discourse relations that they signal. These corpora can be used to provide insight on how explicit discourse relations are affected by the translation process. Furthermore, from the French side of Europarl ConcoDisco we have created the FrConcoDisco corpora: the first PDTB-style discourse annotated corpora. To our knowledge, FrConcoDisco are the first discourse annotated corpora where French discourse connectives are labeled with PDTB discourse relations. Moreover, FrConcoDisco are significant in terms of size as they are more than 25 times larger than the PDTB (Prasad et al., 2008a). These corpora are described in Chapter 4 and in (Laali and Kosseim, 2017b).

(2) **The *ConcoLeDisCo* lexicon:** As shown in Figure 1.1d, *ConcoLeDisCo* is a lexicon of French discourse connectives associated with PDTB discourse relations. While a manually constructed lexicon of discourse connectives already exists for French (LEXCONN; Roze et al., 2012), as we show in (Laali and Kosseim, 2017a), *ConcoLeDisCo* has a different coverage than LEXCONN, and hence is complementary to it. Moreover, *ConcoLeDisCo* constitutes the first lexicon of French discourse connectives mapped to the PDTB relation set[6]. The creation of this lexicon is described in Chapter 6 and in (Laali and Kosseim, 2017a).

In addition to these two main resources, we have **developed the *CLaC DC Disambiguator***. The *CLaC DC Disambiguator* is a pipeline for the disambiguation of discourse connectives. We trained this pipeline for both English and French discourse connectives[7]. To best of our knowledge, the *CLaC DC Disambiguator* is the first tool for the disambiguation of French discourse connectives. We trained the French version of the *CLaC DC Disambiguator* on both a manually annotated corpus extracted from the French Discourse Treebank (FDTB; Danlos et al., 2015) and the induced FrConcoDisco-*Intersection* corpus. The *CLaC DC Disambiguator* achieved an F1-score of 0.766 and 0.546 when trained on these two corpora respectively and tested on the FDTB corpus. The

[6]As discussed in Section 2.1.3, LEXCONN uses a different set of discourse relations than the PDTB.

[7]As explained in Chapter 3, the English version of the *CLaC DC Disambiguator* disambiguates the discourse-usage and also discourse relations of English discourse connectives, but the French version of the *CLaC DC Disambiguator* only disambiguates the discourse-usage of French discourse connectives.

development of *CLaC DC Disambiguator* for English and French discourse connectives was published in (Laali et al., 2015, 2016; Laali and Kosseim, 2016). The features used in this classifier are discussed in Chapter 3 and our method to train it on the FrConcoDisco corpora is described in Chapter 4.

On the theoretical side, we have proposed two novel approaches for discourse annotation projection:

(1) We have **proposed a method to refine the naive method of discourse annotation projection** by filtering unsupported annotations. We have shown that unsupported annotations are typically extracted from parallel sentences where discourse relations are changed from explicit to implicit ones during the translation. Our approach is based on the intersection between statistical word-alignment models and can automatically identify 65% of unsupported projected annotations, which is significantly better than the naive discourse annotation projection. Filtering unsupported annotations using our approach improves the F1-score of the *CLaC DC Disambiguator* by 15% compared to the naive approach used in discourse annotation projection. Our refined approach is described in detail in Chapter 4 and in (Laali and Kosseim, 2017b).

(2) We have also **proposed a novel approach for annotation projection that is independent of statistical word-alignment models.** This approach, explained in Chapter 6 and in (Laali and Kosseim, 2014), is based on sentence alignments followed by the use of statistical tests to mine the sentence aligned parallel corpus. Results show that the proposed approach is more robust to longer French discourse connectives than approaches based on statistical word-alignment models. As shown in (Laali and Kosseim, 2014), this approach can be used to add new discourse connectives to manually constructed lexicons such as LEXCONN (Roze et al., 2012).

1.7 Overview of the Thesis

This thesis is organized as follow: **Chapter 2** briefly explains related work necessary to better appreciate the rest of the thesis. **Chapter 3** describes the development of the *CLaC DC Disambiguator* classifier to automatically disambiguate discourse connectives and reports its performance for English discourse connectives. **Chapter 4** proposes our approach for discourse annotation projection. Typically in annotation projection, it is assumed that linguistic annotations can be projected from one side onto the other side of parallel sentences. In this chapter, we show that this assumption is not always true for discourse annotations because the realization of discourse relations is often changed from explicit to implicit and vice versa during the translation. **Chapter 5** explains how parallel texts and Europarl ConcoDisco can be used to map French discourse connectives to PDTB discourse relations. As a result of this approach, we induced the *ConcoLeDisCo* lexicon where French discourse connectives are mapped to the PDTB relations that they can signal. **Chapter 6** describes our method to extract a list of French discourse connectives from parallel texts and hence eliminate the dependency to statistical word-alignment models. Finally, **Chapter 7** wraps up the thesis and presents conclusions and future work.

Chapter 2

Related Work

This chapter is divided into two main sections: Section 2.1, which describes the discourse resources currently available in the research community and Section 2.2, which focuses on different applications that may benefit from discourse annotation projection.

2.1 Discourse Resources

Our main focus in this section is to introduce two types of discourse resources: discourse annotated corpora (Section 2.1.1) and lexicons of discourse connectives (Section 2.1.2). Next, we describe the discourse resources available specifically for French (Section 2.1.3).

2.1.1 Discourse Annotated Corpora

The content of a text derives from different sources of information. Three major of these sources are semantic and rhetorical information (Hovy, 1995). Semantic information describes an information about the world and/or a perception of it. More precisely, in logic, this semantic information are truth values with respect to the world. The other source of information is the rhetorical intentions of the writer which describes the intention of the writer to relate different parts of text (Mann et al., 1992).

To coherently organize texts and communicate with the reader, the writer semantically and rhetorically connect different part of texts with different relations (e.g. *Justify*, *Elaboration*) which are referred to discourse relations. These relations creates the discourse structure of the text. Discourse structure of texts has been studied from different perspectives, such as linguistic (Halliday, 1985), computational linguistic (Mann and Thompson, 1987; Hobbs, 1990), psychology (Sanders et al., 1992), logic (Asher, 1993), etc. Hence, various theories have been proposed for analyzing the discourse structure of texts, such as Rhetorical Structure Theory (RST; Mann and Thompson, 1987), Segmented Discourse Representation Theory (SDRT; Asher and Lascarides, 2003) and Discourse Lexicalized Tree Adjoining Grammar (D-LTAG; Webber et al., 2003).

Regardless of discourse theories, annotating the discourse structure of texts is very costly and requires expert human annotators. Consequently, only a few discourse theories possess a formal annotation manual and a large manually annotated corpus. In this section, we present an overview of four discourse theories and their associated discourse annotated corpora. A complete discussion of discourse theories is beyond of the scope of this thesis, however, the interested reader may follow the references provided.

Most discourse annotated corpora were initially proposed for English (Carlson et al., 2001; Reese et al., 2007; Prasad et al., 2008a). Subsequently, the annotation schema of some of these corpora were adopted for other languages to build similar corpora for these languages by exploiting the discourse annotation experience with English (e.g. (Zhou et al., 2012)). In the following sections, after briefly overviewing discourse theories, we introduce the corresponding discourse annotated corpora for English and then present similar corpora for other languages.

2.1.1.1 Rhetorical Structure Theory

Rhetorical Structure Theory (RST; Mann and Thompson, 1987) proposed the notion of a nucleus-satellite view on rhetorical relations, in which the span of the satellite text plays a subordinate role to the main nucleus text. RST schemas are recursive (i.e. embedded discourse relations are allowed). This leads to textual discourse structures to be represented as trees in RST. Figure 2.1 shows the RST tree of (Ex. 6). The arrows in the figure are labelled with the name of the rhetorical relation and point to the nucleus span.

(Ex. 6) 1. [Title:] The Perception of Apparent Motion

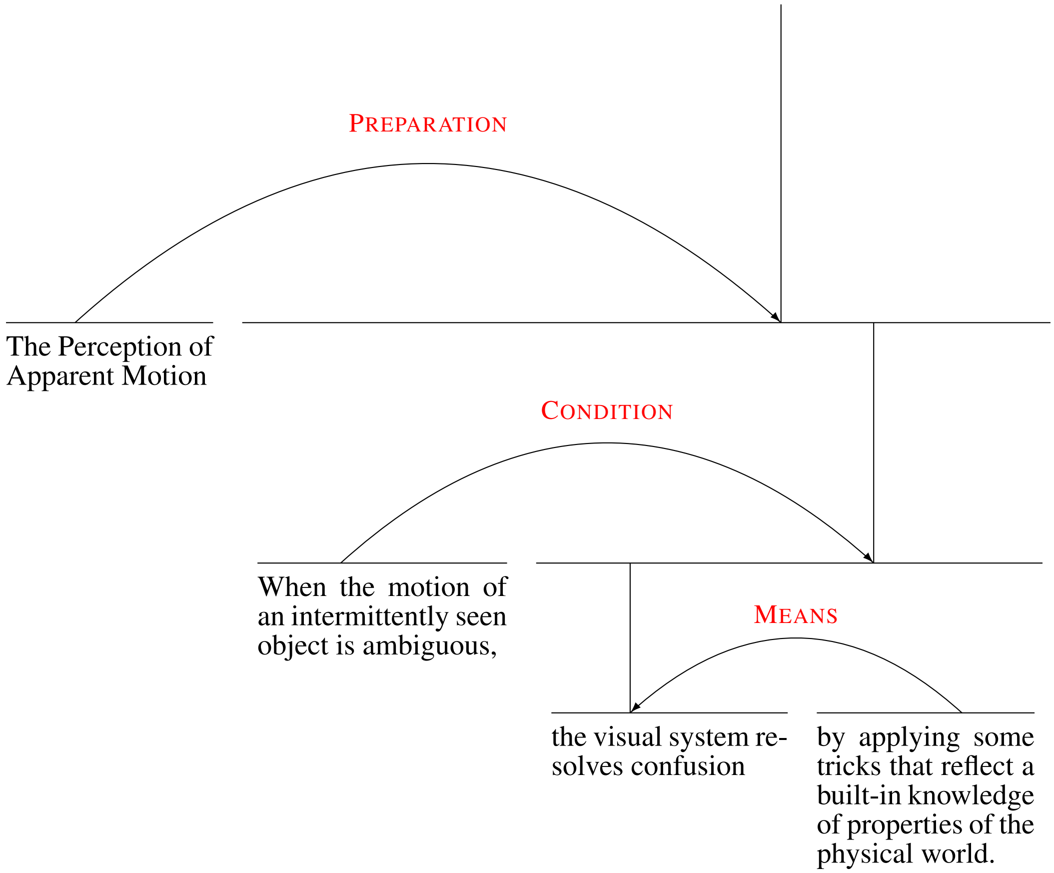


Figure 2.1: RST discourse tree for (Ex. 6)

2. [Abstract:] When the motion of an intermittently seen object is ambiguous, the visual system resolves confusion by applying some tricks that reflect a built-in knowledge of properties of the physical world.[1]

The Column *RST Relations* in Table 2.1 shows the original set of 23 discourse relations that have been defined based on the intention of writer/speaker (Mann and Thompson, 1988). Later, Carlson and Marcu (2001) extended these relations and defined 78 discourse relations. These are shown in the Column RST-DT Relations in Table 2.1. See Mann and Thompson (1987); Carlson et al. (2001); Taboada and Mann (2006) for more details about RST.

For English, there exist two corpora manually annotated with RST: the RST Discourse Treebank (RST-DT; Carlson et al., 2001) and the Discourse Relations Reference Corpus (Taboada and Renkema, 2008). The RST-DT (Carlson et al., 2001) is one of the first discourse annotated corpora

[1]The example was taken from (Taboada and Mann, 2006).

Original RST Relations	RST-DT Relations		
Elaboration	analogy	interpretation-s	temporal-before
Circumstance	antithesis	manner	temporal-same-time
Solutionhood	attribution	means	Analogy
Volitional Cause	attribution-n	otherwise	Cause-Result
Volitional Result	background	preference	Comment-Topic
Non-Volitional Cause	circumstance	problem-solution-s	Comparison
Non-Volitional Result	comment	purpose	Conclusion
Purpose	comparison	question-answer-s	Consequence
Condition	concession	reason	Contrast
Otherwise	conclusion	restatement	Contrast
Interpretation	condition	rhetorical-question	Disjunction
Evaluation	consequence-s	statement-response-s	Evaluation
Restatement	contingency	summary-s	Interpretation
Summary	definition	temporal-same-time	Inverted-Sequence
Sequence	elaboration-additional	topic-drift	List
Contrast	elaboration-general-specific	topic-shift	Otherwise
Motivation	elaboration-object-attribute	cause	Problem-Solution
Antithesis	elaboration-part-whole	consequence-n	Proportion
Background	elaboration-process-step	evaluation-n	Question-Answer
Enablement	elaboration-set-member	interpretation-n	Reason
Evidence	enablement	problem-solution-n	Sequence
Justify	evaluation-s	question-answer-n	Statement-Response
Concession	evidence	result	Temporal-Same-Time
	example	statement-response-n	Topic-Comment
	explanation-argumentative	summary-n	Topic-Drift
	hypothetical	temporal-after	Topic-Shift

Table 2.1: The set of 23 RST relations proposed by Mann and Thompson (1988) and the expanded list of 78 RST relations proposed by Carlson and Marcu (2001).

and the largest one that is based on RST. This corpus contains the annotations of 385 texts from the *Wall Street Journal (WSJ)*. On the other hand, the Discourse Relations Reference Corpus includes 65 texts (each one tagged by one annotator) of several types and from several sources (21 articles from the Wall Street Journal extracted from the RST-DT, 30 movies and books' reviews extracted from the epinions.com website, and 14 diverse texts, including letters, websites, magazine articles, newspaper editorials, etc.).

RST corpora have been also developed for other languages. While most of these corpora are rather small for computational applications, they are still large enough to show the applicability of the RST annotation schema for other languages. These corpora include Rhetalho (50 texts) (Pardo and Seno, 2005) and the CorpusTCC (100 texts) (Pardo et al., 2008) for Portuguese, the Potsdam Commentary corpus (175 German newspaper commentaries) (Stede, 2004; Stede and Neumann, 2014) for German, the Discourse-Annotated Dutch Text Corpus (80 texts) for Dutch and the RST Spanish Treebank (267 texts) (Da Cunha et al., 2011).

Wolf and Gibson (2005) questioned the adequacy of a tree-like structure for modelling discourse relations. They claim that a more complex structure such as a graph structure is required to represent discourse relations of texts. To show their framework, they released graph-based discourse annotations of 135 articles in a corpus called the Discourse Graphbank.

2.1.1.2 Segmented Discourse Representation Theory

Segmented Discourse Representation Theory (SDRT; Asher and Lascarides, 2003) is a more recent discourse theory which focuses on extending existing theories of sentence semantics to the discourse level. SDRT uses a graph-based representation, with long distance attachments. In SDRT, discourse relations are divided into two categories: subordinating and coordinating discourse relations which appear to echo the nucleus-satellite view in RST. Moreover, SDRT also distinguishes veridical from non-veridical relations. For veridical relations, the content of both arguments of relations have to be true, whereas for non-verdical relations at least one of arguments does not need to be true. Table 2.2 shows the set of 14 discourse relations defined in SDRT and their categories (Reese et al., 2007). See Asher and Lascarides (2003); Lascarides and Asher (2007); Muller et al. (2012) for more details about SDRT.

Coordinating Relations		Subordinating Relations	
Veridical	**Nonveridical**	**Veridical**	**Nonveridical**
Continuation	Consequence	Background	Attribution
Narration	Alternation	Elaboration	
Result		Explanation	
Contrast		Commentary	
Parallel		Source	
Precondition			

Table 2.2: The set of 14 discourse relations defined in SDRT.

Figure 2.2 shows the discourse representation of (Ex. 7) using SDRT. Intuitively, π_i represents the discourse entities refered to in (Ex. 7) and K_{π_i} indicates the constraints (properties, relations) on those discourse entities. Each discourse relation (e.g. *Elaboration*, *Narration*) also adds more restriction on the discourse entities. In Figure 2.2, while *Elaboration* is a subordinate discourse relation, *Narration* is a coordinate discourse relation.

(Ex. 7) π_1 . John had a great evening last night.

π_2 . He had a great meal.

π_3 . He ate salmon.

π_4 . He devoured lots of cheese.

π_5 . He won a dancing competition.[2]

A few discourse annotated corpora are based on SDRT. These include the DISCOR corpus (Reese et al., 2007) for English, ANNODIS (Afantenos et al., 2012) and CASOAR (Farah et al., 2016) for French, as well as the SDRT discourse annotated corpus for Arabic (Keskes, 2015). All these corpora are publicly available, except for the DISCOR corpus.

2.1.1.3 Discourse Tree Banks

Webber and Joshi (1998) have proposed a tree-adjoining grammar for discourse called Discourse Lexicalized Tree Adjoining Grammar (D-LTAG; Webber et al., 2003) which aims to extend syntax beyond the sentence. As with LTAG (Joshi and Schabes, 1997), D-LTAG uses lexicalized tree

[2]The example was taken from (Lascarides and Asher, 2007).

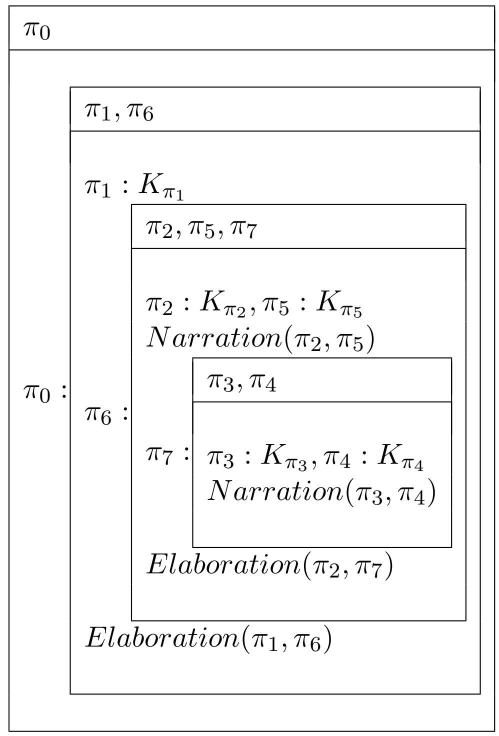


Figure 2.2: The discourse structure of (Ex. 7) in the SDRT framework.

structure elements to describe the discourse structure. This approach provides a uniform way to process texts at both the clause level and at the discourse level and opening up the possibility of sentence processing and low-level discourse processing being carried out in an integrated fashion. From D-LTAG, the Penn Discourse Treebank (Prasad et al., 2008a) project was born.

In 2008, Prasad et al. (2008a) released the Penn Discourse Treebank (PDTB). This corpus is currently the largest publicly available discourse annotated corpus and has been adopted by many languages. Following the view in D-LTAG, the PDTB treats lexical elements called discourse connectives as discourse-level predicates that take two clausal arguments representing abstract objects such as events, states and propositions. If a discourse relation is expressed without any explicit discourse connective, annotators inserted an *inferred discourse connective* which conveys the same discourse relation between the text spans. As a consequence of this annotation schema, discourse relations are divided into two categories: explicit discourse relations (former) and implicit discourse relations (latter). A set of 41 discourse relations which are hierarchically organized in three levels (see Figure 2.3) is used in the PDTB. Such a hierarchical organization helps to increase the inter-annotator agreement, by allowing the annotators to select a tag at the level they are comfortable with. The full annotation guideline of this corpus is available in (Prasad et al., 2008b). See (Webber et al., 2003; Miltsakaki et al., 2004; Prasad et al., 2004, 2008a,b) for more detailed information about the PDTB.

(Ex. 8) and (Ex. 9) show the PDTB annotations for an explicit discourse relation and an implicit discourse relation respectively. Following the PDTB standard, in these examples, the discourse connective is underlined, the first argument of the discourse connective is in *italic*, the second argument is in **bold** and the relation is marked at the end of the sentences in parentheses. In (Ex. 8), a *CONTINGENCY:Cause:result* discourse relation is explicitly signaled by the explicit discourse connective *so*. On the other hand, in (Ex. 9), the *EXPANSION:List* relation is implicit between *the first argument* and **the second argument**. In this example, the discourse connective *and* has been inferred by the reader and inserted between the two discourse arguments.

(Ex. 8) *In addition, its machines are typically easier to operate,* so **customers require less assistance from software.** (*CONTINGENCY:Cause:result*)

(Ex. 9) But other than the fact that besuboru is played with a ball and a bat, it's unrecognizable: *Fans politely return foul balls to stadium ushers;* Implicit = AND **the strike zone expands depending on the size of the hitter;** (*EXPANSION:List*)

In the PDTB, only low-level discourse structures are indicated and relations between two text spans are tagged. In other words, no embedding discourse relations exist in the corpus.

The PDTB contains a large number of texts and has a high inter-annotator agreement. Currently, the PDTB covers all the *Wall Street Journal* corpus (2159 articles) and contains 1 million words. Due to its large size, this corpus was used in different discourse-related applications such discourse parsing (Xue et al., 2015, 2016).

The PDTB's approach for annotating discourse relations has been widely adopted to create discourse treebanks in other languages such as Turkish (Zeyrek et al., 2010), Chinese (Zhou et al., 2012), Arabic (Al-Saif and Markert, 2010), Czech (Mladová et al., 2008), Hindi (Oza et al., 2009) and French (Danlos et al., 2015). However, the scope of some of these corpora is limited due to

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

Figure 2.3: Hierarchy of discourse relations in the PDTB

the high-cost of manually developing PDTB-style corpora. For example, the scope of discourse annotations was limited to explicit discourse relations in The Leeds Arabic Discourse Treebank (Al-Saif and Markert, 2010).

2.1.1.4 Differences and Commonalities across Discourse Theories

The discourse theories discussed in Sections 2.1.1.1-2.1.1.3, exhibit two major differences in their underlying assumptions:

(1) *Representation of Discourse Structure*: Different theories and corpora allow for different structures to represent discourse. RST (see Section 2.1.1.1) assumes a tree representation that covers the entire text; the Discourse Graphbank (see Section 2.1.1.1) uses general graphs that allow multiple parents and crossing; while SDRT (see Section 2.1.1.1) uses directed acyclic graphs that allow for multiple parents, but does not not for crossing. Finally, the PDTB (see Section 2.1.1.3) does not represent the full discourse structure of texts. Instead, discourse structures are flat and may not be fully connected. Nevertheless, the PDTB does not impose any constraints on the text spans as realizations of Arg1 and Arg2, including single- or multi-paragraph long texts. This allows the PDTB to be theory-neutral with respect to discourse structures.

(2) *Basis Used to Define Discourse Relations*: While SDRT and PDTB use the content of the arguments to define discourse relations; RST provides definitions for the relations in terms of the intended effects on the hearer/reader.

In spite of these differences, there are also strong commonalities between these frameworks. In particular, all theories make a distinction between relations that relate facts about the world and relations where the semantic content of the discourse arguments involve an implicit belief. For example, consider the sentence (Ex. 10). In this example, there is no causal relation between John's sending of the message and John not being at work, but rather the sending of the message caused the speaker/writer to *believe* that John is not at work.

(Ex. 10) John is not at work today, because he sent me a message to say he was sick.[3]

[3] The example was taken from (Bunt and Prasad, 2016).

This distinction is referred to as ‘content-metatalk’ in SDRT and ‘semantic-pragmatic‘ in RST. The PDTB also defines a few such pragmatic discourse relations for *CONTINGENCY* and *COMPARISON* relations (see Figure 2.3).

The similarities across frameworks have motivated several studies to unify the annotation of discourse relations (e.g. (Hovy, 1990; Maier and Hovy, 1993; Hovy, 1995; Zitoune and Taboada, 2015; Scheffler and Stede, 2016; Bunt and Prasad, 2016; Demberg et al., 2017)). For example, Maier and Hovy (1993) organize discourse relations in three categories based on three metafunctions of languages proposed by Halliday (1985), namely *ideational*, *interpersonal* and *textual*:

(1) *Ideational relations*: These relations convey semantic information between *abstract objects* in the world of our imagination. Recognizing these relations by the reader/listener will increase their knowledge about the world.

(2) *Interpersonal relations*: These relations affect the reader’s/listener’s belief, attitude, the ability to understand or desire to perform an action.

(3) *Textual relations*: These relations serve to organize the text itself. For example, they allow to conjunct different pieces of text logically.

Using these main categories, Maier and Hovy (1993) have been able to merge discourse relations from different theories collected by Hovy (1990) and organized them into a hierarchy of 44 discourse relations.

It is important to recognize that while in most discourse theories, the inventory of discourse relations is assumed to be fixed, it is also well-accepted that such an inventory should be open and allow for further expansion (Sanders et al., 1992; Maier and Hovy, 1993; Bunt and Prasad, 2016). For example, Kittredge et al. (1991) have argued that to model the discourse structure of texts in *sub-languages*, it is necessary to define highly domain-specific relations.

This concludes our discussion on discourse frameworks and annotated corpora. In the next section, we will discuss lexicons of discourse connectives which is the second resource that we want to extract from parallel texts.

2.1.2 Lexicons of Discourse Connectives

Discourse connectives are terms like *however*, *because* and *while* that explicitly signal a discourse relation within texts. One of the main characteristics of discourse connectives is that they relate two different abstract objects in a discourse such as events, states or propositions (Asher and Lascarides, 2003), also referred to as *discourse arguments* (Prasad et al., 2008a) or *elementary discourse units (EDUs)* (Mann and Thompson, 1987). The usage of discourse connectives does not always signal a discourse relation and may be ambiguous at two levels: first, they can be used in *discourse-usage* or *non-discourse-usage*, and second, they may be used to signal more than one discourse relation (see Chapter 3 for more details).

Even if there is no consensus on the formal definition of discourse connectives, all discourse theories recognize the central role of connectives in the identification of discourse relations (Asr and Demberg, 2012; Drenhaus et al., 2014; Millis et al., 1995; Murray, 1995, 1997).

One approach to identifying discourse connectives is to apply linguistic tests. For example, Roze et al. (2012) proposed the following guidelines for the identification of discourse connectives:

(1) Discourse connectives cannot be part of a subject, an object or an adverbial.

(2) Discourse connectives cannot be substituted (partly or entirely) by an entity (person, event, discourse unit) of the context.

(3) Discourse connectives are lexically fixed and invariable.

Despite the common function of discourse connectives to link the content of two different textual units, the grammatical category of discourse connectives is syntactically heterogeneous. The most frequent categories of discourse connectives are coordinating and subordinating sentence conjunctions, but discourse connectives also include other syntactically categories such as multi-word items with conjunction-like behaviour (e.g. *as soon as*, *as long as*), and single- or multi-word adverbials that show anaphoric, rather than syntactic, linking behavior (e.g., *for example*, *in addition*, *on the contrary*).

The PDTB restricts discourse connectives to three main grammatical categories: 1) subordinating conjunctions (e.g. *because*, *when*, *since*, *although*), 2) coordinating conjunctions (e.g. *and*, *or*, *nor*) and 3) adverbial phrases and prepositional phrases such as (e.g. *however*, *otherwise*, *then*, *as a result*, *for example*). According to the PDTB, other lexical elements that signal discourse relations and do not fall in these three grammatical categories are called AltLex. (Ex. 11) to (Ex. 14)[4] illustrate the use of subordinates, coordinates, adverbials and AltLexes to signal discourse relations respectively.

(Ex. 11) *Knowing a tasty – and free – meal* when **they eat one,** the executives gave the chefs a standing ovation. (TEMPORAL:Synchrony)

(Ex. 12) *Those looking for real-estate bargains in distressed metropolitan areas should lock in leases* or **buy now.** (EXPANSION:Alternative:disjunctive)

(Ex. 13) *Chairman Krebs says the California pension fund is getting a bargain price that wouldn't have been offered to others.* In other words, **The real estate has a higher value than the pending deal suggests.** (EXPANSION:Restatement:equivalence)

(Ex. 14) *After trading at an average discount of more than 20% in late 1987 and part of last year, country funds currently trade at an average premium of 6%.* AltLex [The reason:] **Share prices of many of these funds this year have climbed much more sharply than the foreign stocks they hold.** (CONTINGENCY:Cause:reason)

Because a single connective may be used to signal a variety of relations (and vice-versa), lexicons of discourse connectives containing a list of discourse connectives associated with the discourse relations that they can signal have been built. For example, according to the PDTB, the discourse connective *while* may signal a *TEMPORAL:Synchronous*, *COMPARISON:Contrast* or an *EXPANSION:Conjunction*. Lexicons of discourse connectives can be very useful for discourse studies (e.g. developing discourse annotated corpora (Prasad et al., 2008a; Danlos et al., 2012; Poláková et al., 2013; Al-Saif and Markert, 2010), automatic discourse analysis (Xue et al., 2015; Lin et al., 2014), etc.). Currently, such lexicons are available for English (Knott, 1996), Spanish (Alonso Alemany et al., 2002), German (Stede and Umbach, 1998), Czech (Mírovsky et al., 2016) and French (Roze et al., 2012).

[4]All examples are taken from (Prasad et al., 2008b).

Similarly to the creation of discourse annotated corpora, building lexicons of discourse connectives is not an easy task. To build such lexicons, an extensive corpus study is typically performed. For example, (Knott, 1996) manually analyzed 226 pages of text to build a lexicon of 200 phrases that can function as discourse connectives. Then, he applied different linguistic tests to associate them with the discourse relations that they signal. Even such a comprehensive study may miss some discourse connectives. For example, (Knott, 1996) did not list the discourse connective *in order to* in his lexicon. Interestingly, *in order to* was not listed in the list of discourse connectives used in the PDTB either, even though there are 50 occurrences of this connective in the Wall Street Journal. Our approach (see Chapter 5 and 6) can reduce the effort needed to build such lexicons by automatically mining parallel texts to find evidence that shows that an expression is a discourse connective and/or a discourse connective may signal a discourse relation.

2.1.3 Discourse Resources For French

To the best of our knowledge, there exist only three publicly discourse resources for French:

(1) LEXCONN (Roze et al., 2012): a lexicon of French discourse connectives and

two discourse annotated corpora:

(2) ANNODIS (Afantenos et al., 2012)

(3) the French Discourse Treebank (FDTB; Danlos et al., 2015) (which was briefly discussed in Section 2.1.1.2).

LEXCONN (Roze et al., 2012) is a manually built lexicon of French discourse connectives. The project was initiated in 2010 and released its first edition of the lexicon in 2012. The latest version, LEXCONN V2.1 (Danlos et al., 2015), contains 371 discourse connectives where 343 are mapped to an average of 1.3 discourse relations taken from various sources including RST (see Section 2.1.1.1), SDRT (see Section 2.1.1.2) and PDTB (see Section 2.1.1.3). Moreover, discourse connectives are categorized based on their syntactic categories and divided into two types: *subordinate* and *coordinate* (cf. Section 2.1.1.2). This project is ongoing as 38 discourse connectives still have not been assigned to any discourse relation. See Table 2.3 for a few entries of LEXCONN.

Discourse Connective	Category	Type	Relation
afin de, afin d’	prep	coord	goal
Exemple: Paul a économisé toute l’année (afin de/pour) pouvoir partir en vacances cet été. **Synonymes:** pour			
alors	adv [position: initiale]	coord	result*
Exemple: Marie a l’air tendue. Alors les nouvelles doivent être mauvaises. **Exemple:** Marie a l’air tendue. Les nouvelles doivent être mauvaises, alors. **Synonymes:** donc			

Table 2.3: Sample of an entry in LEXCONN

The ANNODIS corpus (Afantenos et al., 2012) is a discourse annotated corpus where both high-level structures (e.g. topical chains) and local structures (i.e. discourse relations between text spans) of texts have been annotated. Two perspectives on discourse were used in the discourse annotation of ANNODIS: a bottom-up view and a top-down view. The bottom-up view incrementally builds a discourse structure from clauses and links them with discourse relations while the top-down view focuses on text-organizing strategies realized at different levels of textual granularity (from less than a paragraph to several sections). The bottom-up approach resulted in the annotation of 86 documents (short Wikipedia articles as well as news articles) based on SDRT with a total of 3199 text segments and 3355 relations.

The second discourse annotated corpus for French is the French Discourse Treebank (FTB; Danlos et al., 2012). Although the FDTB is based on the PDTB, it differs at a theoretical level. The FDTB plans to provide a full coverage of texts so that the textual discourse structures are fully connected. This is not the case in the PDTB. Moreover, Danlos et al. defined a new hierarchy of discourse relations based on a mixture of the relations in RST (see Section 2.1.1.1), SDRT (see Section 2.1.1.2) and the PDTB (see Section 2.1.1.3) to annotate discourse relations. Currently, the first version of the FDTB (Danlos et al., 2015) contains more than 10,000 instances of LEXCONN’s French discourse connectives annotated as *discourse-usage* in two syntactically annotated corpora: the Sequoia Treebank (Candito and Seddah, 2012) and the French Treebank (FTB) (Abeillé et al., 2000). Out of 343 discourse connectives listed in LEXCONN, only 229 connectives appeared in the FDTB. Moreover, to date, discourse connectives have not been annotated with discourse relations in the FDTB. Figure 2.4 shows a sample annotation in the FDTB.

```
<ARTICLE id="1016">
<SENT id="flmf3_11000_11499ep-11025">Les syndicats ont évidemment
    été " surpris " par une opération rondement menée , qui doit
   faire l' objet de réunions des comités d' entreprise ,
   mercredi 17 janvier et vendredi 19 à UTA . </SENT>
<SENT id="flmf3_11000_11499ep-11026">La fédération CGT des
   transports s' est élevée contre " l' absence de concertation "
    <CONN>et</CONN> estime que les salariés " n' ont rien de bon
   à attendre de cette restructuration " . </SENT>
<SENT id="flmf3_11000_11499ep-11027">À Air France , les repré
   sentants syndicaux au conseil d' administration , reçus
   vendredi 12 au soir par la direction , estiment n' avoir
   obtenu pour l' instant des informations " très formelles " sur
    les implications économiques ou sociales ; toutefois , CFDT
   et CFTC sont plutôt satisfaits , <CONN>tandis que</CONN> FO
   affirme avoir obtenu des assurances sur l' emploi . </SENT>
<SENT id="flmf3_11000_11499ep-11028">À UTA , <CONN>en revanche</
   CONN> , les syndicats , reçus par leur PDG vendredi , dé
   noncent avec " indignation " le manque de concertation . </
   SENT>
<SENT id="flmf3_11000_11499ep-11029"><CONN>Cependant</CONN> , le
   SNPC ( navigants commerciaux ) estiment que la situation ne
   peut être pire que celle des derniers mois . </SENT>
</ARTICLE>
```

Figure 2.4: A sample annotation of discourse connectives in the FDTB.

2.2 Applications

In this thesis, we explore the use of discourse annotation projection in order to induce a PDTB-style discourse annotated corpus for French. In this section, we situate our work with respect to three NLP tasks that can benefit from our work.

2.2.1 Inducing Discourse Resources

Annotation projection has been widely used in the past to build natural language applications and resources. It has been applied for part-of-speech tagging (Yarowsky et al., 2001), word sense disambiguation (Bentivogli and Pianta, 2005) and dependency parsing (Tiedemann, 2015). As discourse relations are semantic and rhetorical in nature, in the translation process, in principle, they should transfer from the source language to the target language. This property of discourse relations makes them an attractive target for annotation projection. As a consequence, annotation projection has been recently used to produce discourse resources (Versley, 2010; Laali and Kosseim, 2014; Hidey and McKeown, 2016). Among these, Versley (2010) projected English discourse connectives to their counterparts in German in a parallel corpus. Doing this, he produced a corpus where discourse vs. non-discourse usage of German discourse connectives are annotated. He then used this corpus to train a discourse parser for German. To evaluate the induced parser, Versley manually annotated discourse relations in a subset of the TüBa-D/Z corpus (Telljohann et al., 2006) (5,000 words). The induced parser achieve an F-score of 68.7% when a list of discourse connectives is given and an F-score 57.5% when the list of discourse connectives are extracted from the parallel texts using a rule-based system. Although Versley (2010) used a list of discourse connectives in generating the corpus, he also tried to automatically induce the discourse connectives from his corpus.

Similarly to previous work that used annotation projection (e.g. (Tiedemann, 2015)), Versley (2010) implicitly assumed that linguistic annotations can be projected from one side onto the other side of parallel sentences. In this thesis, we pay special attention to parallel sentences for which this assumption does not hold and therefore, the projected annotations are unreliable (see Chapter 4). Moreover, Versley (2010) did not explicitly evaluate the induced discourse annotated corpus or the list of discourse connectives, but rather focused on the evaluation of the parser. In this thesis, we propose a linguistic test which we refer to as the *translatable* test to evaluate the induced annotated corpus using crowdsourcing (see Chapter 4). Moreover, not only did we extract a list of discourse connectives, but we associated these discourse connectives to discourse relations and induced a lexicon of French discourse connectives (see Chapter 6). Finally, Versley (2010) has solely employed statistical word-alignment models to find discourse connectives. However, our results show (see Chapter 6) that statistical word-alignment models is not sufficient to align discourse connectives. To address this problem, we propose a new approach which is based on sentence alignments followed by the use of statistical tests to mine the sentence aligned parallel corpus (see Chapter 6).

2.2.2 Machine Translation Systems

While recently, Machine Translation (MT) has dramatically improved the quality of automatically translated texts at the sentence level (Chung et al., 2016; Luong and Manning, 2016; Firat et al., 2016), these systems do not typically preserve discourse phenomena (Meyer and Webber, 2013; Li et al., 2014b; Scarton, 2016). For example, pronouns typically do not map well across languages and their translations depend on many factors such as gender, number, case, formality, or humanness. The differences in where pronouns can be used in different languages often leads to incorrect translations. To exemplify this problem, let us consider the translation of the English pronoun *it* into French. There are many French candidate translations for *it* such as *il*, *elle*, or *cela* which should be picked based on the antecedent of the pronoun. Finding the antecedent of pronouns is an important topics in discourse analysis and is highly related to the discourse structure of texts (Asher and Lascarides, 2003).

Most current approaches to statistical machine translation assume that sentences in a text are independent and do not account for inter-sentential discourse properties. Moreover, metrics such as the BLEU score (Papineni et al., 2002) used for the evaluation of MT systems disregard document-wide discourse information (Scarton, 2016). However, considering discourse relations and textual discourse structure, in general, can help machine translation systems in several ways. For example, Chinese allows very long sentences and often express multiple discourse relations in a single sentence (Li et al., 2014b). These long sentences are typically translated into multiple sentences when they are translated into English. Another example is the case where discourse connectives are highly ambiguous (e.g. *while* can signal a *TEMPORAL:Synchronous* or a *COMPARISON:Contrast* according to the PDTB (Prasad et al., 2008b)) or where the target language uses other syntactic construction than a connective to convey the discourse relation. Meyer and Poláková (2013) showed that training a phrase-base machine translation system such as Moses (Koehn et al., 2007) on an English-Czech parallel corpus where discourse connectives were annotated with PDTB discourse relations leads to translation performance improvement between 4-60% for these cases.

In this thesis, we add discourse annotations on both sides of parallel texts. The annotated parallel texts are a valuable resource for identifying differences between languages, with the goal of achieving better translation models that use discourse annotations (cf. (Meyer and Poláková, 2013)).

2.2.3 Contrastive Discourse Studies

Contrastive linguistics is the study of two or more languages, for applied or theoretical purposes (Johansson, 2000). Currently, most work in contrastive linguistics has focused on aspects of the grammatical system, examining phonological, morphological, lexical and syntactic similarities and differences across languages (Taboada and de los Ángeles Gómez-González, 2012) (see (Johansson, 2007) for a history of contrastive linguistics). Recently, linguists have also showed interest in cross-lingual analysis of discourse phenomena. Much of these studies use parallel corpora and corpus linguistics techniques to study language (Taboada and de los Ángeles Gómez-González, 2012; Zufferey and Degand, 2014; Zufferey and Gygax, 2015; Hoek and Zufferey, 2015). A complete survey of contrastive linguistics is beyond of the scope of this thesis. In this section, we only summarize two families of contrastive linguistics that are related to our work and focus on the translation of discourse connectives in parallel texts:

(1) Linguistic studies on the meaning of discourse relations and discourse connectives.

(2) Cognitive studies on the use of explicit and implicit discourse relations.

2.2.3.1 Linguistic Studies on the Meaning of Discourse Relations and Discourse Connectives

Discourse connectives play an important role in the identification of discourse relations. As suggested by Knott (1996), discourse connectives can be considered as linguistic evidence for discourse relations and by analyzing their usage in texts, we can define a hierarchy of discourse relations. Similarly, studies on the translation of discourse connectives in parallel texts can enrich the definition of discourse relations.

Zufferey and Cartoni (2012) studied two important characteristics of the *Cause* discourse relation: 1) the notion of *domain of use* and 2) the *information of the status* of the *Cause* segments. According to Zufferey and Cartoni (2012), the domains of use for the *Cause* discourse relation can be real-world uses (Ex. 15), epistemic uses (Ex. 16) or speech act uses (Ex. 17).

(Ex. 15) *The snow is melting* <u>because</u> **the sun is shining.**

(Ex. 16) *John must be ill,* <u>because</u> **he did not come to work today.**

(Ex. 17) *Is anybody coming to the party?* <u>Because</u> **it is time to go.**[5]

Regarding the information of the status of the *Cause* segments, the status can either be new or given if the speaker considers that the listener is not aware of the cause or it is part of the common ground respectively. For example, in (Ex. 18), the speaker introduces a given information to indicate why the report is important and in (Ex. 19), the speaker provides a new information that justifies why she welcomes the President.

(Ex. 18) Madam President, *this is a very technical but important report* <u>since</u> **we are dealing with the question of food safety and hygiene.**

(Ex. 19) *I welcome the President-in-Office to Parliament officially* <u>since</u> **it is the first time I have had this direct contact with him.** [5]

To study these characteristics, Zufferey and Cartoni (2012) manually annotated these characteristics for three English and three French causal discourse connectives (*because, since, as, parce*

[5] All examples are taken from (Zufferey and Cartoni, 2012).

que, car, puisque) in parallel texts and showed that the translation of these discourse connectives is directly influenced by these characteristics.

Zufferey and Degand (2014) studied the meaning of discourse connectives in five Indo-European languages of the Germanic and Romance families: English, French, German, Dutch and Italian. To do so, they constructed a small parallel corpus (around 2,500 words for each language) and projected English discourse connectives to their translation in the other languages. Then, they associated a PDTB discourse relation to each discourse connective independently of their translation in other languages. The disagreement between annotators provides insight to refine the PDTB discourse relation hierarchy and its annotation manual for annotating discourse relations for multilingual purposes.

2.2.3.2 Cognitive Studies on the Use of Explicit and Implicit Discourse Relations

As noted in Section 2.1.1.3, discourse relations can either be explicitly marked by discourse connectives or implicitly conveyed. An important question in discourse studies, from both a theoretical and an applied point of view, is how speakers choose between the two options to signal discourse relations (Taboada, 2009; Asr and Demberg, 2012; Das and Taboada, 2013; Drenhaus et al., 2014; Zufferey and Gygax, 2015; Hoek and Zufferey, 2015; Yung et al., 2017). To answer this question, one hypothesis is that readers and listeners have certain expectations about discourse relations and those discourse relations that are in line with readers' and listeners' expectations are more often implicit than the ones that are not. This hypothesis has been traditionally studied in monolingual corpora (Asr and Demberg, 2012; Das and Taboada, 2013), but recently, researchers have shown an interest in testing this hypothesis in parallel texts (Hoek and Zufferey, 2015).

Hoek and Zufferey (2015) analyzed the implicitness of discourse relations from a multilingual perspective. To do so, they randomly selected around 1,000 parallel sentences that contain one of *although*, *because*, *also*, or *if* discourse connectives from Europarl Direct (Koehn, 2005; Cartoni et al., 2013). Then, Hoek and Zufferey (2015) manually analyzed the parallel sentences based on how the discourse connectives were translated: explicitly, implicitly, or by means of a paraphrase or syntactic construction. According to their results, the existing hypotheses about readers'/listeners' expectations are not sufficient to explain the implicitness of discourse relations. Hoek and Zufferey (2015) proposed that the rate of implicitness of discourse relations depends on the cognitive complexity of discourse relations.

As indicated in Section 1.6, an important contribution of our thesis is the automatic annotation of explicit discourse relations on both sides of parallel sentences. Cognitive studies on the use of explicit and implicit discourse relations can benefit from Europarl ConcoDisco to validate their hypothesis on a larger corpus for variety of discourse connectives and discourse relations.

2.3 Conclusion

In this chapter, we have described two important discourse resources, namely discourse annotated corpora and lexicons of discourse connectives. We have also listed the discourse resources currently available in the research community for English and other languages. In particular, we have reviewed three discourse resources for French: LEXCONN, ANNODIS and the FDTB. We also discussed why the PDTB framework is the most suitable framework for our work.

In Sections 2.2, we have introduced three applications that can benefit from discourse annotation projection: 1) the induction of discourse resources, 2) machine translation and 3) contrastive discourse studies.

In the next chapter, we present our pipeline to disambiguate discourse connectives. We extensively use this pipeline in the rest of thesis in our approach to discourse annotation projection.

Chapter 3

On the Disambiguation of Discourse Connectives

With respect to discourse organization, discourse connectives constitute the most basic way of signaling the speaker's or writer's intentions. They provide an important clue to disambiguate discourse relations whose interpretations would be opaque without them (Asr and Demberg, 2012; Drenhaus et al., 2014; Millis et al., 1995; Murray, 1995, 1997). Discourse connectives can be ambiguous at two levels:

(1) they can be used in *discourse-usage* or *non-discourse-usage*, and

(2) they may be used to signal more than one discourse relation.

In this chapter, we focus on our first research questions (see Section 1.2):

(Q. 1) Can English discourse connectives be automatically annotated?

(Q. 1) is important because, as we will see in Chapter 4, we have projected annotations of English discourse connectives onto the French side to build Europarl ConcoDisco and FrConcoDisco-*Intersection*. Therefore, being able to automatically disambiguate discourse connectives allow us to estimate the quality of these two corpora.

We also try answer another research question related to discourse connectives:

(Q. 2) Are discourse connectives easier/more difficult to disambiguate across languages?

(Q. 2) is not among our main research questions, however, it is important for our thesis because it motivates the bootstrapping expansion of our approach (we leave this project as feature work, see Chapter 7). More specifically, if some English discourse connectives are easier to be disambiguated than their French translation or vice versa, it would be possible to develop two classifiers for each language, then use these two classifiers to feed each other to improve their performance using parallel texts.

To answer **(Q. 1)**, we have developed a pipeline of two classifiers to disambiguate discourse connectives. This pipeline is a part of the CLaC discourse parser (Laali et al., 2015, 2016). The CLaC discourse parser is not only able to disambiguate discourse connectives, it also marks the two discourse arguments of discourse connectives and labels explicit and implicit discourse relations. The CLaC discourse parser ranked sixth out of 16 teams at the CoNLL 2015 shared-task (Xue et al., 2015) and sixth out of 14 teams at the CoNLL 2016 shared-task (Xue et al., 2016) on shallow discourse parsing. The parser is publicly available at `https://github.com/mjlaali/CLaCDiscourseParser`.

To answer **(Q. 2)**, we used the same pipeline but trained it for French discourse connectives. We refer to this parser as the *CLaC DC Disambiguator*. This work has been published in (Laali and Kosseim, 2016) and a pre-trained version of the parser is publicly available at `https://github.com/mjlaali/french-dc-disambiguation`. This classifier is used in Chapter 4 when we extrinsically evaluate the induced discourse annotated corpus for French.

3.1 Background

As mentioned before, discourse connectives can be ambiguous at two levels:

(1) they can be used in *discourse-usage* or *non-discourse-usage*, and

(2) they may be used to signal more than one discourse relation.

Discourse connectives are used in discourse-usage when they relate two abstract objects. For instance, (Ex. 20) to (Ex. 22) show examples of discourse-usage of *and*, *for example*, and *when*.

(Ex. 20) *Most balloonists seldom go higher than 2,000 feet* and **most average a leisurely 5-10 miles an hour.** (*EXPANSION:Conjunction*)

(Ex. 21) *Electronic gimmicks are key.* **Premark International Inc.,** for example**, peddles the M8.7sp Electronic Cycling Simulator, a $2,000 stationary cycle.** (*EXPANSION:Instantiation*)

(Ex. 22) *Most oil companies,* when **they set exploration and production budgets for this year,** *forecast revenue of $15 for each barrel of crude produced.* (*TEMPORAL:Synchronous*)[1]

However, these words/phrases do not always signal a discourse relation and may serve other functions such as to relate two non-abstract objects. This is the case, for example with the use of *and* in (Ex. 23) that connects two noun phrases, the use of *for example* in (Ex. 24) to modify a noun phrase or the use of *when* in (Ex. 25) to relativize extracted adjuncts.

(Ex. 23) Dr. Talcott led a team of researchers from the National Cancer Institute *and* the medical schools of Harvard University and Boston University.

(Ex. 24) These mainly involved such areas as materials – advanced soldering machines, *for example* – and medical developments derived from experimentation in space, such as artificial blood vessels.

(Ex. 25) Equitable of Iowa Cos., Des Moines, had been seeking a buyer for the 36-store Younkers chain since June, *when* it announced its intention to free up capital to expand its insurance business.[1]

Discourse connectives may also be ambiguous as they may signal different discourse relations. For example, *while* may signal a *TEMPORAL:Synchronous* as in (Ex. 26); a *COMPARISON:Contrast* as in (Ex. 27) or an *EXPANSION:Conjunction* as in (Ex. 28).

(Ex. 26) The league is the brainchild of Colorado real estate developer James Morley – once a minor-leaguer himself – who says *he had the idea last January* while **lying on a beach in Australia.** (*TEMPORAL:Synchronous*)

[1]All examples were taken from PDTB (Prasad et al., 2008a).

(Ex. 27) That's *because pollination,* while **easy in corn because the carrier is wind**, *is more complex and involves insects as carriers in crops such as cotton*. (*COMPARISON:Contrast*)

(Ex. 28) *In the past year, one inside director resigned,* while **three others retired.** (*EXPANSION: Conjunction*)[1]

Most previous work on the disambiguation of discourse connectives have focused on English discourse connectives (Marcu, 2000; Pitler and Nenkova, 2009; Lin et al., 2014). One of earliest and pioneer work on the disambiguation of discourse connectives, Pitler and Nenkova (2009), showed that four syntactic features (see Section 3.2 for details about the features) and the connective itself can disambiguate the usage of discourse connectives with an accuracy of 95.04% and the discourse relation signaled by discourse connectives with an accuracy of 94.15% at the first-level of the PDTB hierarchy (i.e. class – see Chapter 2 for more information about the PDTB hierarchy) within the PDTB corpus (Prasad et al., 2008a). Pitler and Nenkova (2009) used the gold-standard parse trees of the Penn Treebank (Marcus et al., 1993).

Later, Lin et al. (2014) used the context of the connective (i.e. the previous and the following word of the connective) and added seven lexico-syntactic features to the feature set proposed by Pitler and Nenkova (2009). In doing so, Lin et al. achieved an F1-score of 95.76% when using the gold-standard parse trees and 93.62% when using a syntactic parser for discourse-usage disambiguation of discourse connectives within the PDTB. Their system can also label the discourse relation signaled by discourse connectives with an F1-score of 80.61% on the second level of the PDTB hierarchy.

On the other hand, the disambiguation of discourse connectives in languages other than English has received much less attention. Due to syntactic differences across languages and different discourse annotation methodologies, the techniques developed for one language may or may not be as effective in another. For example, English discourse connectives include mostly subordinating conjunctions (e.g. *when*) or coordinating conjunctions (e.g. *but*). In addition, only a few connectives are disjoint (e.g. *On the one hand ... On the other hand*). This is not the case for Chinese which uses many more disjoint connectives (Zhou and Xue, 2012). Inspired by Pitler and Nenkova (2009), Alsaif and Markert (2011) proposed an approach for the disambiguation of Arabic Discourse connectives. Alsaif and Markert have shown that the features proposed by Pitler and Nenkova (2009) work well for Arabic with an accuracy of 91.2% to the usage of Arabic discourse connectives. Moreover, they further improved the result of their system by considering Arabic-specific morphological features and achieved an accuracy of 92.4%.

Today, due to the availability of discourse annotated corpora such as the French Discourse Treebank (FDTB; Danlos et al., 2015), it is possible to analyze how the features developed for English behave when applied to French.

3.2 Overview of the CLaC DC Disambiguator

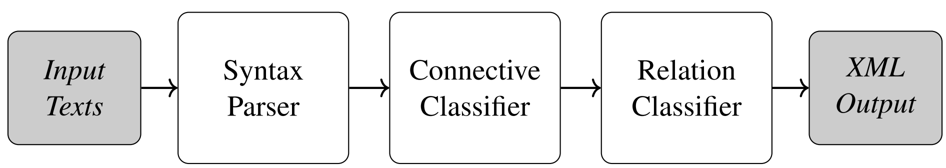


Figure 3.1: Pipeline for the disambiguation of discourse connectives.

We developed the *CLaC DC Disambiguator*, a pipeline for the disambiguation of discourse connectives, based on the UIMA framework (Ferrucci and Lally, 2004) and we used ClearTK (Bethard et al., 2014) to add machine learning functionality to the UIMA framework. Figure 3.1 shows the pipeline. Motivated by Lin et al. (2014), the *CLaC DC Disambiguator* consists of three components: the *Syntactic Parser*, the *Connective Classifier* and the *Relation Classifier*.

The *Syntax Parser* uses the Berkeley syntactic parser (Petrov and Klein, 2007) to add syntactic information (i.e. POS tags, constituent parse trees and dependency parses) to the input texts in the UIMA framework. It is also possible to configure this component so that it reads syntactic information from an external JSON file in the CoNLL 2015/2016 shared-task format (Xue et al., 2015, 2016).

Next, the *Connective Classifier* annotates discourse connectives within a text. Figure 3.2 shows the input and output of the *Connective Classifier* for (Ex. 29).

(Ex. 29) *We would stop index arbitrage* when **the market is under stress**. (*TEMPORAL:Synchronous*)[2]

Input: We would stop index arbitrage when the market is under stress.

Output:

```
<Document>
We would stop index arbitrage <DiscourseConnective>when</DiscourseConnective>
   the market is under stress.
</Document>
```

Figure 3.2: Example of input and output of the *Connective Classifier*.

Once discourse connectives have been classified as *discourse-usage*, the *Relation Classifier* labels the discourse relation signaled by the annotated discourse connectives. Figure 3.3 shows the input and the output of the *Relation Classifier* for (Ex. 29).

Section 3.3 and 3.4 will discuss the *Connective Classifier* and the *Relation Classifier* in detail.

3.3 Connective Classifier

3.3.1 Dataset Preparation

In order to build the *Connective Classifier* for English and French, we used the Penn Discourse Treebank (PDTB; Prasad et al., 2008a) and the French Discourse Treebank (FDTB; Danlos et al., 2015) for gold discourse annotations (see Chapter 2 for more information about these two corpora). To prepare these two corpora for our experiments, we used the annotated discourse connectives

[2]This example was taken from the PDTB.

Input:

```
<Document>
We would stop index arbitrage <DiscourseConnective>when</DiscourseConnective>
   the market is under stress.
</Document>
```

Output:

```
<Document>
We would stop index arbitrage <DiscourseConnective DiscourseRelation="TEMPORAL:
   Synchronous">when</DiscourseConnective> the market is under stress.
</Document>
```

Figure 3.3: The input and output of the *Relation Classifier*.

of these corpora as positive instances and all other occurrences of the connectives were used as negative instances. Table 3.1 shows the size of the datasets extracted from both the FDTB and the PDTB. As Table 3.1 shows, the dataset extracted from the FDTB is more biased toward negative examples than the dataset extracted from the PDTB. While the ratio of positive to negative examples is 0.38 (= 14K/37K) for the dataset extracted from the PDTB and this ratio is 0.25 (= 10K/40k) for the dataset extracted from the FDTB.

	Positive Examples	Negative Examples	# Words
PDTB	14K	37K	931K
FDTB	10K	40K	557K

Table 3.1: Statistics of the datasets extracted from the FDTB and the PDTB

Table 3.2 shows the distribution of the discourse connectives in both corpora along with their frequency. 63% (24% + 39%) of the French discourse connectives appear less than 10 times. This constitutes a large portion of French discourse connectives if we compare this number to its English counterpart in the PDTB (i.e. 18% = 3% + 15%). The more biased dataset for French entails that it will be more difficult to learn an accurate model for the disambiguation of French discourse connectives.

	PDTB (English)		**FDTB (French)**	
Frequency	**Number of DCs**	**%**	**Number of DCs**	**%**
$f = 1$	3	3%	55	24%
$1 < f < 10$	15	15%	89	39%
$f \geq 10$	82	82%	85	37%
Total	**100**	**100%**	**229**	**100%**

Table 3.2: Distribution of discourse connectives in the FDTB and the PDTB

3.3.2 Methodology

Algorithm 1 shows how we train the *Connective Classifier*. The algorithm takes four inputs. The first input is a list of discourse connectives: for English, we used the 100 discourse connectives listed in the Penn Discourse Treebank (Prasad et al., 2008a), and for French, we used the 371 discourse connectives listed in LEXCONN V2.1 (Danlos et al., 2015). The remaining inputs are to the algorithm are the input text, its gold annotations (see Section 3.3.1) and its syntactic tree generated by the Syntax Parser (see Section 3.2). Using these four inputs, Algorithm 1 trains a binary classifier to tag the *discourse-usage* of discourse connectives.

Algorithm 1: Train-Connective-Classifier

Input: dcs: a list of discourse connectives.

Input: $text$: the input texts.

Input: $syntaxTrees$: syntactic trees generated by the Syntax Parser.

Input: $annotations$: annotations of discourse connectives listed in dcs.

Output: $trainedClassifier$: the classifier that was trained using the datasets.

1 $instances = \{\}$;
2 **foreach** $dc \in dcs$ **do**
3   **foreach** $matched \in MatchesInText(dc, text)$ **do**
4         $features = GetFeatures(matched, text, syntaxTrees)$;
5         $\{features, GetLabel(matched, annotations)\} \rightarrow instances$;
6   **end**
7     $trainedClassifier \leftarrow Train(classifier, instances)$;
8 **end**

For each discourse connective, we first search the input texts for terms that match any expression in our list of discourse connectives (Line 2-3). Then, we compute 10 features for each match of the discourse connective (Line 5). These features, listed in Table 3.3, consist of the six features proposed by (Pitler et al., 2009) (#1 – #6 in Table 3.3) and four of the features proposed by (Lin et al., 2014) (#7 – #10 in Table 3.3). For example, given (Ex. 29) and its parse tree (shown in Figure 3.4), the value of these features are shown in the column labeled "Example" in Table 3.3.

Finally, we gather all these features and the label of the matched expression (either *discourse-usage* or *non-discourse-usage*) (Line 5) and use them to train a classifier (Line 7). For our experiments, we used the off-the-shelf implementation of the C4.5 decision tree classifier (Quinlan, 1993) available in WEKA (Hall et al., 2009) and trained a binary classifier to label discourse-usage and non-discourse usage of discourse connectives.

At inference time, we use Algorithm 2. Similarly to Algorithm 1, this algorithm also takes a

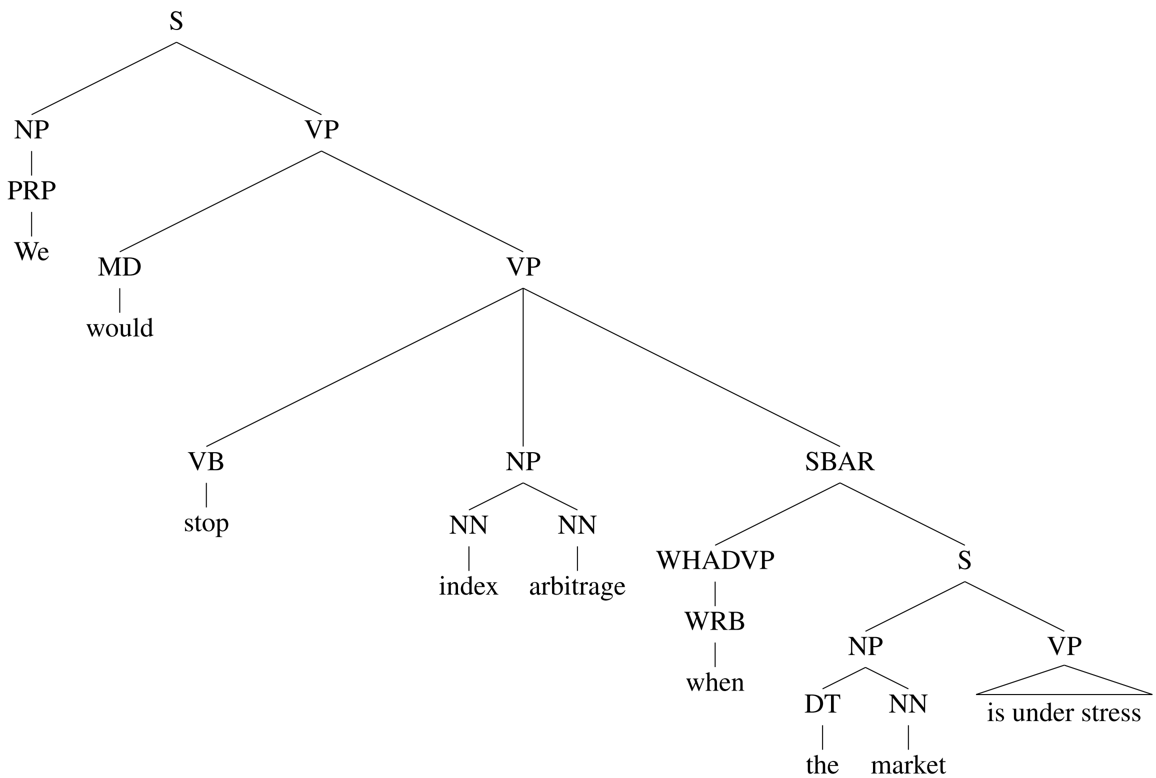


Figure 3.4: The parse tree for (Ex. 29) (available in the PDTB)

Description	Example
1. The discourse connective text in lowercase.	*when*
2. The categorization of the case of the connective: *all lowercase*, *all uppercase* and *initial uppercase*.	all lowercase
3. *SelfCat*: The highest node in the parse tree that covers the connective words but nothing more.	WHADVP
4. The parent of *SelfCat*	SBAR
5. The left sibling of *SelfCat*	null
6. The right sibling of *SelfCat*	S
7. The left word of the connective.	*arbitrage*
8. The POS of the left word of the connective.	NN
9. The right word of the connective.	*the*
10. The POS of the right word of the connective.	DT

Table 3.3: Features used for the disambiguation of discourse connectives.

list of discourse connectives and a text as inputs. Using the classifier trained using Algorithm 1, it generates labels of all matches of discourse connectives. Algorithm 2 is similar to Algorithm 1, however, after calculating the features, it feeds these features to the classifier to obtain the label of a discourse connective match (Line 5).

Algorithm 2: Label-Connectives

Input: dcs: a list of discourse connectives.

Input: $text$: the input texts.

Input: $classifier$: a trained classifier.

Output: $annotations$: the classifier that was trained in the datasets.

1 $annotations = \{\}$;

2 **foreach** $dc \in dcs$ **do**

3   **foreach** $matched \in MatchesInText(dc, text)$ **do**

4         $features = GetFeatures(matched, text, syntaxTrees)$;

5         $\{matched, Prediction(classifier, features)\} \rightarrow annotations$;

6   **end**

7 **end**

3.3.3 Evaluation

We evaluated the *Connective Classifier* in two settings: 1) *in-domain settings*: when the train dataset and the test dataset have the same domain, and 2) *out-of-domain settings:* when the test dataset has a different domain than the train dataset. These evaluations show how the *Connective Classifier* is robust to domain variation.

For *in-domain settings*, we report results using 10-fold cross-validation over the extracted datasets (see Table 3.1). For these experiments, we used Sections 2–21 of the PDTB and the FTB section of FDTB. We chose these sections because they share the same domain and therefore the classifiers are trained and tested on a homogeneous dataset. Moreover, Sections 2–21 of the PDTB have been recommended by both the PDTB manual and the CoNLL 2015/2016 shared-tasks for training.

Table 3.4 shows the overall performance of the classifier for the disambiguation of English and French discourse connectives. The results show that while the accuracies of the classifiers are similar for both English and French discourse connectives (94.6% and 94.4% respectively), the F1-score of the English classifier is higher than the F1-score of the French classifier (90.8% and 86.9% respectively). As Table 3.2 and Table 3.1 show, more French discourse connectives have a frequency higher than 10 and the French dataset is more biased towards non-discourse usage. These two characteristics are likely the reason for the lower F1-score for the French classifier.

Dataset	Precision	Recall	F1-score	Accuracy
Extracted from the PDTB (English)	87.0%	94.9%	90.8%	94.6%
Extracted from the FDTB (French)	86.1%	87.7%	86.9%	94.4%

Table 3.4: Overall performance of classifiers to disambiguate English and French discourse connectives.

For *out-of-domain settings*, we tested the classifiers on the CoNLL 2015/2016 blind test set (Xue, 2005) for the English classifier and the Sequoia section of the FDTB for the French classifier. The CoNLL 2015/2016 blind test set was extracted from Wikipedia and its domain significantly differ from the PDTB. Similarly, the text of the Sequoia section of the FDTB was extracted from Wikipedia and ANNODIS (Afantenos et al., 2012) which have different domain from the French Treebank. This evaluation can estimate the performance of the classifiers on texts with different domains.

Table 3.5 reports the performance of the classifiers with *out-of-domain settings*. As shown in Table 3.5, the F1-score of the English classifier slightly drops by 1.1% (=90.8% - 89.7%) which shows that it is robust when applied to texts with a different domain. It seems the French classifier is more sensitive to texts with a different domain as its F1-score drops by 8.5% (=86.9% - 78.4%). This can be explained by the low performance of the Berkeley parser or the smaller size of the FDTB (see Table 3.1).

Dataset	Precision	Recall	F1-score
CoNLL 2015/2016 Blind Test Set (English)	86.5%	89.7%	88.1%
Sequoia Section of the FDTB (French)	77.4%	79.4%	78.4%

Table 3.5: Performance of classifiers to disambiguate English and French discourse connectives when applied to texts with a different domain.

3.3.4 Cross-lingual Analysis of English and French Discourse Connectives

3.3.4.1 Entropy of French Discourse Connectives

To show the differences between English and French discourse connectives, we first compared the ambiguity of discourse connectives in the two languages by calculating the entropy of each discourse connective. Table 3.6 shows the top three most ambiguous and the top three least ambiguous discourse connectives (based on entropy) in the PDTB and the FDTB[3]. The full list of connectives with their entropy is available in Appendix B and Appendix C. As Table 3.6 shows, in English, ambiguous connectives which are used as often in a discourse/non-discourse context (yielding an entropy of 1.0) include *in contrast* and *as a results*, while in French, ambiguous connectives include the discourse connectives *effectivement* and *sinon*. On the other hand, in English, the non-ambiguous connectives (with entropy=0.0) include *on the other hand*, *particularly* and *upon*, while in French, they include *toutefios*, *à* and *à propos*.

Table 3.6 also shows the weighted average entropy of discourse connectives for each language. The entropy of French discourse connectives is 0.39 while the entropy of English discourse connectives is 0.51. This seems to indicate that the disambiguation of French discourse connectives can be considered a slightly easier task than the disambiguation of English discourse connectives.

[3]To achieve statistically reliable results, we did not consider discourse connectives that appeared less than 20 times.

PDTB (English)		
Discourse Connective	**Entropy**	**Freq.**
in contrast	1.00	22
besides	1.00	30
as a result	1.00	133
...	...	...
on the other hand	0.00	28
particularly	0.00	124
upon	0.00	40
Avg. Entropy	**0.51**	

(a) Entropy of English discourse connectives

FDTB (French)		
Discourse Connective	**Entropy**	**Freq.**
effectivement	1.00	27
sinon	1.00	27
d' une part	1.00	28
...	...	...
toutefois	0.00	135
à	0.00	9880
à propos	0.00	35
Avg. Entropy	**0.39**	

(b) Entropy of French discourse connectives

Table 3.6: Entropy of top three most/least ambiguous discourse connectives in the PDTB and the FDTB

To make a more detailed comparison, it would be preferable to align French and English discourse connectives with the same meaning and then compare the entropy of the mapped discourse connectives. Unfortunately, discourse connectives are language specific and cannot be easily aligned. To the best of our knowledge, a cross-lingual alignment of discourse connectives is available only for casual discourse connectives (Zufferey and Cartoni, 2012). Zufferey and Cartoni (2012) manually aligned a few hundred occurrences of *Causal* discourse connectives with their translation in the Europarl (Koehn, 2005) parallel texts. Then, they created an English-French dictionary for these discourse connectives based on the similarities and discrepancies between the discourse connectives and their most appropriate translation.

DC	English Translations	Entropy
because	*car, parce que*	0.98
since	*puisque, étant donné que, car*	0.80
as	*car, étant donné que, puisque, dans la mesure où*	0.59

(a) English discourse connectives

DC	French Translations	Entropy
parce que	*because*	0.55
puisque	*since, as, because*	0.25
car	*because, as, since, for*	0.05

(b) French discourse connectives

Table 3.7: Entropy of discourse connectives that signal a *Cause* relation in the FDTB and the PDTB

Table 3.7 shows the entropy of the French and English discourse connectives that signal the *Cause* relation identified by Zufferey and Cartoni (2012) and their most likely translations[4]. As

[4]Note that some translations of discourse connectives such as *étant donné que* are not considered discourse connectives in the FDTB and the PDTB because they do not satisfy the formal definition of discourse connectives. Therefore, we do not list their entropy in Table 3.7.

Table 3.7 shows, there does not seem to be a direct relationship between the entropy of the mapped discourse connectives. For example, while the French discourse connective *car* has an entropy of 0.05 (i.e. *car* is more than 99% of the time used in discourse-usage in the FDTB), its translations in English (i.e. *because*, *since*, and *as*) are very ambiguous.

The disparity between the entropy of discourse connectives in the FDTB and the PDTB can be explained by the differences between the languages. Regardless of its source, this disparity shows that for a specific discourse relations (e.g. the *Cause* discourse relation), annotating texts within a language (e.g. French) may be easier than in another language (e.g. English) because of the use of less ambiguous discourse connectives to signal these relations (e.g. *car* vs *because*). This disparity motivates discourse annotation projection (see Chapter 4).

3.3.4.2 Performance of the Classifier for Each Discourse Connective

The overall accuracy of the classifiers (see Table 3.4) shows that the effectiveness of the features is similar for both English and French. However, if we analyze the results for each connective, many seem to be very well classified with the features used; while a few are more difficult to disambiguate. In a further analysis, we compared the performance of classifier for each discourse connective for both languages. If we use as a baseline the assignment of the most likely class based only on the discourse connective text (the first feature in Table 3.3), many connectives obtained statistically significant improvements with all features. Table 3.8a and Table 3.8b show the accuracy of the classifiers for the English and French discourse connectives which achieved the greatest improvements over the baseline. All differences between the accuracies are statistically significant using Student t test with $P < .05$ and marked with $\Uparrow$. As Table 3.8a and Table 3.8b show, for these connectives, the classifier can disambiguate *discourse-usage* versus *non-discourse-usage* with a much better accuracy than the baseline. For example, the English classifier can disambiguate *as a result*, which is among the top tree ambiguous English discourse connectives, with an accuracy of 98.5%, showing a 45.1% improvement over the baseline classifier.

While the accuracy of the classifier is high for many discourse connectives, there are a few discourse connectives that the classifier cannot disambiguate. The five discourse connectives[5] that

[5]To achieve statistically reliable results, we did not consider discourse connectives that appeared less than 20 times.

Discourse Connective	Freq.	Entropy	Baseline	Accuracy	Diff.	
as a result	133	1.00	53.4%	98.5%	45.1%	⇑
instead	176	1.00	54.0%	98.3%	44.3%	⇑
besides	30	1.00	53.3%	93.3%	40.0%	⇑
because	1062	0.98	58.8%	95.1%	36.3%	⇑
until	302	0.98	57.6%	92.7%	35.1%	⇑

(a) English discourse connectives.

Discourse Connective	Freq.	Entropy	Baseline	Accuracy	Diff.	
si	502	0.77	22.5%	86.1%	63.5%	⇑
tant que	21	0.96	61.9%	100.0%	38.1%	⇑
en attendant	30	0.95	63.3%	100.0%	36.7%	⇑
aussi	533	0.97	59.3%	89.9%	30.6%	⇑
au lieu de	37	0.88	70.3%	100.0%	29.7%	⇑

(b) French discourse connectives.

Table 3.8: Accuracy of the classifiers for the English and French discourse connectives that achieved the greatest improvement over the baseline.

achieve the lowest accuracy are listed in Table 3.9a and Table 3.9b for English and French respectively. Again the differences between accuracies were evaluated with the Student t test, with $P < .05$ considered statistically significant and marked with $\Downarrow$ and lack of statistical increase is indicated by $\oslash$ in the table. Most of the discourse connectives in Table 3.9a and Table 3.9b have very high entropy. For some of these discourse connectives, we even see a drop in the accuracy of the classifier compared to the baseline. For example, the French classifier shows a drop of 37.5% for the discourse connective *simplement*. Typically, these discourse connectives have a low frequency and the classifier cannot learn a good model to disambiguate them.

3.4 Relation Classifier

In Section 3.3, we detailed the *Connective Classifier* (see Figure 3.1). In this section, we focus on the *Relation Classifier* (see Figure 3.1) that disambiguates the discourse relation signalled by discourse connectives.

For our experiments, we excluded French discourse connectives and only focused on the disambiguation of English discourse connectives. This is because, to date, there exists no large-scale

Discourse Connective	**Freq.**	**Entropy**	**Baseline**	**Accuracy**	**Diff.**	
though	288	0.94	63.9%	66.7%	02.8%	⊘
later	221	0.93	65.6%	66.5%	00.9%	⊘
ultimately	45	0.94	64.4%	64.4%	00.0%	⊘
finally	73	0.97	60.3%	60.3%	00.0%	⊘
in the end	20	0.99	40.0%	40.0%	00.0%	⊘

(a) English discourse connectives.

Discourse Connective	**Freq.**	**Entropy**	**Baseline**	**Accuracy**	**Diff.**	
par exemple	97	0.95	62.9%	62.9%	00.0%	⊘
simplement	32	0.00	100.0%	62.5%	-37.5%	⇓
maintenant	81	0.93	65.4%	58.0%	-07.4%	⊘
non plus	41	0.00	100.0%	56.1%	-43.9%	⇓
tout de même	21	0.99	57.1%	42.9%	-14.3%	⊘

(b) French discourse connectives.

Table 3.9: Accuracy of the classifier for discourse connectives with the least accuracy.

discourse annotated corpus for French where French discourse connectives are annotated with discourse relations[6]. Hence, we cannot train nor evaluate a French *Relation Classifier*.

3.4.1 Dataset Preparation

For our experiment, we used the dataset provided by the CoNLL 2014/2015 shared tasks (Xue et al., 2015, 2016). This dataset is based on the PDTB, however, a subset of PDTB discourse relations has been used in this dataset. This set of relations contains 14 relations that are primarily based on the second-level types of the PDTB (see Figure 2.3) and a selected number of third-level subtypes. This set of relations was created by the CoNLL orgonizers to collapse together very similar discourse relations that are hard to distinguish and thus difficult to annotate (such as *CONTINGENCY:Cause:reason* and *CONTINGENCY:Pragmatic cause*)(Xue et al., 2015). Table 3.10 shows the set of discourse relations specified by the CoNLL 2015/2016 shared-tasks with their correspondences to the PDTB discourse relations. For detailed information about this list see (Xue et al., 2015).

[6]Currently, only the discourse-usage of French discourse connectives is annotated in the FDTB and the discourse connectives have not been annotated with discourse relations.

	CoNLL Relation	PDTB Relation
1.	*TEMPORAL:Synchronous*	*same*
2.	*TEMPORAL:Asynchronous:precedence*	*same*
3.	*TEMPORAL:Asynchronous:succession*	*same*
4.	*CONTINGENCY:Cause:reason*	*CONTINGENCY:Cause:reason + CONTINGENCY:Pragmatic cause*
5.	*CONTINGENCY:Cause:result*	*same*
6.	*CONTINGENCY:Condition*	*CONTINGENCY:Condition + CONTINGENCY:Pragmatic condition + Subtypes of CONTINGENCY:Condition + Subtypes of CONTINGENCY:Pragmatic Condition*
7.	*COMPARISON:Contrast*	*COMPARISON:Contrast + COMPARISON:Pragmatic contrast + Subtypes of COMPARISON:Contrast*
8.	*COMPARISON:Concession*	*COMPARISON:Concession + COMPARISON:Pragmatic concession + Subtypes of COMPARISON:Concession*
9.	*EXPANSION:Conjunction*	*EXPANSION:Conjunction + EXPANSION:List*
10.	*EXPANSION:Instantiation*	*same*
11.	*EXPANSION:Restatement*	*EXPANSION:Restatement + Subtypes of EXPANSION:Restatement*
12.	*EXPANSION:Alternative*	*EXPANSION:Alternative:conjunctive + EXPANSION:Alternative:disjunctive*
13.	*EXPANSION:Alternative:chosen alternative*	*same*
14.	*EXPANSION:Exception*	*same*

Table 3.10: The 14 discourse relations specified in the CoNLL 2015/2016 shared-tasks with their correspondences to the PDTB discourse relations.

3.4.2 Methodology

The *Relation Classifier* uses the set of discourse relations specified by the CoNLL 2015/2016 shared-tasks (Xue et al., 2015, 2016). To label the discourse relation of each discourse connective, the *Relation Classifier* uses the same algorithms used for the *Connective Classifier* (i.e. Algorithm 1 and Algorithm 2). Therefore, we used the same 10 features in Table 3.3. As with the *Connective Classifier*, we used the off-the-shelf implementation of the C4.5 decision tree classifier (Quinlan, 1993) available in WEKA (Hall et al., 2009) for our experiments.

3.4.3 Evaluation

As with discourse-usage disambiguation, we first report results using 10-fold cross-validation on Sections 2–21 of the PDTB. The *Relation Classifier* identifies discourse relations signaled by discourse connectives with an accuracy of 81.0% within the PDTB. This is a high accuracy if we compare it with the annotator agreement reported for the PDTB as reported in Table 3.11 (Prasad et al., 2008a). As shown in Table 3.10, the list of the relations used in the CoNLL 2015/2016 shared-tasks are mostly chosen from the second-level types and some third-level subtypes of the PDTB relations. Therefore, we can compare the accuracy of the *Relation Classifier* (81.0%) with either the agreement at the type level (84%) or the agreement at the subtype level (80%).

CLASS	**Type**	**subtype**
94%	84%	80%

Table 3.11: Inter-annotator agreement reported for the PDTB.

If we break down the overall performance of the *Relation Classifier* for each discourse relation, we see that while the classifier can reliably identify most of the discourse relations such as *EXPANSION:Instantiation* with an F1-score above 90%, our features are not as effective for a few discourse relations. Table 3.13 shows the precision, recall and F1-score of the classifier for each discourse relations using 10-fold cross-validation. The top three discourse relations with lowest F1-score are *COMPARISON:Concession*, *EXPANSION:Restatement* and *TEMPORAL:Synchronous*. To understand relations that are confused with these three relations, we computed the confusion matrix.

As shown in Table 3.14, most errors come from *COMPARISON:Concession (R1)* that are miss-labeled as *COMPARISON:Contrast (R2)*. This accounts for 822 classifications out of 1093, for a total of 75.2%. These two relations are semantically very close and are very hard to distinguish even for human annotators (Zufferey and Degand, 2014). *EXPANSION:Restatement (R11)* also shows a high level of confusion (see Table 3.14). There are very few instances of this relation in the PDTB (126 in total) and it seems that the classifier could not learn a proper model to identify this relation. Finally, *TEMPORAL:Synchronous (R14)* relation are mostly confused for *CONTINGENCY:Cause:reason (R3)*. This is mainly because of the connective *when* which can signal both *TEMPORAL:Synchronous* and *CONTINGENCY:Cause:reason* at the same time. Table 3.12 shows all discourse relations signalled by *when* with a frequency $\geq$ 10 in the PDTB. According to the PDTB, as shown in Table 3.12, most of time when the connective *when* signals *CONTINGENCY:Cause:reason*, the connective also signals another discourse relation. For example, 65 occurrences of *when* in the PDTB signals both *CONTINGENCY:Cause:reason* and *TEMPORAL:Synchronous* at the same time. Since the *Relation Classifier* cannot output multiple discourse relations, it tends to not label *when* with *CONTINGENCY:Cause:reason* and labels *when* with its most likely relation (i.e. *TEMPORAL:Synchronous*).

Relation	**Frequency**
TEMPORAL:Synchronous	477
TEMPORAL:Asynchronous:succession	157
CONTINGENCY:Condition	124
CONTINGENCY:Cause:reason and *TEMPORAL:Asynchronous:succession*	65
CONTINGENCY:Condition and *TEMPORAL:Synchronous*	50
CONTINGENCY:Cause:reason and *TEMPORAL:Synchronous*	39
CONTINGENCY:Condition and *TEMPORAL:Synchronous*	10

Table 3.12: All discourse relations signalled by *when* with a frequency $\geq$ 10.

To estimate the performance of the *Relation Classifier* on texts with different domains, we trained the classifier on Sections 2–21 of the PDTB and tested it on the CoNLL 2015/2016 blind test set (Xue, 2005) which is extracted from Wikipedia. Table 3.15 shows the precision, recall and F1-score of the *Relation Classifier* with and without error propagation from the *Connective Classifier*. As Table 3.15 shows, the F1-score of drops from 79.7% (see Table 3.13) to 74.3% when tested on the CoNLL 2015/2016 blind test set. The F1-score drops further to 63.0% when the errors

Discourse Relation	Precision	Recall	F1-score
COMPARISON:Concession	59.3%	16.1%	25.3%
COMPARISON:Contrast	73.2%	93.2%	82.0%
CONTINGENCY:Cause:reason	91.5%	66.2%	76.8%
CONTINGENCY:Cause:result	99.1%	71.0%	82.8%
CONTINGENCY:Condition	93.8%	79.4%	86.0%
EXPANSION:Alternative	94.1%	87.9%	90.9%
EXPANSION:Alternative:chosen alternative	90.1%	91.9%	91.0%
EXPANSION:Conjunction	90.9%	93.4%	92.2%
EXPANSION:Exception	88.9%	61.5%	72.7%
EXPANSION:Instantiation	99.1%	96.2%	97.6%
EXPANSION:Restatement	62.7%	41.3%	49.8%
TEMPORAL:Asynchronous:precedence	89.0%	91.9%	90.4%
TEMPORAL:Asynchronous:succession	87.5%	63.5%	73.6%
TEMPORAL:Synchronous	54.6%	84.9%	66.5%
Weighted Avg:	**82.1%**	**81.0%**	**79.7%**

Table 3.13: Precision, recall, and F1-score of the *Relation Classifier* for each discourse relation using 10-fold cross-validation on Sections 2–21 of the PDTB.

from the *Connective Classifier* are propagated. While the overall F1-score of the *Relation Classifier* is not high when errors are propagated, many discourse connectives are still reliably disambiguated. Table 3.16 shows 18 discourse connectives with an F1-score higher than 80.0%.

3.5 Conclusion

In this chapter, we have described our pipeline to disambiguate discourse connectives. The pipeline consists of two main components: 1) the *Connective Classifier* and 2) the *Relation Classifier*. For these two classifiers, we used the same set of 10 features.

Our experiments on the French Discourse Treebank (FDTB) and the Penn Discourse Treebank (PDTB) show that overall the *Connective Classifier* can effectively disambiguate English and French discourse connectives between *discourse-usage* and *non-discourse-usage* with an F1-score of 90.8% for English and 86.9% for French. The fact that the same features proposed for English can be used almost as effectively for French and Arabic (Alsaif and Markert, 2011) suggests that lexicalized discourse connectives share certain common structural features cross-linguistically and that these structures are potentially an important component in discourse processing. However, our

		Classified Relation													
	True Relation	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	R_9	R_{10}	R_{11}	R_{12}	R_{13}	R_{14}
R_1	***COMPARISON:Concession***	176	**862**	0	0	35	0	1	7	0	0	1	0	1	10
R_2	***COMPARISON:Contrast***	105	3144	0	1	7	1	3	33	0	0	5	4	4	68
R_3	***CONTINGENCY:Cause:reason***	0	1	773	1	1	0	0	13	0	0	0	3	54	322
R_4	***CONTINGENCY:Cause:result***	0	1	1	427	1	0	0	145	0	0	1	19	1	5
R_5	***CONTINGENCY:Condition***	5	2	3	0	951	2	0	40	0	0	1	18	6	170
R_6	***EXPANSION:Alternative***	0	0	0	0	1	174	4	14	1	0	0	3	0	1
R_7	***EXPANSION:Alternative:chosen alternative***	1	1	0	0	0	0	91	0	0	0	5	0	0	1
R_8	***EXPANSION:Conjunction***	1	141	0	0	0	4	0	4149	0	0	10	19	1	115
R_9	***EXPANSION:Exception***	1	2	0	0	0	2	0	0	8	0	0	0	0	0
R_{10}	***EXPANSION:Instantiation***	0	0	0	0	0	0	0	2	0	227	7	0	0	0
R_{11}	***EXPANSION:Restatement***	0	5	0	0	3	2	1	**56**	0	2	**52**	3	0	2
R_{12}	***TEMPORAL:Asynchronous:precedence***	0	7	1	2	0	0	0	48	0	0	0	736	3	4
R_{13}	***TEMPORAL:Asynchronous:succession***	1	0	56	0	1	0	0	43	0	0	0	14	554	204
R_{14}	***TEMPORAL:Synchrony***	7	**128**	11	0	14	0	1	14	0	0	1	8	9	1086

Table 3.14: Confusion matrix for the *Relation Classifier*.

	Precision	Recall	F1-score
Without error propagation	72.7%	76.1%	74.3%
With error propagation	61.9%	64.2%	63.0%

Table 3.15: Precision, recall, and F1-score of the *Relation Classifier* when trained on Sections 2–21 of the PDTB and tested on the CoNLL 2015/2016 blind test set.

	Discourse Connective	Precision	Recall	F1-score
1.	*in addition*	100.0%	100.0%	100.0%
2.	*for example*	100.0%	100.0%	100.0%
3.	*furthermore*	100.0%	100.0%	100.0%
4.	*so that*	100.0%	100.0%	100.0%
5.	*additionally*	100.0%	100.0%	100.0%
6.	*afterwards*	100.0%	100.0%	100.0%
7.	*by then*	100.0%	100.0%	100.0%
8.	*in short*	100.0%	100.0%	100.0%
9.	*moreover*	100.0%	100.0%	100.0%
10.	*on the other hand*	100.0%	100.0%	100.0%
11.	*therefore*	100.0%	100.0%	100.0%
12.	*also*	88.1%	96.1%	91.9%
13.	*because*	82.4%	100.0%	90.3%
14.	*so*	87.5%	87.5%	87.5%
15.	*then*	87.5%	87.5%	87.5%
16.	*before*	76.2%	94.1%	84.2%
17.	*or*	71.4%	100.0%	83.3%
18.	*until*	80.0%	80.0%	80.0%

Table 3.16: Discourse connectives with an F1-score higher than or equal to 80.0%.

analysis also shows that the features are not as effective for all connectives. Some high entropy connectives such as *as a result* have a very high accuracy whereas others such as *finally* or *in the end* require additional features.

Our experiments on the PDTB show that the *Relation Classifier* can identify the discourse relation signaled by English discourse connectives with near-human performance. However, as with the *Connective Classifier*, our analysis shows that the features are not as effective for all discourse relations. While the performance of the *Relation Classifier* are high for most discourse relations such as *EXPANSION:Instantiation*, other discourse relations such as *COMPARISON:Concession* need additional features to disambiguate.

To estimate the performance of our pipeline on texts with different domain, we evaluated it on the CoNLL 2015/2016 blind test set. Our experiments show that the *Connective Classifier* is robust as its F1-score slightly drops from 90.8% to 88.1%. We also showed that even if the *Relation Classifier* performance drops from 79.7% to 74.3% on the CoNLL 2015/2016 blind test set, many discourse connectives such as *also* whose discourse relations can be efficiently disambiguated on texts with a different domain.

Finally, our comparison between English and French discourse connectives show that some discourse connectives are easier to be disambiguated in French than English. As discussed in at the beginning of this chapter, this motivates a bootstrapping expansion of our approach (see Chapter 7).

In next chapter, we use our pipeline developed in this chapter to annotate English discourse connectives within parallel texts and then project these annotations from English texts onto French texts.

Chapter 4

Discourse Annotation Project

Annotation projection is a promising approach to quickly build initial discourse treebanks using parallel texts. In this chapter, we develop a method to project discourse annotations of English discourse connectives onto French discourse connectives. To annotate English discourse connectives, we used the *CLaC DC Disambiguator* presented in the previous chapter. Figure 4.1 shows the input and output of our method where the English discourse connective *since* was automatically labeled by the *CLaC DC Disambiguator*.

In this chapter, we try to address research questions **(Q. 2)** (see Section 1.2):

(Q. 2) **How can annotations of discourse connectives be automatically projected withing parallel texts in order to induce PDTB-style discourse annotated corpora?**

To answer **(Q. 2)**, we have developed a novel approach based on the intersection between statistical word-alignment models to align occurrences of French discourse connectives to their English translation. Then, we used these alignments to project annotations from English texts onto French texts. We experimented with different statistical word-alignment models and induced the Europarl ConcoDisco corpora where English and French discourse connectives are aligned to each other. The Europarl ConcoDisco-Intersection corpus, which contains the most accurate alignments, is publicly available at `https://github.com/mjlaali/Europarl-ConcoDisco`. Moreover, from the French side of the Europarl ConcoDisco corpora, we created the first PDTB-style discourse annotated corpus for French, which we refer to as the FrConcoDisco corpora.

CONTINGENCY.Cause.reason
EN: I would ask that they reconsider, $\boxed{\text{since}}$ this is not the case.
FR: Je demande que cette décision soit reconsidérée car ce n'est pas le cas.

(a) The input of discourse annotation projection.

CONTINGENCY.Cause.reason
EN: I would ask that they reconsider, $\boxed{\text{since}}$ this is not the case.

Annotation Projection

FR: Je demande que cette décision soit reconsidérée $\boxed{\text{car}}$ ce n'est pas le cas.
CONTINGENCY.Cause.reason

(b) The output of discourse annotation projection.

Figure 4.1: Example of the projection of discourse annotations from English to French texts within parallel texts.

To evaluate the FrConcoDisco corpora, we have used both an intrinsic and an extrinsic evaluation. Our intrinsic evaluation shows that our approach can project discourse annotations with a precision of 0.914. For the extrinsic evaluation, we used the FrConcoDisco corpora to train a classifier to identify the discourse-usage of French discourse connectives. This classifier can identify the discourse-usage of French discourse connectives with an F1-score of 0.546, which is 15% better than the F1-score of the classifier trained on the non-filtered annotations. This work has been published in (Laali and Kosseim, 2017b).

4.1 Introduction

Annotation projection has been widely used in the past to build natural language applications and resources (Yarowsky et al., 2001; Bentivogli and Pianta, 2005; Tiedemann, 2015; Versley, 2010; Laali and Kosseim, 2014; Hidey and McKeown, 2016) (see Section 2.2.1 for related work). Annotation projection exploits parallel sentences and projects annotations from a source language to a target language. By parallel sentences, we mean two sentences that are a translation of each other in two different languages. The main assumption of annotation projection is that because parallel sentences are a translation of each other, semantic and rhetorical annotations should, in principle, transfer from the source language to the target language (Versley, 2010; Laali and Kosseim, 2014;

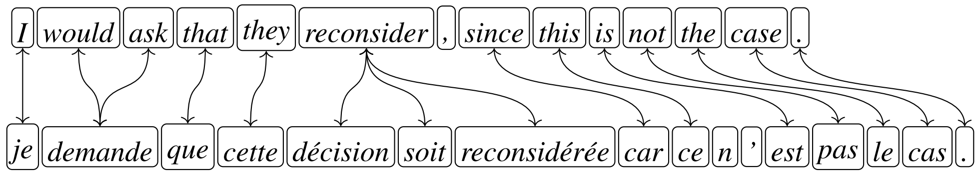


Figure 4.2: Example of the alignment between English and French words generated from a statistical word-alignment model.

Hidey and McKeown, 2016). Hence, these annotations can be projected from one side onto the other side of parallel sentences.

In this chapter, we will project explicit discourse relations within parallel texts. As discourse relations are semantic and rhetorical in nature, they are an attractive target for annotation projection.

Typically annotation projection relies on statistical word-alignment models (Tiedemann, 2015; Versley, 2010; Laali and Kosseim, 2014; Hidey and McKeown, 2016). Essentially, statistical word-alignment models are unsupervised models that map words to their most likely translation in parallel sentences (Brown et al., 1993). Figure 4.2 shows an example of word-alignments generated from a statistical word-alignment model. For example, in Figure 4.2, the English discourse connective *since* has been aligned to its best translation *car* in French. Based on this alignment, the annotation of the English discourse connective *since* (i.e. *CONTINGENCY.Cause.reason*) can be projected onto the French discourse connective *car* as shown in Figure 4.1.

As we show in this chapter, a naive approach for aligning English and French discourse connectives is not accurate enough to build discourse annotated corpora and may generate unsupported discourse annotations. This is because statistical word-alignment models tend to generate noisy alignments when discourse connectives are not reproduced in the target language, or in other words, when discourse relations are changed from explicit relations to implicit ones during the translation process. Moreover, because no counterpart translation exists for these discourse connectives, it is difficult to reliably annotate them and any induced annotation would be unsupported. (Ex. 30) shows parallel sentences where the French discourse connective *mais*[1] has been dropped in the English translation, hence the discourse relation *COMPARISON:Concession* is changed from an explicit relation in French to an implicit one in English.

[1]Free translation: *but*

(Ex. 30) **FR:** *Comme tout le monde dans cette Assemblée, j'aspire à cet espace de liberté, de justice et de sécurité, <u>mais</u> je ne veux pas qu'il débouche sur une centralisation à outrance, le chaos et la confusion.*

EN: *Like everybody in this House, I want freedom, justice and security. I do not want to see these degenerate into over-centralisation, chaos and confusion.*

Note that, as many previous work have done (Prasad et al., 2010; Versley, 2010; Meyer, 2011; Popescu-Belis et al., 2012; Cartoni et al., 2013; Laali and Kosseim, 2014; Hidey and McKeown, 2016), we still assume that discourse relations are preserved during the translation process. However in contrast to them, we do not assume that the realization of discourse relations is the same in the source and target languages and the relations may change from explicit relations to implicit ones or vice-versa.

Changing the realization of discourse relations during the translation process is a known phenomenon in the Machine Translation community (Cartoni and Meyer, 2012; Popescu-Belis et al., 2012; Meyer and Webber, 2013) and in discourse studies (Zufferey and Cartoni, 2012; Taboada and de los Ángeles Gómez-González, 2012; Zufferey and Degand, 2014; Zufferey and Gygax, 2015; Hoek and Zufferey, 2015; Zufferey, 2016) (see Section 2.2 for a more detailed discussion). For example, according to (Meyer and Webber, 2013), up to 18% of explicit discourse relations are changed to implicit ones in the English/French portion of the newstest2010+2012 dataset (Callison-Burch et al., 2010, 2012).

In this chapter, we also propose an approach to identify dropped discourse connectives during the translation in order to identify noisy word-alignments and unsupported annotations. In previous work, to extract dropped discourse connectives, scholars either manually annotated parallel sentences (Zufferey and Cartoni, 2012; Zufferey and Gygax, 2015; Zufferey, 2016) or used a heuristic-based approach using a dictionary (Meyer and Webber, 2013; Cartoni et al., 2013) to verify the translation of discourse connectives proposed by statistical word alignment models such as IBM models (Brown et al., 1993). In contrast to previous works, our approach automatically identifies dropped discourse connectives by intersecting statistical word-alignments without using any additional resources such as a dictionary.

As a by-product of our approach for annotation projection, we generated a PDTB-style discourse annotated corpus for French which we refer to as FrConcoDisco-*Intersection*. As discussed in Chapter 2, there currently exist two publicly available discourse annotated corpora for French:

(1) *The French Discourse Treebank (FDTB)* (Danlos et al., 2015): This corpus contains more than 10,000 instances of LEXCONN's French discourse connectives annotated as *discourse-usage*. However, to date, these French discourse connectives have not been annotated with discourse relations.

(2) *ANNODIS* (Afantenos et al., 2012): This corpus includes annotations of discourse relations, however, the size of the corpus is small and only contains 3355 relations. While this corpus uses SDRT, we use the PDTB-style annotations in the FrConcoDisco-*Intersection* corpus.

In the rest of this chapter, we explain our approach in detail. Section 4.2 explains our methodology to build the Europarl ConcoDisco and FrConcoDisco-*Intersection* and then Section 4.3 presents our approach to evaluate the FrConcoDisco-*Intersection* corpus. Finally Section 4.4 concludes our findings.

4.2 Methodology

4.2.1 Dataset Preparation

For our experiment, we have used the English-French part of the Europarl parallel corpus (Koehn, 2005) which contains around two million parallel sentences and around 50 millions words in each side. To prepare this dataset for our experiment, we used the *CLaC DC Disambiguator* presented in Chapter 3 to identify English discourse connectives and the discourse relation that they signal. Recall that the *CLaC DC Disambiguator* has been learned on Section 02-20 of the PDTB and can disambiguate the usage of the 100 English discourse connectives listed in the PDTB with an F1-score of 88.1% and label them with their PDTB relation with an F1-score of 74.3% when tested on the blind test set of the CoNLL 2016 shared task (Xue et al., 2016).

The *CLaC DC Disambiguator* was used because its performance is very close to that of the state of the art system (Oepen et al., 2016) (i.e. 91% and 77% respectively), but is more efficient at running time than (Oepen et al., 2016). Note that since the CoNLL 2016 blind test set was extracted from Wikipedia and its domain and genre differ significantly from the PDTB, the 88.1% and 74.3% F1-scores of the *CLaC DC Disambiguator* can be considered as an estimation of its performance on texts with a different domain/genre such as Europarl.

In addition to disambiguate English discourse connectives, we used the Moses statistical machine translation system (Koehn et al., 2007) to align English and French words. As a part of its translation model, Moses can use a variety of statistical word-alignment models. For example, Figure 4.3 shows word-alignments for the French discourse connective *d'autre part* where the alignment model found a 1:2 alignment between *d'* and *on the* then three 1:1 alignments. In this case, the English translation of *d'autre part* will be considered to be *on the other hand*.

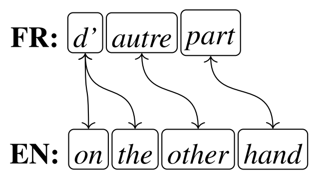


Figure 4.3: Word-alignments for the French discourse connective *d'autre part*.

Previous works on annotation projection only experimented with the *Grow-diag* model Och and Ney (2003) (see (Versley, 2010; Tiedemann, 2015) for example). However, in this work we experimented with different models to identify their effect on the annotation projection task. For our experiment, we trained an IBM 4 word-alignment model (Brown et al., 1993) in both directions and generated two word-alignments:

(1) *Direct* word-alignment which includes word-alignments when the source language is set to French and the target language is set to English.

(2) *Inverse* word-alignment which is learned in the reverse direction of *Direct* word-alignment (i.e. the source language is English and the target language is French).

In addition to these two word-alignments, we also experimented with:

(3) *Intersection* word-alignment which contains alignments that appear in both the *Direct* word-alignment and in the *Inverse* word-alignment. This creates less, but more accurate alignments.

(4) *Grow-diag* word-alignment which expands the *Intersection* word-alignment with the alignments that lie in the union of the *Direct* word-alignment and the *Inverse* word-alignment and that satisfy the heuristic proposed by Och and Ney (2003). This heuristic creates more, but less supported alignments.

4.2.2 Discourse Annotation Projection

Algorithm 3 shows how we project discourse relations from the English side onto the French side. The inputs to our algorithm is a pair of parallel sentences ($sent_{en}, sent_{fr}$) along with its word-alignments ($alignments$), and the annotations of the English discourse connectives ($annotations_{en}$) within the parallel sentences that have been prepared in Section 4.2.1. Moreover, the algorithm needs as input a list of French discourse connectives. For this, we used the list of 371 French discourse connectives in LEXCONN (Roze et al., 2012).

As Algorithm 3 shows, we first identified all occurrences of the 371 French discourse connectives listed in LEXCONN (Roze et al., 2012), in the French side of the parallel texts and marked them as French candidate discourse connectives (Lines 2-3). Then, we automatically identify the translation of these French candidate discourse connectives by concatenating all the English words that were aligned with each word of the French candidate discourse connectives (Line 4). If a French candidate discourse connective has been translated into English in the parallel sentence and has been aligned to English texts (Line 5), we consider it as a supported candidate and label it according to the annotation of its English translation identified by the word alignments (Lines 6-12) as follows:

(1) *Discourse-Usage (or NDU)*: If the English translation was part of a PDTB English discourse connective and was marked by the *CLaC DC Disambiguator* then we project the English annotations and assume that the French candidate discourse connective signals the same relation as the English discourse connective (Line 8).

(2) *Non-Discourse-Usage (or NDU)*: If the English translation was not part of a PDTB English discourse connective or was not marked by the *CLaC DC Disambiguator*, then we project the English NDU label and assume that the French candidate discourse connective is not used in a discourse usage and label it as *NDU* (Line 10).

Algorithm 3: Project-Discourse-Annotation

Input: $(sent_{en}, sent_{fr})$: a pair of parallel sentences.

Input: $alignments$: alignments between English and French words in $(sent_{en}, sent_{fr})$.

Input: $annotations_{en}$: annotations of English discourse connectives in $sent_{en}$.

Input: DC_{fr}: a list of French discourse connectives.

Output: $annotations_{fr}$: annotations of French discourse connectives in $sent_{fr}$.

$annotations_{fr} = \{\}$;
foreach $dc \in DC_{fr}$ **do**
    foreach $candidate \in Occurences(dc, sent_{fr})$ **do**
        $trans = GetTranslation(candidate, sent_{en}, alignments)$;
        if $trans \neq nil$ **then**
            $relation = GetAnnotation(trans, annotations_{en})$;
            if $relation \neq nil$ **then**
                $label = (DU, relation)$;
            else
                $label = NDU$;
            end
            $CreateAnnotation(candidate, label) \rightarrow annotations_{fr}$;
        end
    end
end

Our algorithm excludes any candidate that has not been translated. More specifically, if the word-alignments contain no alignments for a French candidate discourse connective, then we assume that the candidate has no translation and there is no annotation to be projected. We refer to such French candidate discourse connectives as unsupported candidates and filter them before the annotation projection.

Table 4.1 shows examples of the input and output of our algorithm for four parallel texts. In (Ex. 31), *aussi* is translated to *also* which the *CLaC DC Disambiguator* tagged as a discourse

	Input		Output
#	**French**	**English**	**Projected Annotation**
(Ex. 31)	*Les États membres ont* ***aussi*** *leur part de responsabilité dans ce domaine et ils ne doivent pas l'oublier.*	*The Member States must* ***also/DU/CONJUNCTION*** *bear in mind their responsibility.*	DU/CONJUNCTION ⇒ included in corpus
(Ex. 32)	*Et quand je parle d'utilisation optimale, j'évoque* ***aussi*** *bien le niveau national que le niveau régional.*	*When I speak of optimum utilisation, I am referring* ***both/NDU*** *to the national and regional levels.*	NDU ⇒ included in corpus
(Ex. 33)	*Pour conclure, je dirai que nous devons faire en sorte que les lignes directrices soient larges, indicatives et souples,* ***afin d'****aider nos gestionnaires de programmes et les utilisateurs des crédits et de valoriser au mieux les potentialités de nos nouveaux domaines de régénération.*	*The conclusion is that we must make the case for guidelines to be broad, indicative and flexible to assist our programme managers and fund-users and to get the maximum potential out of our new fields of regeneration.*	None ⇒ not included in corpus
(Ex. 34)	*Vous me direz que la croissance ou la pénurie, ce n'est pas* ***pour*** *tout le monde.*	*You will tell me that situations of growth or shortage do not affect everyone alike.*	None ⇒ not included in corpus

Table 4.1: Examples of discourse connective annotation projection in parallel sentences. French candidate discourse connectives and their correct English translation are in bold face[4].

connective signaling a *EXPANSION:Conjunction* relation. By projecting this annotation, we induce that *aussi* should also be used in discourse usage and signals a *EXPANSION:Conjunction* relation. On the other hand, in (Ex. 32), *aussi* is translated to *both* which is not recognized as a discourse connective, therefore, this French candidate discourse connective is assumed to be used in a NDU.

(Ex. 33) and (Ex. 34) in Table 4.1 illustrate two cases of unsupported French candidate discourse connectives. In (Ex. 33), the explicit French discourse connective *afin d'*[2] signals a *CONTINGENCY:Cause:reason* relation, however it has been dropped in the English translation and replaced by the use of *to + infinitive* (*to assist*) to implicitly convey the *CONTINGENCY:Cause:reason* relation. This example shows how the realization of discourse relations may be changed from explicit to implicit during the translation process. In (Ex. 34), the French candidate discourse connective *pour*[3] does not signal a discourse relation but again, it has no English translation. In both examples, since there is no English translation of the French candidate discourse connectives, they will be filtered because there is no annotation that can be reliably projected onto them.

Our approach is different from previous work as we identify unsupported French candidate discourse connectives before the projection and filter them out. For example, Versley (2010) assumed

[2]Free translation: *in order to*
[3]Free translation: *for*

that French candidate discourse connectives are used in either a NDU or a NDU. Anytime there is not enough evidence to label a French candidate discourse connective as a NDU (e.g. its translation is not part of an English discourse connective), the candidate is assumed to be a NDU. This means that in (Ex. 32), (Ex. 33) and (Ex. 34), all French candidate discourse connectives would be tagged as NDU in Versley (2010)'s approach. On the other hand, our approach only labels the French candidate discourse connective in (Ex. 32) as NDU and filters out the French candidate discourse connectives in (Ex. 33) and (Ex. 34) as they cannot be reliably annotated.

4.2.3 Building the Europarl ConcoDico Corpora and FrConcoDisco Corpora

Automatically aligning French candidate discourse connectives to their English counterparts allowed us to automatically project discourse annotations from English onto French for each of the four word-alignment models. As a result, we created four different corpora from Europarl where French candidate discourse connectives are aligned to their English translation and are labeled with either NDU and the discourse relation that they signal or NDU. We called these corpora: the *Europarl ConcoDisco* corpora. For comparative purposes, we also extracted a corpus without filtering unsupported candidates, which we refer to as Europarl ConcoDisco-*Naive-Grow-diag*. In total, we threfore generated: 1) Europarl ConcoDisco-Intersection, 2) Europarl ConcoDisco-Grow-diag, 3) Europarl ConcoDisco-Direct, 4) Europarl ConcoDisco-Inverse and 5) Europarl ConcoDisco-Naive-Grow-diag.

Figure 4.4 shows a sample of the Europarl ConcoDisco-*Intersection* corpus. Each pair of parallel sentences contains annotations of English discourse connectives (automatically marked by the *CLaC DC Disambiguator*) and annotations of French candidate discourse connectives (as a result of annotation projection) encapsulated in *DiscourseConnective* XML elements. For French candidate discourse connectives, if *DiscourseConnective* elements does not indicate a sense, it means that the French candidate discourse connective is not used in a discourse usage (i.e. it was aligned to an English text that does not signal a discourse relation).

Since our focus is to build a PDTB-style discourse annotated corpus, for the rest of this chapter, we only focus on the French side of the Europarl ConcoDisco corpora, which we refer to as the

[4]All examples are extracted from the Europarl parallel corpus.

FrConcoDisco corpora. Table 4.2 shows statistics of the five FrConcoDisco corpora that we generated. As the table shows, all corpora contain about 1 million French candidate discourse connectives that are labelled as true French discourse connective and for which a PDTB discourse relation is assigned, and around 5 million candidates in non-discourse-usage. Compared to the FDTB, these corpora are approximately 100 times larger and French discourse connectives are associated with PDTB relations.


```xml
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="translator.xsl"?>

<DOCUMENT>
  <ParallelChunk annotation_id="0" docOffset="0">
    <en>Resumption of the session</en>
    <fr>Reprise de la session</fr>
  </ParallelChunk>
<Speaker annotation_id="26" id="1" name="President">
    <ParallelChunk annotation_id="26" docOffset="1">
      <en>I declare resumed the session of the European Parliament adjourned on Friday 17
          December 1999, <Alignment alignment="132" annotation_id="121">
          <DiscourseConnective annotation_id="121" sense="Expansion.Conjunction">and</
              DiscourseConnective>
        </Alignment> I would like once again to wish you a happy new year in the hope that
             you enjoyed a pleasant festive period.</en>
      <fr>Je déclare reprise la session du Parlement européen qui avait été interrompue le
           vendredi 17 décembre dernier <Alignment alignment="121" annotation_id="132">
          <DiscourseConnective annotation_id="132" sense="Expansion.Conjunction">et</
              DiscourseConnective>
        </Alignment> je vous renouvelle tous mes vux <DiscourseConnective annotation_id="
            167">en</DiscourseConnective> espérant <DiscourseConnective annotation_id="179
            ">que</DiscourseConnective> vous avez passé de bonnes vacances.</fr>
    </ParallelChunk>
<ParallelChunk annotation_id="234" docOffset="2">
      <en>
        <DiscourseConnective annotation_id="234" sense="Comparison.Concession">Although</
            DiscourseConnective>, <Alignment alignment="219" annotation_id="244">
          <DiscourseConnective annotation_id="244" sense="Temporal.Synchrony">as</
              DiscourseConnective>
        </Alignment> you will have seen, the dreaded 'millennium bug' failed to
            materialise, still the people in a number of countries suffered a series of
            natural disasters that truly were dreadful.</en>
      <fr>
        <Alignment alignment="244" annotation_id="219">
          <DiscourseConnective annotation_id="219" sense="Temporal.Synchrony">Comme</
              DiscourseConnective>
        </Alignment> vous avez pu le constater, le grand "bogue de l'an 2000" ne s'est pas
             produit. En revanche, les citoyens d'un certain nombre de nos pays ont été
            victimes de catastrophes naturelles qui ont vraiment été terribles.</fr>
    </ParallelChunk>
<ParallelChunk annotation_id="426" docOffset="3">
      <en>You have requested a debate on this subject in the course of the next few days,
          during this part-session.</en>
      <fr>Vous avez souhaité un débat <DiscourseConnective annotation_id="466">à</
          DiscourseConnective> ce sujet dans les prochains jours, au cours de cette pé
          riode de session.</fr>
    </ParallelChunk>
```


Figure 4.4: A sample of the Europarl ConcoDisco-*Intersection* corpus.

Corpus	# DU	# NDU	Total
FrConcoDisco-*Intersection*	988K	3,926K	4,914K
FrConcoDisco-*Grow-diag*	1,074K	5,191K	6,265K
FrConcoDisco-*Direct*	1,045K	4,279K	5,324K
FrConcoDisco-*Inverse*	1,090K	5,579K	6,668K
FrConcoDisco-*Naive-Grow-diag*	1,074K	5,839K	6,913K

Table 4.2: Statistics of the FrConcoDisco and FrConcoDisco-*Naive-Grow-diag* corpora.

As Table 4.2 shows, the FrConcoDisco corpora contain significantly different numbers of NDUs. For example, the *Inverse* word-alignment model generated 1,653 thousands more NDU labels than the *Intersection* word-alignment model (5,579K versus 3,926K). Section 4.3.1.2 discusses this difference and its relation to unsupported French candidate discourse connectives.

4.3 Evaluation

To evaluate our approach to filtering unsupported annotations, we proceeded with two methods: 1) an intrinsic evaluation of both NDU/NDU labels and the PDTB relations assigned to the French discourse connectives in the FrConcoDisco corpora (see Section 4.3.1) and 2) an extrinsic evaluation of NDU/NDU labels using the task of disambiguation of French discourse connective usage (see Section 4.3.2).

4.3.1 Intrinsic Evaluation

To intrinsically evaluate the approach, we first built a gold-standard dataset using crowdsourcing (see Section 4.3.1.1), and then compared the FrConcoDisco corpora against this gold-standard dataset (see Section 4.3.1.2).

4.3.1.1 Building a Gold-Standard Dataset

To evaluate if French candidate discourse connectives have the same discourse annotations as their translation, we designed a linguistic test, which we call the *Translatable Test*, inspired by the *Substitutability Test* of Knott (1996, p. 71). To identify if two discourse connectives signal the same relation, Knott (1996) compared a set of sentences where the only difference was the discourse connectives used. If the two sentences conveyed the same meaning then he assumed that the two discourse connectives signal the same relation in that context. For example, the first two sentences in (Ex. 35) (marked with a ✓) convey the same meaning, and therefore we can conclude that *so* and *thereby* signal the same relation in these two sentences. However, the third sentence (marked with a ×) does not convey the same meaning and therefore, it does not support that *in short* can signal the same relation as the other two connectives[5].

(Ex. 35) ✓ *She left the country before the year was up;* ***so*** *she lost her right to permanent residence.*

✓ *She left the country before the year was up; she* ***thereby*** *lost her right to permanent residence.*

× *She left the country before the year was up;* ***in short*** *she lost her right to permanent residence.*

The *Substitutability Test* has also been used by Roze et al. (2012) as one of their linguistic tests to associate discourse relations to French discourse connectives.

Inspired by the *Substitutability Test* test, we designed the *Translatable Test*. Since parallel sentences are a translation of each other, we can assume that they convey the same meaning and we therefore only need to verify if there is an English expression that is a good substitution for the French discourse connective candidate. If this is the case, then we conclude that the French discourse connective candidate should have the same discourse annotation (discourse usage and relation) as their English substitution. Otherwise, we conclude that the French discourse connective candidate cannot be reliably annotated.

To build a gold-standard dataset, we first randomly selected parallel sentences from a random Europarl file[6] containing French candidate discourse connectives. For each French candidate discourse connective, we selected at most 10 parallel sentences to keep the number of sentence pairs tractable and to avoid any bias towards frequent French candidate discourse connectives. This approach generated 696 pairs of parallel sentences for 149 French discourse connectives, similar to the examples in Table 4.1. Then, we used the CrowdFlower platform[7] to run the *Translatable Test*

[5] All sentences are taken from (Knott, 1996).

[6] ep-00-01-17.txt

[7] https://www.crowdflower.com/

on the dataset. To do so, we highlighted the French candidate discourse connectives in each pair of parallel sentences (as shown in the column *French* in Table 4.1) and asked annotators to identify (i.e. copy and paste) the English expression that is the best translation of the French candidate discourse connective or to indicate if the French candidate discourse connective has no translation. Figure 4.5 shows a screenshot of the website designed by us for running the CrowdFlower experiment.

To ensure more accurate results, we limited the annotators to bilingual English-French speakers by setting non-English language skills required on the CrowdFlower website. Moreover, we manually aligned 80 qualifying questions using three bilingual English-French speakers with a background in discourse analysis and filtered annotators whose accuracy was below 0.80 against these test questions. Out of 211 initial annotators, only 33 passed our qualifying questions and proceeded with the actual annotation task. We used the webservice[8] provided by Freelon (2010) to calculate the Krippendorff's Alpha agreement (Krippendorff, 2004) between the 33 annotators. The agreement between annotators was 0.787 which shows a strong agreement according to Krippendorff (2004, pp. 241-243).

The CrowdFlower annotations allowed us to create a corpus of 696 pairs of sentences which we refer to it as the *CrowdFlower gold-standard* dataset. Table 4.3 shows statistics of this dataset. According to the crowdsourced annotators, 31.61% of French candidate discourse connectives can be substituted by an English discourse connective which was marked by the *CLaC DC Disambiguator* and therefore are used in a NDU (as in (Ex. 31) of Table 4.1); while 53.74% can be substituted by an English expression which does not signal any discourse relation according to the *CLaC DC Disambiguator* (as in (Ex 32) of Table 4.1) and is therefore used in a NDU. Finally, 14.66% of the French candidate discourse connectives have no English translation (as in (Ex. 33) or (Ex. 34) of Table 4.1), hence they cannot be reliably annotated. Recall that, as opposed to previous work such as (Versley, 2010), our approach specifically addresses this significant proportion of explicit relations translated as implicit ones.

[8] http://dfreelon.org/utils/recalfront/recal3/

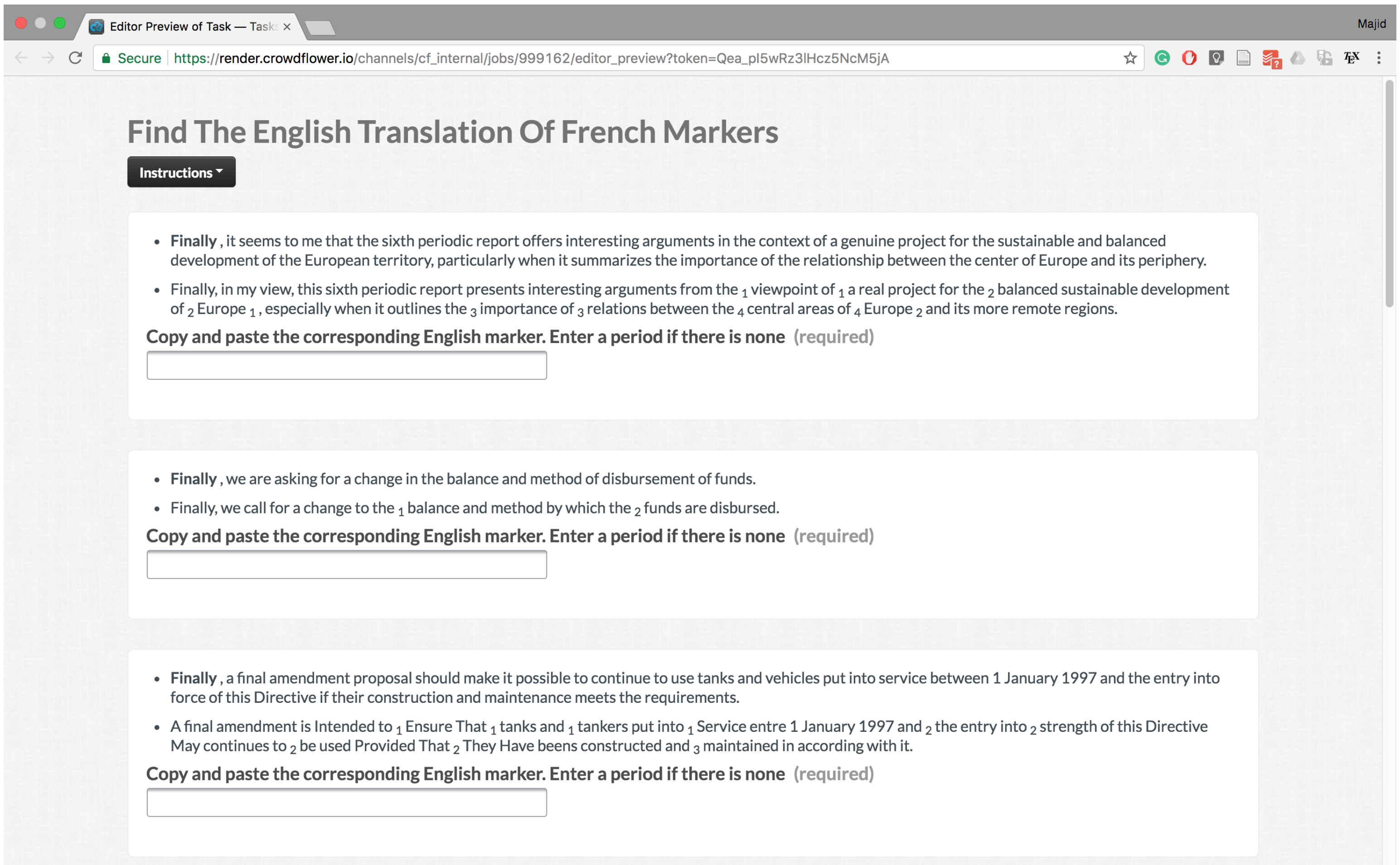


Figure 4.5: A screenshot of the website designed by us for running the CrowdFlower experiment.

French Candidate Discourse Connectives			
Total	**Actual DU**	**Actual NDU**	**Dropped in English**
696 (100%)	220 (31.61%)	374 (53.74%)	102 (14.66%)

Table 4.3: Statistics of the CrowdFlower gold-standard dataset.

4.3.1.2 Evaluation of the FrConcoDisco Corpora

To evaluate the performance of the four word-alignment models in the identification of the English translation of French candidate discourse connectives, we compared the FrConcoDisco corpora generated by the models (see Section 4.2.2) against the CrowdFlower gold-standard dataset (see Section 4.3.1.1). Note that this evaluation shows the performance of the word-alignment models for the *Translatable Test*, and therefore can be also considered as an intrinsic evaluation of the discourse relations assigned to the French candidate discourse connectives[9]. Table 4.4 shows the precision (P) and recall (R) for both NDU and NDU labels, as well as the overall annotations (OA) of the four FrConcoDisco corpora. As Table 4.4 shows, the FrConcoDisco-*Intersection* corpus achieves the highest precision for both NDU labels (0.934) and NDU labels (0.902), at the expense of recall. For example, while the FrConcoDisco-*Intersection* corpus achieves a higher overall precision than the FrConcoDisco-*Naive-Grow-diag* corpus (0.914 versus 0.815), its overall recall is lower (0.845 versus 0.955).

Corpus	**DU**		**NDU**		**OA**	
	P	**R**	**P**	**R**	**P**	**R**
FrConcoDisco-*Intersection*	0.934	0.895	0.902	0.816	0.914	0.845
FrConcoDisco-*Grow-diag*	0.906	0.923	0.814	0.904	0.847	0.911
FrConcoDisco-*Direct*	0.902	0.918	0.883	0.866	0.890	0.886
FrConcoDisco-*Inverse*	0.891	0.927	0.801	0.928	0.832	0.928
FrConcoDisco-*Naive-Grow-diag*	0.906	0.923	0.771	0.973	0.815	0.955

Table 4.4: Precision (P) and recall (R) of the four FrConcoDisco and the FrConcoDisco-*Naive-Grow-diag* corpora against the CrowdFlower gold-standard dataset for NDU/NDU labels and overall (OA).

Because the *Intersection* model suffers from sparsity issues (many words are aligned to null), the *Grow-diag* model is typically used for annotation projection (Tiedemann, 2015; Versley, 2010).

[9]Because we do not have gold discourse annotations for Europarl, we can estimate the quality of the discourse annotations of the English side by evaluating the performance of the *CLaC DC Disambiguator*on texts with a different domain such as the blind dataset of CoNLL shared task (see Section 4.2.1).

However, Table 4.4 shows that the *Intersection* model is more suitable for discourse annotation projection due to its higher precision. Because the FrConcoDisco corpora are much larger than existing discourse corpora (with around 5 million annotations), a higher precision is preferable in our case.

A further error analysis shows that the main advantage of the *Intersection* model is when French candidate discourse connectives are dropped during the translation (i.e. explicit relations that are changed to implicit ones – see the column *Dropped* in Table 4.3). For example in (Ex. 30), *mais* has been dropped in the English translation. This causes both the *Grow-diag* and the *Inverse* models to incorrectly align *mais* to *and*. Hence, when we project the discourse relation for either of these two models, *mais* will be incorrectly marked as *NDU* because *and* is not an English discourse connective. However, *mais* signals a *COMPARISON:Contrast* relation. Therefore, a false-negative instance is generated for *mais*.

Table 4.5 shows the performance of each alignment model for the identification of dropped French candidate discourse connectives against the CrowdFlower gold-standard dataset. While the *Intersection* model identifies the most dropped discourse connectives (65% out of the 102 dropped candidates), the *Inverse* word alignment is the worst model as it identifies only 6% of the dropped candidates and the naive *Grow-diag* approach clearly identifies none. Note that the alignment models tend to label dropped French candidates discourse connectives as NDU more often than as NDU when they cannot identify candidates that were dropped during the translation; therefore, dropped French candidate discourse connectives may artificially increase the number of NDU labels. This also explains why the number of NDU labels for the *Intersection* word-alignment is the lowest among the word-alignment models (see Table 4.2).

4.3.2 Extrinsic Evaluation

To extrinsically evaluate the effect of unsupported annotations on the quality of the FrConcoDisco corpora models, we used the corpora to train a binary classifier in order to detect the discourse usage of French discourse connectives. Since the classifiers only differ by the training set used, by comparing the results of the classifiers, we indirectly assessed the quality of the corpora.

For our experiment, we used the French Discourse Treebank (FDTB) (Danlos et al., 2015).

Corpus	Dropped Candidate DC		
	Identified	Not identified and labeled as DU	NDU
FrConcoDisco-*Intersection*	64%	8%	28%
FrConcoDisco-*Grow-diag*	20%	11%	69%
FrConcoDisco-*Direct*	48%	13%	39%
FrConcoDisco-*Inverse*	6%	17%	77%
FrConcoDisco-*Naive-Grow-diag*	0%	11%	89%

Table 4.5: Accuracy of the four FrConcoDisco and the FrConcoDisco-*Naive-Grow-diag* corpora in the identification of dropped candidate discourse connectives (unsupported candidates) against the CrowdFlower gold-standard dataset.

Recall from Chapter 2 that the FDTB marks French discourse connectives in two syntactically annotated corpora: the Sequoia Treebank (Candito and Seddah, 2012) and the French Treebank (FTB) (Abeillé et al., 2000). We assigned NDU labels to the French discourse connectives marked in the FDTB and NDU labels for all other non-discourse occurrences of the French discourse connectives in the FDTB. Table 4.6 shows statistics of the FDTB.

Corpus	# Words	# DU	# NDU
FTB	557,149	10,437	40,669
Sequoia	33,205	544	2,255
Total	**579,243**	**10,735**	**42,924**

Table 4.6: Statistics of the FDTB.

In our experiments, as with Chapter 3, we used the same classifier used in the *CLaC DC Disambiguator* (Laali et al., 2016) for disambiguating the usage of English discourse connectives and trained it on the four FrConcoDisco corpora, the FrConcoDisco-*Naive-Grow-diag* corpus and the FTB section of the FDTB. We reserved the Sequoia section of the FDTB for the evaluation of the trained classifiers. The text of the Sequoia section of the FDTB is extracted from Wikipedia and the ANNODIS corpus (Afantenos et al., 2012). This allowed us to compare the classifiers on datasets of different domains/genres than the training datasets, therefore, introducing no bias toward any of the training datasets.

Table 4.7 shows the precision, recall and the F1-score of the classifiers. While the precision of classifiers trained on the FrConcoDisco corpora is high (0.831~0.857) and actually higher than the one trained on the manually annotated FTB, their recall is much lower (0.309~0.406). We also observed that the classifiers trained on FrConcoDisco-*Naive-Grow-diag* and on FrConcoDisco-*Grow-diag* have the same performance. This is because the Grow-diag models created many false-negative instances for a set of French discourse connectives. Hence, the classifiers trained on this model labeled all occurrence of these French discourse connectives as NDU. In addition, FrConcoDisco-*Naive-Grow-diag* also added more false-negative instances to the same set of French discourse connectives so the classifier labeled all those French discourse connectives as NDU.

Among the classifiers trained on the FrConcoDisco corpora, the one based on the *Intersection* model again achieved the best performance with an F1-score of 0.546. This confirms that the trade-off between precision and recall achieved by the *Intersection* model makes it the most appropriate for discourse annotation projection.

The low recall of the classifiers trained on the FrConcoDisco corpora is an indication of a large number of false-negative instances. As discussed in Section 4.3.1.2, an important source of false-negative instances is due to French candidate discourse connectives that are dropped in the translation. Table 4.7 shows this by illustrating the same behaviour as in Table 4.5. As these two tables show, the more accurate a word alignment model is at pruning dropped French candidate discourse connectives, the higher recall the classifier will achieve using the dataset extracted from this word alignment model. In our case, the *Intersection* model is the most accurate model in the identification of dropped candidate discourse connectives with an accuracy of 65% (see Table 4.5), and the classifier trained on the FrConcoDisco-Intersection also achieves the highest recall (i.e. 0.406). This classifier achieves a 15% relative improvement in F1-score compared to the one that was trained on FrConcoDisco-*Naive-Grow-diag*. This shows the adverse effect of unsupported annotations on the classifiers.

To investigate further the low recall of the classifiers, we manually analyzed the results of three French discourse connectives with a low recall and a high frequency in the CrowdFlower gold-standard dataset: *enfin*, *afin de* and *ainsi*[10]. We observed that while 96% of the French candidate discourse connectives for these English discourse connectives were properly aligned to their translation, 59% of them were incorrectly labeled as NDU because their English translation were not

[10]Free translation: *enfin* ≈ *finally*, *afin de* ≈ *in order to*, *ainsi* ≈ *so*.

properly annotated. This happened for three main reasons:

(1) The English translation is an English discourse connective, but because it is either infrequent in the PDTB (e.g. *finally*) or its NDU usage dominates its NDU usage (e.g. *for*), the English discourse connective cannot be reliably annotated.

(2) The English translation is an English discourse connective, but it is not listed in the PDTB (e.g. *in order to*).

(3) The English translation is not an English discourse connective, but it signals a discourse relations (e.g. *this would ensure that* or *in this way*). Such expressions are called *AltLex* in the PDTB. We excluded AltLex from our analysis because to our knowledge, no English discourse parser can currently annotate them reliably.

Training Corpus	**P**	**R**	**F1**
FTB	0.777	0.756	0.766
FrConcoDisco-*Intersection-Intersection*	0.831	0.406	0.546
FrConcoDisco-*Intersection-Grow-diag*	0.837	0.331	0.474
FrConcoDisco-*Intersection-Direct*	0.834	0.397	0.538
FrConcoDisco-*Intersection-Inverse*	0.857	0.309	0.454
FrConcoDisco-*Naive-Grow-diag*	0.837	0.331	0.474

Table 4.7: Performance of the classifiers trained on different corpora against the Sequoia test set.

4.4 Conclusion

In this chapter, we have addressed the issue of noisy word-alignments and showed the applicability of discourse annotation projection. We showed that discourse annotations may not always be reliably projected in parallel sentences when discourse relations are changed from explicit to implicit ones during the translation. We proposed a novel approach based on the intersection between statistical word-alignment models to identify unsupported annotations. This approach was able to identify 65% of the unsupported annotations, hence allowing the automatic induction of more precise corpora. As a by-product of our approach, we automatically induced the FrConcoDisco-*Intersection* corpus: the first PDTB style discourse corpora for French. We showed that our approach to filtering unsupported annotations improves the F1-score of a classifier that labels the NDU and the NDU of French discourse connectives by 15% compared to when the unsupported annotations are not filtered.

Chapter 5

Automatic Mapping of French Discourse Connective to Discourse Relations

Building a lexicon of discourse connectives, where each connective is mapped to the discourse relations it can signal, is not an easy task. In this chapter, we present an approach to exploit the Europarl ConcoDisco corpora developed in the previous chapter (see Section 4.2.3), in order to map French discourse connectives to discourse relations. Using this approach, we created *ConcoLeDisCo*, a lexicon of French discourse connectives associated with their PDTB relations. When evaluated against LEXCONN, *ConcoLeDisCo* achieves a recall of 0.81 and an average precision of 0.68 for the *COMPARISON.Concession* and *CONTINGENCY.Condition* relations. *ConcoLeDisCo* is publicly available at `https://github.com/mjlaali/ConcoLeDisCo`. This work has been presented at the SIGdial 2017 conference (Laali and Kosseim, 2017a).

This chapter and next chapter address research question **(Q. 4)** (see Section 1.2):

(Q. 4) **How can lexicons of discourse connectives for the target language be induced from parallel texts?**

To properly answer **(Q. 4)**, we divide this question into two questions:

(Q. 4.a) **How can discourse connectives be mapped to discourse relations using parallel texts?**

(Q. 4.b) **How can a list discourse connectives be induced from parallel text?**

In this chapter, we address **(Q. 4.a)**. More specifically, we assume that a list of French discourse connectives is given and focus on mapping French discourse connectives to PDTB discourse relations. In next chapter, we will present a novel approach to relax this assumption and induce a list of French discourse connectives from parallel texts to answer **(Q. 4.b)**.

5.1 Introduction

To date, to build lexicons of discourse connectives, it is necessary to have linguists manually analyze the usage of individual discourse connectives through a corpus study. This is an expensive endeavour both in terms of time and expertise. As indicated in Section 2.1.3, LEXCONN (Roze et al., 2012) was initiated in 2010 and released its first edition in 2012. The latest version, LEXCONN V2.1 (Danlos et al., 2015), contains 343 discourse connectives mapped to an average of 1.3 discourse relations. This project is still ongoing as 37 discourse connectives still have not been assigned to any discourse relation. Because of this, only a limited number of languages currently possess such lexicons (see Section 2.1.2 for a list of lexicons of discourse connectives for different languages).

In this chapter, we propose an approach to automatically map French discourse connectives to their associated PDTB discourse relations using the Europarl ConcoDisco corpora developed in Chapter 4. To map French discourse connectives to discourse relations any of the Europarl ConcoDisco corpora could have been used, however, in this chapter, we report our results based on the Europarl ConcoDisco-*Naive-Grow-diag* corpus. We chose the Europarl ConcoDisco-*Naive-Grow-diag* corpus because this approach achieves the highest recall when we projected discourse annotations (see Table 4.4). As we see in Section 5.2, the number of mappings between discourse connectives and discourse relation is manageable using our approach, and therefore, it is possible to manually analyze all mappings. This means that, in this context, a higher recall is preferable.

Our approach can also automatically identify the usage of a discourse connective where the discourse connective signals a specific discourse relation. This can help linguists study a discourse connective in parallel texts and/or find evidence for an association between discourse relations and discourse connectives.

Our approach is based on statistical word alignment models (see Chapter 4) and makes no assumption about the target language except the availability of a parallel corpus with another language for which a discourse parser exists; hence the approach is easy to expand to other languages.

As a result of our approach, we generated *ConcoLeDisCo*[1], a lexicon mapping French discourse connectives to their associated Penn Discourse Treebank (PDTB) discourse relations (Prasad et al., 2008a). To our knowledge, *ConcoLeDisCo* is the first lexicon of French discourse connectives mapped to the PDTB relation set. When compared to LEXCONN, *ConcoLeDisCo* achieves a recall of 0.81 and an average precision of 0.68 for the *COMPARISON.Concession* and *CONTINGENCY.Condition* discourse relations.

5.2 Methodology

5.2.1 Dataset Preparation

For our experiments, we used the Europarl ConcoDisco-*Naive-Grow-diag* corpus (see Chapter 4). Any of the Europarl ConcoDisco corpora could have been used, but we chose this corpus because it has the highest recall compared to the other Europarl ConcoDisco corpora (see Table 4.4). A higher recall is more preferable because, for this task, an expert human annotator can manually analyze all induced mappings between French discourse connectives and discourse relations, and eventually flag noisy mappings as opposed to manually identifying missing mapping.

Recall that the Europarl ConcoDisco-*Naive-Grow-diag* corpus contains alignments between the 371 French discourse connectives from LEXCONN V2.1 (Danlos et al., 2015) and the 100 English discourse connectives from the PDTB (Prasad et al., 2008a) within the English-French part of Europarl (Koehn, 2005). Moreover, English discourse connectives were automatically annotated with the subset of 14 PDTB discourse relations that was used in the CoNLL shared task (Xue et al., 2015) using the classifiers presented in Chapter 3. See Chapter 4 for a detailed discussion on how the Europarl ConcoDisco-*Naive-Grow-diag* corpus has been constructed.

[1]*ConcoLeDisCo* is publicly available at `https://github.com/mjlaali/ConcoLeDisCo`.

5.2.2 Mapping Discourse Relations

To label French discourse connectives with a PDTB discourse relation, we assumed that if a French discourse connective is aligned to an English discourse connective tagged with a discourse relation *Rel*, then it should signal the same discourse relation *Rel*. To have statistically reliable results, we ignored French discourse connectives that appeared 50 times or less in Europarl. Out of the 371 French discourse connectives listed in LEXCONN, seven do not appear in Europarl and 55 have a frequency 50 or lower. This means that 89% (309/371) of the French discourse connectives have a frequency higher than 50 and were thus used in the analysis. A manual inspection of the infrequent discourse connectives shows that they are either informal (e.g. *des fois que*) or rare expression (e.g. *en dépit que*). Table 5.1 shows the distribution of the LEXCONN French discourse connectives in Europarl.

	Frequency			
	$= 0$	**≤ 50**	**> 50**	**Total**
# French Discourse Connectives	7	55	309	371

Table 5.1: Distribution of LEXCONN French discourse connectives in the Europarl corpus.

We used the Europarl ConcoDisco-*Naive-Grow-diag* corpus to extract the number of alignments between French discourse connectives and English discourse connectives to create a table that contains the frequency of the alignments between English and French discourse connectives. We refer to this table as the *Connective Translation Table*. Table 5.2 shows a few entries of this table for the French discourse connective *même si*. As the table shows, *même si* was aligned to three different English discourse connectives: *although*, labeled by the classifier as a *COMPARISON.Contrast* or as a *COMPARISON.Concession* and to *even if* and *even though* which were not tagged.

French Connective	**English Connective**	**Relation**	**Freq**
même si	*even if*	-	2538
même si	*even though*	-	1895
même si	*although*	*COMPARISON:Contrast*	1446
même si	*although*	*COMPARISON:Concession*	858

Table 5.2: A few entries of the Connective Translation Table extracted from alignments of the Europarl ConcoDisco-*Naive-Grow-diag* corpus for the connective *même si*.

The Connective Translation Table contains 1,970 entries made of a French discourse connective, an English discourse connective and a discourse relation. From these, we computed the number of times a French discourse connective was aligned to each discourse relation, then, created *ConcoLeDisCo*: tuples of the type <*FR-DC, Rel, Prob*>, where *FR-DC* and *Rel* indicate a French discourse connective and a discourse relation and *Prob* indicates the probability that *FR-DC* signals *Rel*. To calculate *Prob*, we divided the number of times *FR-DC* is associated to *Rel* by the frequency of *FR-DC* in Europarl. In total, the approach generated a lexicon of 900 such tuples, a few of which are shown in Table 5.3. *ConcoLeDisCo* is available in Appendix D and an electronic version is available on https://github.com/mjlaali/ConcoLeDisCo.

FR-DC	**Relation**	**Prob**
si	*COMPARISON:Condition*	0.27
même si	*COMPARISON:Concession*	0.08
lorsque	*COMPARISON:Condition*	0.05
néanmoins	*COMPARISON:Concession*	0.07

Table 5.3: A few entries of *ConcoLeDisCo*. (See Appendix D for the entire lexicon)

5.3 Evaluation

To evaluate *ConcoLeDisCo*, because LEXCONN uses a different inventory of discourse relations than the PDTB, we only considered the discourse relations that are common across these inventories: *COMPARISON.Concession* and *CONTINGENCY.Condition*. According to LEXCONN, 61 French discourse connectives can signal a *COMPARISON.Concession* or a *CONTINGENCY.Condition* discourse relation. Out of these, 44 have a frequency higher than 50 in Europarl. These discourse connectives are listed in Table 5.4.

5.3.1 Automatic Evaluation

To measure the quality of *ConcoLeDisCo*, we ranked the <*FR-DC, Rel, Prob*> tuples based on their probability and measured the quality of the ranked list using 11-point interpolated average precision (Manning and Schutze, 2008). This curve shows the highest precision at the 11 recall levels of 0.0, 0.1, 0.2, ..., 1.0. This method allows us to evaluate the ranked list without considering any arbitrary cut-off point. As Figure 5.1 shows, the approach retrieved 50% of the French discourse connectives in LEXCONN with a precision of 0.81.

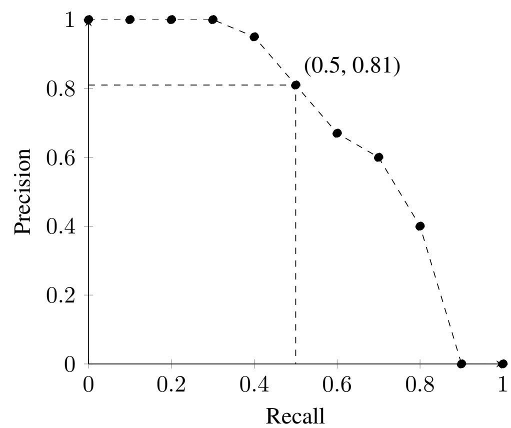


Figure 5.1: 11-Point Interpolated Average Precision curve.

In addition, we also computed Average Precision (AveP) (Manning and Schutze, 2008); the average of the precision obtained after seeing a correct LEXCONN entry in *ConcoLeDisCo*. More specifically, given a list of ranked tuples:

$$AveP = \frac{1}{N}\sum_{i=1}^{N} Precision(DC_i) \tag{1}$$

where N is the number of LEXCONN French discourse connectives that signals the *COMPARISON.Concession* or *CONTINGENCY.Condition* discourse relations (i.e. 44), DC_i is the rank of the i^{th} LEXCONN discourse connective in *ConcoLeDisCo*, and $Precision(DC_i)$ is the precision at the rank DC_i of the ranked tuples. It can be shown that $AveP$ approximates the area under the interpolated precision-recall curve (Manning and Schutze, 2008). The proposed approach identified 36 (81%) of these 44 French discourse connectives with an $AveP$ of 0.68.

5.3.2 Manual Evaluation

In addition to the quantitative evaluation, we also performed a manual analysis of the false-positive errors to see if they really constituted errors. To do so, we looked at the tuples with a probability higher than 0.01 but which did not appear in LEXCONN. Fourteen such cases, shown in Table 5.5, were found.

For example, while the French connective *à défaut de* (#1 in Table 5.5) signals a *CONTINGENCY.Condition* discourse relation in Sentence (1) below, only the EXPLANATION[2] and the *COMPARISON.Concession* discourse relations were associated with this connective in LEXCONN.

(1) **FR:** À défaut de se montrer très ambitieux, notre industrie, nos chercheurs et nos experts ne disposeront purement et simplement pas du brevet moderne dont ils ont besoin.

EN: If we are anything less than ambitious in this field, we shall simply not provide our industry, our research and development experts with the modern patent which they need.

To evaluate if these 14 cases were true mistakes, we randomly selected five English-French parallel sentences from Europarl that contained the French discourse connective and one of its English discourse connective translations signaling the discourse relation. Then, we showed the French discourse connectives within their sentence to two native French speakers and asked them to confirm if the discourse relation identified was indeed signaled by the French discourse connectives or not. The Kappa agreement between the two annotators was 0.72. For 9 French connectives, both annotators agreed that in at least one of the five sentences, the discourse relation was signaled by the connective. This indicates that 64% (9/14) are in fact true-positives, i.e. correct mappings that are not listed in LEXCONN. Table 5.5 shows the 14 pairs of <FR-DC/English translation, Discourse relation> used in the manual evaluation and indicates the newly discovered mappings with a ✓.

We also observed that if multiple explicit connectives occur in the same clause (e.g. *certes* and *mais*), one of them can affect the discourse relation signaled by the other. This is an interesting phenomenon as it seems to indicate that connectives are not independent. For example, in Sentence (2), the combination of *certes* and *mais* signals a *COMPARISON.Concession* discourse relation. But according to LEXCONN, neither *certes* nor *mais* can signal a *COMPARISON.Concession* discourse relation.

(2) **FR:** Cela coûte certes un peu plus cher, mais est sans conséquence pour l'environnement.

EN: Although it is a little more expensive, it does not harm the environment.

[2] EXPLANATION is not among the PDTB discourse relations and has only been defined in SDRT (see Chapter 2). The most similar PDTB relation to EXPLANATION is *CONTINGENCY.Cause.reason*.

The same phenomenon was also reported for English in the PDTB corpus (Prasad et al., 2008b, p. 5).

5.4 Conclusion

In this chapter, we proposed a novel approach to automatically map PDTB discourse relations to French discourse connectives. Using this approach, we generated *ConcoLeDisCo*: a lexicon of French discourse connectives mapped to their PDTB discourse relations. When compared with LEXCONN, our approach achieved a recall of 0.81 and an Average Precision of 0.68 for the *COMPARISON.Concession* and *CONTINGENCY.Condition* discourse relations. A manual error analysis of the false-positives showed that the approach identified new discourse relations for 9 French discourse connectives which are not included in LEXCONN.

In this chapter, we used LEXCONN to extract a list of discourse connectives to build *ConcoLeDisCo*. In the next chapter, we present an automatic approach to extract such a list from parallel texts; which complements the approach described in the current chapter to build an end-to-end extractor of lexicons of discourse connectives from parallel texts.

	French Connective	Relations
1	*dire encore qu', dire encore que, dire qu', dire que*	*COMPARISON.Concession*
2	*dans la mesure où*	*CONTINGENCY.Condition*
3	*dans l'hypothèse où*	*CONTINGENCY.Condition*
4	*pourvu qu', pourvu que*	*CONTINGENCY.Condition*
5	*dès lors qu', dès lors que*	*CONTINGENCY.Condition*
6	*à condition d', à condition de*	*CONTINGENCY.Condition*
7	*le jour où*	*CONTINGENCY.Condition*
8	*du moment qu', du moment que*	*CONTINGENCY.Condition*
9	*à supposer qu', à supposer que*	*CONTINGENCY.Condition*
10	*bien qu', bien que*	*COMPARISON.Concession*
11	*si ce n'est qu', si ce n'est que*	*COMPARISON.Concession*
12	*malgré qu', malgré que*	*COMPARISON.Concession*
13	*tout en*	*COMPARISON.Concession*
14	*même en, notamment en, qu'en*	*CONTINGENCY.Condition*
15	*s', si*	*CONTINGENCY.Condition*
16	*en supposant qu', en supposant que*	*CONTINGENCY.Condition*
17	*soit dit en passant*	*COMPARISON.Concession*
18	*et dire qu', et dire que*	*COMPARISON.Concession*
19	*a fortiori s', a fortiori si, que s', que si, surtout s', surtout si*	*CONTINGENCY.Condition*
20	*s', si*	*COMPARISON.Concession*
21	*en même temps qu', en même temps que*	*COMPARISON.Concession*
22	*quand bien même*	*COMPARISON.Concession*
23	*en dépit du fait qu', en dépit du fait que*	*COMPARISON.Concession*
24	*aussi longtemps qu', aussi longtemps que*	*CONTINGENCY.Condition*
25	*pour peu qu', pour peu que*	*CONTINGENCY.Condition*
26	*à défaut d', à défaut de*	*COMPARISON.Concession*
27	*même quand*	*CONTINGENCY.Condition*
28	*alors même qu', alors même que*	*COMPARISON.Concession*
29	*quand*	*CONTINGENCY.Condition*
30	*pour autant qu', pour autant que*	*CONTINGENCY.Condition*
31	*à condition qu', à condition que*	*CONTINGENCY.Condition*
32	*quoiqu', quoique*	*COMPARISON.Concession*
33	*en*	*CONTINGENCY.Condition*
34	*à partir du moment où*	*CONTINGENCY.Condition*
35	*cependant qu', cependant que*	*COMPARISON.Concession*
36	*dans le cas où*	*CONTINGENCY.Condition*
37	*malgré le fait qu', malgré le fait que*	*COMPARISON.Concession*
38	*pourtant*	*COMPARISON.Concession*
39	*encore qu', encore que*	*COMPARISON.Concession*
40	*même s', même si*	*COMPARISON.Concession*
41	*dès qu', dès que*	*CONTINGENCY.Condition*
42	*tant qu', tant que*	*CONTINGENCY.Condition*
43	*au cas où*	*CONTINGENCY.Condition*
44	*si tant est qu', si tant est que*	*CONTINGENCY.Condition*

Table 5.4: 44 French connectives with a frequency higher than 50 in Europarl.

Fr Connective / En Translation	Relation	Jdg	Fr Connective / En Translation	Relation	Jdg
à défaut de *if*	*CONTINGENCY.* *Condition*	✓	*tout de même* *nonetheless*	*COMPARISON.* *Concession*	✓
cependant *nonetheless*	*COMPARISON.* *Concession*	✓	*toutefois* *nonetheless*	*COMPARISON.* *Concession*	✓
faute de *if*	*CONTINGENCY.* *Condition*	✓	*pour autant* *if*	*CONTINGENCY.* *Condition*	×
malgré tout *nonetheless*	*COMPARISON.* *Concession*	✓	*sinon* *if*	*CONTINGENCY.* *Condition*	×
néanmoins *nonetheless*	*COMPARISON.* *Concession*	✓	*certes* *although*	*COMPARISON.* *Concession*	×
nonobstant *although*	*COMPARISON.* *Concession*	✓	*lorsque* *if*	*CONTINGENCY.* *Condition*	×
quand même *nonetheless*	*COMPARISON.* *Concession*	✓	*pour que* *if*	*CONTINGENCY.* *Condition*	×

Table 5.5: Error analysis of the potential false positive entries. ✓indicates newly discourse mappings which are not included in LEXCONN.

Chapter 6

Inducing a List of French Discourse Connectives

As discussed in Chapter 2, building a lexicon of discourse connectives is a valuable resource and is an important step towards building PDTB-style corpora. Nevertheless, building these lexicons is a time-consuming and expensive task and as a consequence, many languages lack such resources. The approach presented in Chapter 5, automatically mapped discourse connectives to discourse relations, but used a pre-existing lexicon of connectives to start the process. In this chapter, our focus is to automatically induce a list of discourse connectives from parallel texts, so that no manually-built lexicon is needed.

This chapter complements the previous chapter to addresses our last research question **(Q. 4)** (see Section 1.2):

(Q. 4) **How can lexicons of discourse connectives for the target language be induced from parallel texts?**

As mentioned in the previous chapter, we divide question **(Q. 4)** into the following two questions:

(Q. 4.a) **How can discourse connectives be mapped to discourse relations using parallel texts?**

(Q. 4.b) **How can a list discourse connectives be induced from parallel text?**

Chapter 5 addressed question **(Q. 4.a)** and this chapter addresses question **(Q. 4.b)**. Answering these two question will allow us to define an approach to automatically build a lexicon of discourse connectives mapped to discourse relations from parallel texts.

To answer **(Q. 4.b)**, we propose a novel approach that exploits collocation extraction techniques. The approach is based on the identification of candidate connectives and ranking them using the Log-Likelihood Ratio. Then, it relies on several filters to filter this list of candidates, namely: Word-Alignment, POS patterns, and Syntactic information.

Using this approach, we have extracted several lists of discourse connectives. Compared to LEXCONN, we have achieved the best result in term of Average Precision (AveP) with the Syntactic Filter. A manual error analysis of the extracted discourse connectives shows that 31 new discourse connectives not listed in LEXCONN were identified.

6.1 Methodology

Our approach to extract discourse connectives consists of two main steps. The first step is the preparation of the parallel corpus with discourse annotations; while the second mines the parallel corpus to identify discourse connectives.

6.1.1 Preparing the Parallel Corpus

Our experiment has focused on building a list of French discourse connectives from English. In order to build the English-French parallel corpus with discourse annotations, we again used the English-French part of the Europarl parallel corpus (Koehn, 2005). To label discourse relations in the parallel text, we have automatically parsed the English side using the PDTB-style End-To-End Discourse parser[1] (Lin et al., 2010). This parser has been trained on Section 02-22 of the PDTB corpus (Prasad et al., 2008a) and can identify and label a discourse connective with PDTB discourse relations at the second-level with 81.19% precision[2] when tested on Section 23 of the PDTB.

[1]At the time of this experiment, since the *CLaC DC Disambiguator* had not been developed yet, we used the PDTB-style End-To-End Discourse parser which was the state-of-the-art discourse parser at the time.

[2]Since the PDTB-style End-to-End Discourse parser uses a different set of discourse relations, this number cannot be compared with the precision of the *CLaC DC Disambiguator*.

After tagging the English text, we kept only parallel sentences whose English translation had exactly one discourse relation. This was done to ensure that no ambiguity would exist in the discourse relation of the French sentences, once we transfer the discourse relation from English to French. In other words, we can label each French sentence with a single discourse relation, that of its English translation. In addition, we have also removed sentences whose discourse relations were expressed implicitly. Although the (Lin et al., 2010) parser is able to identify both implicit and explicit discourse relations, we have only considered relations expressed with a discourse connective. This has been done, since not only the precision of the parser in detecting discourse relations in the absence of discourse connectives is very low (24.54%), but also we would not expect implicit relations to help us to identify new discourse connectives in French. In other words, this would be only useful if a translator inserts a new French discourse connective that was not present in the translation of explicit discourse relations[3]. Therefore, we would not expect that too many new discourse connectives would exist in the translation of sentences with an implicit discourse relation.

Table 6.1 provides statistics on the original English-French Parallel Corpus and the corpus extracted with exactly one explicit discourse relation per sentence. Initially, the Europarl parallel corpus contained 2,054K sentences (57 million and 63 million words in the English and the French sides respectively). However, after removing the sentences with no relations or more than one discourse relation, the corpus was reduced to 543K sentences automatically annotated with a single discourse relation. The English sentences contain 14 million words, while the French counterparts contain 15 million words.

	# Parallel Sentences	**# English Words**	**# French Words**
Original Europarl Corpus	2,054K	57M	63M
Extracted Corpus	543K	14M	15M

Table 6.1: Statistics on the parallel corpora created.

Although this new annotated corpus represents only 26% of the original French Europarl, the corpus still represents a large annotated corpus with respect to existing discourse-annotated corpora. For example, the corpus is almost 14 times bigger than the PDTB. Therefore, due to the large size

[3]Also note that our experiment shows that only at most 14.66% of the time a discourse connective may be inserted in translation texts (see Table 4.3).

of the corpus, it can be expected that eventual errors in the corpus (e.g. sentences whose discourse relations have been changed during the translation) should not affect the results significantly.

6.1.2 Mining the Parallel Corpus

Once the aligned corpus has been built, we have used Algorithm 4 to mine the French side and build a lexicon of potential French discourse connectives. The inputs of our algorithm are a list of French sentences ($sents$) along with the discourse relations signalled within these sentences ($relations$). We have extracted these two inputs from the aligned corpus. Our algorithm has two parameters: 1) $maxLength$ is a maximum length of French discourse connectives that the algorithm will generate. 2) $threshold$ is a minimum frequency for French discourse connectives in the input $sents$. For our experiments, because the French discourse connectives listed in LEXCONN have a maximum length of 6 words, we have set $maxLength$ to this value. Moreover, based on our analysis on the corpus (see Section 6.2.3), we have set the value of $threshold$ to 10.

In our algorithm, for each pair of French sentence and the relation signalled within sentences (Line 2-4), we have extracted n-grams from the French sentences as potential candidates to be discourse connectives (Line 6). Then, we have stored each potential candidate with its discourse relation as a pair (Line 7). For example, in (Ex. 36), the French sentence contains an *EXPANSION:Alternative* relation.

(Ex. 36) Donc, d'un point de vue judiciaire, il convient de prendre des mesures. (*EXPANSION:Alternative*)

We have therefore produced the following pairs from this French sentence:

(1) (Donc, ALTERNATIVE), (d, ALTERNATIVE), ...

(2) (Donc d, ALTERNATIVE), (d un, ALTERNATIVE), ...

(3) (Donc d un, ALTERNATIVE), (d un point, ALTERNATIVE), ...

(4) ...

Algorithm 4: Build-Lexicon-French-DC

Input: $sents$: a list of French sentences.

Input: $relations$: a list of relations signalled in $sents$.

Input: $maxLength$: a maximum length for French discourse connectives.

Input: $threshold$: a minimum frequency for the French discourse connectives.

Output: $tuples$: a ranked list of potential French discourse connectives.

$pairs = \{\}$;
for $i \leftarrow 1$ **to** $Length(sents)$ **do**
    $relation = relations[i], sent_{fr} = sents[i]$;
    for $begin \leftarrow 1$ **to** $Length(sent_{fr})$ **do**
        for $len \leftarrow 1$ **to** $maxLength$ **do**
            $ngram = GetNGrams(begin, len, sent_{fr})$;
            $\{ngram, rel\} \rightarrow pairs$;
        end
    end
end
$tuples = \{\}, N = Length(pairs)$;
foreach $(ngram, rel) \in pairs$ **do**
    $O_{1,1} = counts((ngram, rel), pairs)$;
    $O_{1,2} = counts((*, rel), pairs)$;
    $O_{2,1} = counts((ngram, *), paris)$;
    $O_{2,2} = N - (O_{1,1} + O_{1,2} + O_{2,1})$;
    if $O_{1,1} > threshold$ **then**
        $LLR = CalculateLLR(O_{1,1}, O_{1,2}, O_{2,1}, O_{2,2})$;
        $\{ngram, rel, LLR\} \rightarrow tuples$
    end
end
$tuples = SortBasedOnLLR(tuples)$

Next, we have used LLR to rank the extracted pair[4] (Line 11-20). LLR evaluates association strength between a pair of events based on their frequency. This measure has been largely used, for example in collocation extraction (e.g. (Seretan, 2010)). According to Evert (2004), LLR is equivalent to the average mutual information that one event conveys about the other. For the sake of completeness, Figure 6.1 shows the formula used to calculate LLR for two binary random variables X and Y. Note that in Figure 6.1, O refers to the observed frequencies, E refers to the expected frequencies and N refers to the total number of observations.

$$LLR(X,Y) = 2 \times \sum_{i=1}^{2}\sum_{j=1}^{2} O_{ij} \times log(\frac{O_{ij}}{E_{ij}})$$

$$E_{ij} = \frac{\sum_{k=1}^{2} O_{ik} \times \sum_{k=1}^{2} O_{kj}}{N}, \; N = \sum_{i=1}^{2}\sum_{j=1}^{2} O_{ij}$$

	$Y = v$	$Y = \neg v$
$X = u$	O_{11}	O_{12}
$X = \neg u$	O_{21}	O_{22}

Figure 6.1: The formula used to calculate Log-Likelihood Ratio (LLR).

In our configuration, our pairs of events consist of the observation of a discourse relation and a discourse connective candidate. We have computed contingency tables of frequencies of these pairs from the pairs (Line 13-16) and then used the NSP package (Pedersen et al., 2011) to calculate the LLR for each pairs that has a frequency higher than the $threshold$ (Line 17-20). Finally, we ranked these pairs based on their LLR score (Line 22).

Once the initial list of discourse connectives has been extracted and ranked based on their LLR score, we have experimented with two types of filters to refine it:

(1) **Word-Alignment Filter:** This filter removes any discourse connective candidate that does not align with any part of an English discourse connective. In other words, as with our approach for discourse annotation projection (see Chapter 4), this filter keeps any consecutive words in the French text if at least one of its composing words aligns to at least one word of an English discourse connective when using a word-alignment model. To have a higher recall, as with building Europarl ConcoDisco-*Grow-diag*, we used *Grow-diag* word alignments[5], a

[4] We have also used other association measures, such as PMI, t-score test, and Chi-square test, but LLR achieved the best results in terms of Average Precision.

[5] We have also experimented with other word-alignment models but their performances were not better. The *Grow-diag* model outperformed the *Direct* word-alignment model and achieved similar results as the *Inverse* word-alignment model.

combination of alignments of the *Direct* word-alignments and the *Inverse* word-alignments based on the heuristic proposed by Och and Ney (2003). We have used MGIZA++ (Gao and Vogel, 2008) to generate *Direct* and *Inverse* word-alignments; then used Moses (Koehn et al., 2007) to compute the *Grow-diag* word alignment. Figure 6.2 presents the *Grow-diag* alignments for two parallel sentences. An alignment between two words is shown by a line connecting them. For example, in these sentences, the connective *therefore* is aligned to the three French words *raison pour laquelle*.

(2) **Syntactic Filters:** As we saw in Chapter 2, discourse connectives are defined as syntactically well-defined terms (Prasad et al., 2008a). The syntactic filters exploit this property and remove any constituent that does not fall into expected syntactic categories. In other words, these filters keep only Prepositional Phrases (PP), Coordinate Phrases (CP) or Adverbial Phrases (ADVP). We have implemented two types of Syntactic Filters. The first one (called **POS Filter**) uses predefined Part-of-Speech (POS) patterns to filter out incorrect candidates. We have manually defined POS patterns based on an analysis of the French discourse connectives in the LEXCONN resource (Roze et al., 2012). Table 6.2 shows the POS patterns we have used along with an example. The second approach (called **Parse Tree Filter**) makes use of the Syntactic Trees to filter unlikely syntactic constituents. Therefore, after parsing all the French sentences, the Syntactic Filter only kept PPs, CPs and ADVPs. We have used the Stanford POS Tagger (Toutanova et al., 2003) and the Stanford PCFG Parser (Green et al., 2011) for POS tagging and parsing the French text, respectively.

POS Pattern	**Example**	**POS Pattern**	**Example**
ADV	alors	P ADV	après tout
C	et	P N	par exemple
P	comme	P P	avant de
ADV C	encore que	V C	considérant que
ADV P	en outre	N D P	de ce fait
C C	parce que	P N P	de manière à
N P	histoire de	P D N	dans ce cas

Table 6.2: POS patterns used in the POS filter.

[5]The examples in this figure are taken from the Europarl parallel corpus.

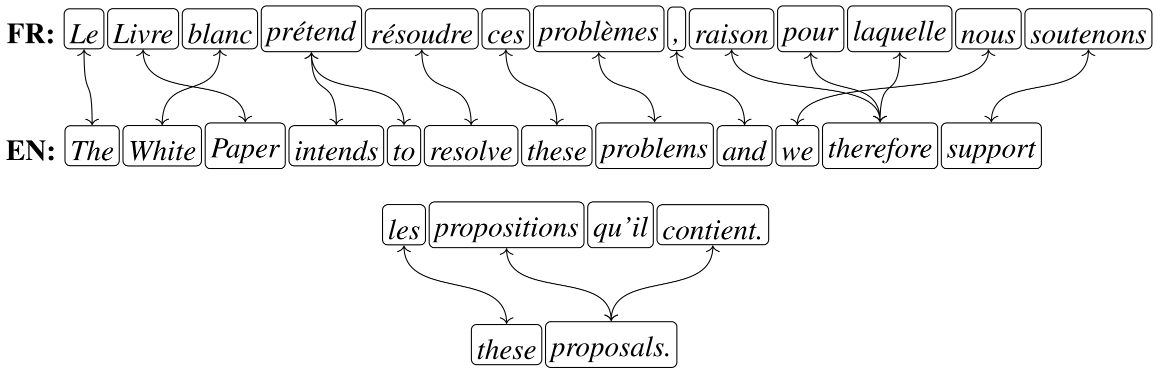


Figure 6.2: Example of word-alignments between English and French texts.[5]

6.2 Evaluation

6.2.1 Gold Dataset

To evaluate our final ranked list of French discourse connectives candidates and compare the four filters, we have used the LEXCONN V1.0 dataset[6] (Roze et al., 2012). Recall from Chapter 2 that LEXCONN V1.0 includes 328 French discourse connectives, 43 less than LEXCONN V2.0. For our experiment, we considered different spellings of the 328 French discourse connective of LEXCONN (e.g. *alors que* and *alors qu'*) as our target expressions. This created 467 target expressions. Table 6.3 provides some statistics about the French connectives in LEXCONN V1.0. We also provide statistics about the discourse connectives in PDTB for comparative purposes. Each row of Table 6.3 indicates the number of discourse connectives and the average number of relations per discourse connective in parenthesis. For example, in LEXCONN, 70 discourse connectives are unigrams and on average they indicate 1.66 different discourse relations. Table 6.3 also shows statistics on the length of discourse connectives (in number of words). It is interesting to note that French tends to have longer discourse connectives than English. Indeed LEXCONN contains 69 discourse connectives that contain four words (e.g. *au même titre que*, *dans l'espoir de*, etc.) while there are only 4 four-gram discourse connectives in English (e.g. *as it turns out* or *on the other hand*).

Although there are fewer relations in PDTB, English discourse connectives tend to be more

[6]At the time of this experience, LEXCONN V2.0 was not publicly available.

[7]As the parser labels relations at the second level of the PDTB hierarchy, we here report only the number of second level relations.

	LEXCONN (French)	PDTB Discourse Connectives (English)
# Discourse relation	29	16[7]
# Total number of discourse connectives	467 (1.29)	133 (3.05)
# Unigram discourse connectives	70 (1.66)	76 (3.50)
# Bigram discourse connectives	169 (1.25)	33 (2.70)
# Trigram discourse connectives	139 (1.22)	18 (2.11)
# Four-gram discourse connectives	69 (1.17)	4 (2.50)
# Five-gram discourse connectives	14 (1.07)	1 (1.00)
# Six-gram discourse connectives	5 (1.20)	0 (-)
# Seven-gram discourse connectives	1 (2.00)	1 (1.00)

Table 6.3: Statistics on discourse connectives in LEXCONN V1.0 and PDTB.

ambiguous. As Table 6.3 shows, each English discourse connective conveys 3.05 relations on average, while this number is 1.29 for French discourse connectives. We also notice that the longer the discourse connective, the less ambiguous it is in terms of discourse relations it can convey. For example, unigram discourse connectives in French convey on average 1.66 relations, however the number of relations decreases when the length of the discourse connective increases, so that for a trigram discourse connective, on average, there are 1.22 relations.

6.2.2 Evaluation Metric

Since our task is very similar to a collocation extraction task, we have used a similar evaluation methodology to evaluate our results. More specifically, we have used the Algorithm 4 and filters defined in Section 6.1.2 to rank the list of potential discourse connectives based on their LLR. Then, we measured the quality of the ranked list of discourse connectives with 11-point interpolated average precision curve (Manning and Schutze, 2008) and Average Precision (AveP) (Manning and Schutze, 2008) (see Section 5.3.1 for details on these metrics.). As Pecina (2010) noted for the evaluation of collocation extraction, since the precision is not reliable at low recall levels and changes frequently at high recall levels, we only considered average precision (AveP) in the interval of <0.1, 0.9> when we are calculating AveP.

Another consideration when evaluating our final ranked lists is how to evaluate discourse connective fragments. For example, when evaluating the candidate *à ce point*, we have to label it as a wrong discourse connective because it is not listed in LEXCONN. However, it is a segment of the French discourse connective *à ce point que* and only one word is missing in the expression. This issue has been also addressed in the field of collocation extraction; in particular, Kilgarriff et al. (2010) suggested to consider a partial collocation as a true positive, since it signals the presence of the longer collocation. However, this was not a decision that human evaluators were comfortable with (Kilgarriff et al., 2010). In our evaluation, we have used two approaches to evaluate fragment discourse connectives. In the first approach, the Exact Match approach, we have considered fragment discourse connectives as an incorrect discourse connective. In the other approach, the Exclude-From-The-List approach, we have removed them from our list, so that when we analyzed the find list, they do not appear as an incorrect discourse connective.

6.2.3 Automatic Evaluation

To evaluate the discourse connective extraction approach, we first analyzed the candidate generation step without any filtering. Table 6.4 provides the frequency distribution of LEXCONN's discourse connectives in the annotated corpus. This table shows that the longer the discourse connectives, the less frequent they are in our corpus. For example, all one-word discourse connectives of LEXCONN appear in the corpus, while 21% of LEXCONN's five-gram and 60% of LEXCONN's six-gram discourse connectives never occur in the corpus. Overall, 14% of all LEXCONN discourse connectives do not appear in the corpus.

	freq > 10	**$10 \geq$ freq > 0**	**freq $= 0$**
# Unigram discourse connectives	93%	7%	0%
# Bigram discourse connectives	76%	16%	8%
# Trigram discourse connectives	60%	24%	16%
# Four-gram discourse connectives	36%	31%	33%
# Five-gram discourse connectives	50%	29%	21%
# Six-gram discourse connectives	20%	20%	60%
Overall	**66%**	**20%**	**14%**

Table 6.4: Distribution of LEXCONN discourse connectives in the extracted corpus.

For our experiments, we set $threshold$ to 10 in Algorithm 4. This threshold removed an additional 20% discourse connectives, so that overall only 66% of LEXCONN's discourse connectives are considered in the corpus. Most of these removed discourse connectives are not common or rather formal expressions in French such as *conséquemment*, *hormis que* or *tout bien considéré*. However, several more informal discourse connectives commonly used in French were also removed, especially discourse connectives of three words or more (e.g. *à part ca*).

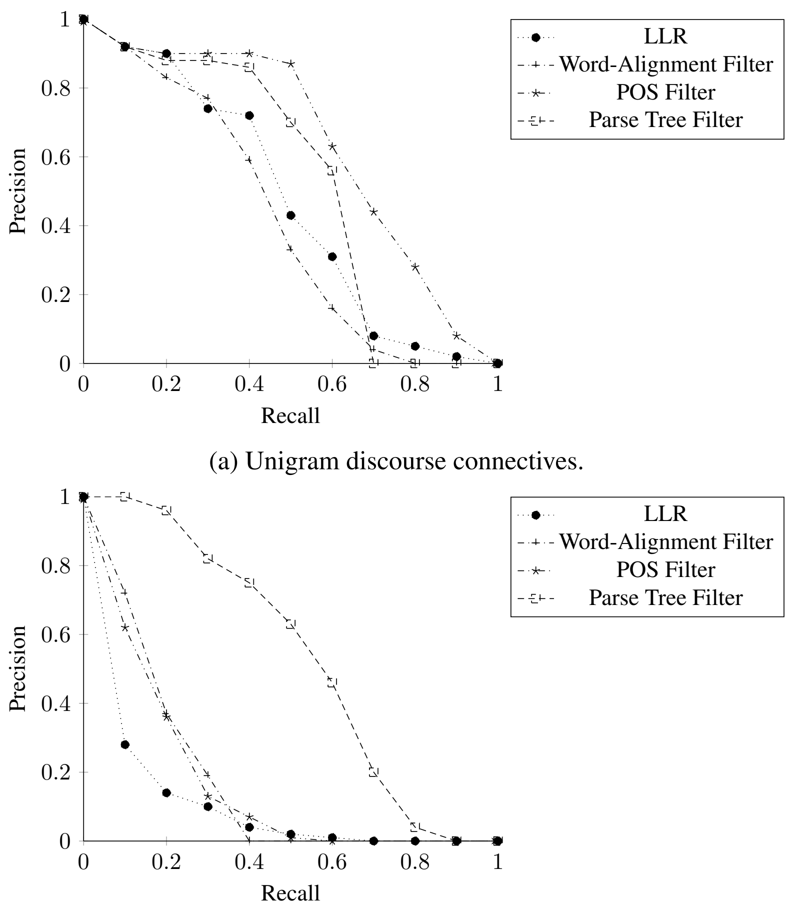
Filter	**AveP with Exact Match**	**AveP with Exclude-From-The-List**
LLR only	0.06	0.07
LLR + Word-Alignment Filter	0.10	0.12
LLR + POS Pattern Filter	0.12	0.14
LLR + Parse Tree Filter	0.39	0.44

Table 6.5: Average Precision of each filter.

Once we calculated the number of available discourse connectives in the corpus, we evaluated the ranked list of discourse connectives after applying each filter. Table 6.5 shows the AveP values of each filter using both the Exact Match and Exclude-From-The-List approaches to judge fragment discourse connectives[8] (see Section 6.2.2). With all four filters, we first used the Frequency Filter and then ranked the candidates using LLR. Our results show that using the POS Pattern Filter outperforms the Word-Alignment Filter. For example, if we consider the Exact Match metric, the AveP value of the Word-Alignment is 0.10 while it is 0.12 for the POS-Pattern Filter. As Table 6.5 shows, the best AveP values are achieved using the Syntactic Filter. For the rest of chapter, we only consider the Exclude-From-The-List approach to judge fragment discourse connectives, since we would like to focus on other sources of errors in the ranked list of discourse connectives in addition to the fragment discourse connectives.

After analyzing the list of discourse connectives generated by all approaches, we noted that the size of a discourse connective affects the performance of our approach. Figure 6.3 shows the performance of each filter when detecting unigram (Figure 6.3a) and bigram (Figure 6.3b) discourse connectives. These figures shows that except for the Parse Tree Filter, the performance of the identification of bigram discourse connectives drops rapidly when compared with the identification of unigram discourse connectives.

[8]When calculating recall points, we only considered the available discourse connectives in the dataset after applying the Frequency Filter (i.e. 66% of the discourse connectives).



(a) Unigram discourse connectives.

(b) Bigram discourse connectives.

Figure 6.3: 11-Point Interpolated Average Precision curve for the extraction of unigram and bigram discourse connectives.

6.2.4 Error Analysis

To better understand why longer discourse connectives are more difficult to identify, we manually analyzed the errors of each filters. The most significant proportion of errors with bigram discourse connectives are composed of a unigram discourse connective and a noisy word. For example, *mais je* is composed of the French discourse connective *mais* and a noisy word *je*. As these errors usually do not create a syntactic well-defined constituent, they can only be filtered out by the Parse Tree Filter.

The POS Pattern Filter cannot detect noisy syntactic components since detecting such components needs contextual syntactic information. When we analyzed negative examples of this filter, we noticed that most of bigram errors are comprised of two words that belong to two different chunks.

For example, in (Ex. 37), the POS pattern “ADV C” extracts *donc que*, but these two words belong to two different syntactic constituents (i.e *ADV* and *Ssub*) as shown in parse tree of Figure 6.4.

(Ex. 37) Je demande donc que l’on soutienne l’Irlande dans ce cas particulier.

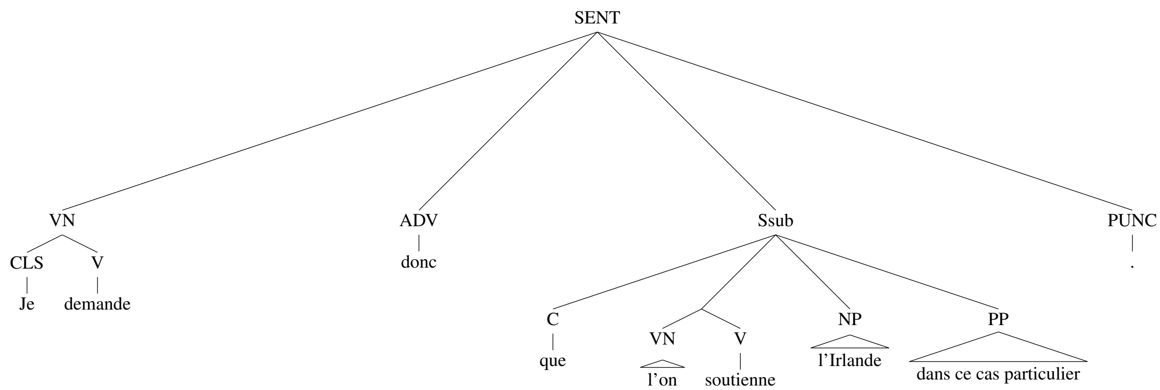


Figure 6.4: The parse tree generated by the Stanford parser for (Ex. 37).

It is interesting to note that the ranked list created with the Parse Tree Filter includes several discourse connectives that do not appear in the LEXCONN lexicon but are nevertheless correct discourse connectives in French. Among the top 100 candidates labeled as an incorrect discourse connective, we have found 31 correct discourse connectives which are not listed in LEXCONN V1.0, such as *toutefois*, *certes* and *au lieu de cela*. The work of (Roze et al., 2012) (or any manually curated list of discourse connectives) constitutes an invaluable resource. However, as Prasad et al. (2010) mentioned, discourse connectives are open-class terms. Therefore, our approach to induce discourse connectives from parallel texts can be used to improve the coverage of such a list.

The results of the Word-alignment show that the Grow-diag word-alignment model cannot align discourse connectives from English onto French. Indeed, our analysis shows that only 176 LEXCONN discourse connectives (38%) were aligned to English discourse connectives. We believe that since a discourse relation can be conveyed with different discourse connectives and human translators can choose between them during the translation, aligning discourse connectives is much harder for alignment models. Moreover, discourse connectives can be also placed at the beginning or at the end of discourse segments, therefore the word-alignment needs to tolerate long-distance alignment to align them.

6.3 Conclusion

In this chapter, we have presented an approach to induce discourse connectives from a parallel text. Our approach extracts a list of discourse connective candidates and ranks them using the Log-Likelihood Ratio. We have also used several filters to prune the final list of discourse connectives: Word-Alignment, POS Patterns and Parse Tree Filters. We have achieved the best result in term of average precision with the Parse Tree Filter. Our analysis shows that the size of discourse connectives affects the quality of the filters. We also found that 31 candidates that labeled as non discourse connective, are indeed correct discourse connectives, yet are not covered in the LEXCONN V1.0 lexicon.

Our analysis also shows an important weakness of discourse annotation projection techniques based on statistical word-alignment models. Indeed a comparison between the Word-Alignment Filter and the the Parse Tree Filter shows that the longer French discourse connectives are, the less efficient statistical word-alignment models are at aligning the connectives. Hence, discourse annotation projection techniques based on solely statistical word-alignment models may not be efficient in projecting discourse annotations on long discourse connectives.

This chapter concludes our analysis of discourse annotation projecting. In the next chapter, we wrap up the thesis and summarize our findings. Then, we present different research avenues to extend our work.

Chapter 7

Conclusion and Future Work

7.1 Summary of the Thesis

Currently, building discourse resources is a time-consuming task and requires human expert annotators. Therefore, many languages suffer from lack of discourse resources. To address this problem, in this thesis, we propose an approach to automatically induce initial discourse resources from parallel texts based on available discourse resources for English.

In Chapter 2, we first defined the two target discourse resources that we want to induce from parallel texts. More specifically, we described 1) discourse annotated corpora and 2) lexicons of discourse connectives. Chapter 2 also listed the discourse resources currently available in the research community for different languages.

Next, in Chapter 3, we explained the development of the *CLaC DC Disambiguator* which we extensively used in our approach to annotate English discourse connectives. When trained on Sections 2–21 of the PDTB, the *CLaC DC Disambiguator* can disambiguate the discourse-usage of English discourse connectives with an F1-score of 90.8% and label their discourse relations with an F1-score of 79.7%. To estimate the performance of the *CLaC DC Disambiguator* on texts with different domains, we tested it on the CoNLL 2015/2016 blind test set. Our experiments show that the F1-scores drop from 90.8% to 88.1% and from 79.7% to 74.3% in labeling discourse-usage and discourse relations of English discourse connectives respectively.

Using the *CLaC DC Disambiguator*, we induced our first discourse resource in Chapter 4. To build a discourse annotated corpus for French, we used the *CLaC DC Disambiguator* to annotate English discourse connectives in parallel texts and aligned them to their counterpart French translations using statistical word-alignment models. We showed that statistical word-alignment models may produce noisy alignments when discourse relations are changed from explicit to implicit ones during the translation. To address this problem, we used a word-alignment model based on the intersection between direct and inverse word-alignment models. Our approach is able to identify 65% of the noisy word-alignments.

By using statistical word-alignment models to align words in parallel texts, we induced the Europarl ConcoDisco corpora where English discourse connectives are aligned to French discourse connectives. From the French side of the Europarl ConcoDisco corpora, we have created the FrConcoDisco corpora, the first PDTB-style discourse annotated corpora. We have evaluated both extrinsically and intrinsically the FrConcoDisco corpora and intrinsically showed that FrConcoDisco-*Intersection* contains the most accurate annotations at the expense recall. On the other hand FrConcoDisco-*Naive-grow-diag* contains more but less accurate annotations.

In Chapter 5 and Chapter 6, we showed how a lexicon of discourse connectives can be extracted from parallel texts. First, we developed an approach to map discourse relations to discourse connectives in Chapter 5. As a result of this approach we built *ConcoLeDisCo*, the first lexicon of French discourse connectives mapped to their PDTB discourse relations. Next, in Chapter 6 we proposed a novel approach to induce a list of French discourse connectives.

7.2 Main Findings and Contributions of the Thesis

Our contributions can be divided into two categories 1) practical contributions and 2) theoretical contributions.

7.2.1 Practical Contributions

We have **developed the *CLaC DC Disambiguator*** (see Chapter 3). We trained the *CLaC DC Disambiguator* on the FDTB to disambiguate French discourse connectives with an F1-score of

0.766. To best of our knowledge, this model is the only publicly available tool for the disambiguation of French discourse connectives.

We mined the Europarl corpus to build two types of discourse resources:

(1) We extracted bilingual and monolingual discourse annotated corpora (see Chapter 3):

    (a) **The Europarl ConcoDisco corpora:** In these corpora, around 1 million occurrences of French discourse connectives are aligned to their English translations and the English discourse connectives are annotated with the PDTB discourse relations that they convey. These corpora are valuable resource for corpus studies on how explicit discourse relations are affected by the translation process.

    (b) **The FrConcoDisco corpora:** The FrConcoDisco are extracted from the French side of the Europarl ConcoDisco corpora. To the best of our knowledge, these corpora are the first PDTB-style discourse annotated corpora for French.

(2) We have also built the *ConcoLeDisCo* lexicon (see Chapter 6). Again, to our knowledge, *ConcoLeDisCo* is the first lexicon of French discourse connectives where connectives are mapped to PDTB discourse relations.

7.2.2 Theoretical Contributions

We proposed two novel approaches in this thesis:

(1) We have proposed a novel approach based on the intersection statistical word-alignment models to identify unsupported annotations when projecting discourse relations (see Chapter 4). Our approach can automatically identify 65% of unsupported projected annotations. To our knowledge, our work is the first that systematically addresses unsupported annotations. This approach helped us to refine the naive method of discourse annotation projection. In particular, filtering unsupported annotations from projected annotations improves the F1-score of *CLaC DC Disambiguator* trained on these annotations by 15%.

(2) We have also proposed a novel approach for annotation projection (see Chapter 6). This approach is based on sentence alignments followed by the use of statistical tests to mine the

sentence aligned parallel corpus without using any statistical word-alignment models. Our results show that this approach is more robust to longer French discourse connectives than approaches based on statistical word-alignment models.

The above contributions have been disseminated in (Laali and Kosseim, 2014; Laali et al., 2015, 2016; Laali and Kosseim, 2016, 2017a,b).

7.3 Directions for Future Research

We believe our work can be expanded in at least three main directions:

(1) Improving discourse annotation projection.

(2) Developing a low-cost manual evaluation of the induced discourse resources.

(3) Exploring the use of the Europarl ConcoDisco corpora in other domains.

We will discuss each direction in more detail in the following sections.

7.3.1 Improving Discourse Annotation Projection

Our approach to discourse annotation projection can be extended in several ways.

First, our approach for projecting the discourse relations signaled by discourse connectives (see Chapter 4) can be extended so that it also projects the annotations of discourse arguments or the annotation of implicit discourse relations. To project the annotations of discourse arguments, we could also use an approach based on statistical word-alignment models to locate the most likely translation of each discourse argument in the target language and mark them as the discourse arguments of the identified relations. This is an interesting extension because recent work in the automatic identification of discourse arguments (Xue et al., 2016) has reached performance levels that made them usable as downstream applications. Because of recent advances in the development of parsers for implicit relations (e.g. (Wang et al., 2017)), it is now possible to consider projecting implicit discourse relations as well. As with explicit relations, we can assume that implicit discourse relations are preserved during the translation. Using a discourse parser for implicit relations (e.g. Wang et al.

(2017)), we can first tag such relations in a source language, then using machine translation systems, we can identify the best translation of the discourse arguments in the target language. Finally, we can project the same discourse relation between the translation of discourse arguments.

Another promising line of research would be to improve the quality of discourse annotation projection using deep-learning techniques. In this thesis, to project discourse annotations, we 1) developed the *CLaC DC Disambiguator* to annotate English discourse connectives and 2) used statistical word-alignment models to align English and French words. Both of these two components can benefit from deep-learning techniques. Deep-learning architectures such as Convolutions Neural Networks (CNN) and Recurrent Neural Networks (RNN) have recently been used to annotate implicit relations (Li et al., 2014a; Xue et al., 2016; Liu et al., 2016; Zhang et al., 2015; Braud and Denis, 2015). These results suggest that deep learning architectures can be more efficient than standard classifiers using hand-crafted features. Using similar neural architectures inside the *CLaC DC Disambiguator* may also lead to a better system to annotate English discourse connectives. Regarding the alignment of English and French words, currently Neural Machine Translation (NMT) systems create better and more natural translations than Statistical Machine Translation (SMT) systems that are based on statistical word-alignment models (Turovsky, 2016). NMT systems typically use an Attention Mechanism (Bahdanau et al., 2015) which creates alignments between words. As NMT systems typically perform better than SMT systems, they may also generate more accurate word alignments.

A third line of research would be to investigate the use of a bootsrapping approach. As shown in Chapter 3, some French discourse connectives are easier to disambiguate than their English counterparts. This motivates a bootstrapping extension to our approach to induce a classifier to annotate French discourse connectives. In our work, we used the *CLaC DC Disambiguator* trained on the PDTB to annotated English discourse connectives, then projected these annotations onto French discourse connectives and finally trained the *CLaC DC Disambiguator* on the induced corpus to annotate French discourse connectives. We could also do the reverse. More specifically, we could use the *CLaC DC Disambiguator* trained on the induced corpus and re-train it to annotate English discourse connectives, hence developing a bootstrapping extension of our approach.

To reduce error propagation through our pipeline of discourse annotation projection, as a fourth line of future work, we could experiment with jointly training the *CLaC DC Disambiguator* for English and French discourse connectives at the same time. To do so, we would need to define a loss function and an optimization mechanism to minimize this loss. The loss function could be defined as a linear combination of the number of incorrect relations identified by the English model on a manually annotated corpus (e.g. the PDTB) and the number of disagreements between the English and the French models on the discourse relations of discourse connectives aligned to each other. To minimize this loss function we could use stochastic gradient decent optimization techniques such as Momentum Optimizer (Sutskever et al., 2013). To use such techniques, it would be necessary to back-propagate through the whole pipeline which can be achieved if we use neural network architectures for the *CLaC DC Disambiguator* and word-alignments (e.g. using an Attention Mechanism).

Finally, although we used the French language in our experiments, our methodology could be applied to other languages. As indicated in Section 1.2, our approach makes no assumption about the target language except the availability of a parallel corpus with another language for which a discourse parser exists; hence the approach is easy to expand to other languages. It would be interesting to evaluate our approach with other languages and eventually induce new resources for other under-studied languages.

7.3.2 Developing a Low-Cost Manual Evaluation of the Induced Discourse Resources

The results of our work can be used to improve the development of French discourse resources such as LEXCONN (Roze et al., 2012) or the FDTB (Danlos et al., 2015). To do so, it is important to manually evaluate the discourse relations in the Europarl ConcoDisco corpora and/or *ConcoLeDisCo*. This could be done using human expert annotators.

However, to avoid the inherent cost of using human expert annotators, we can use crowdsourcing by designing linguistic tests that native speakers are capable to perform. In Chapter 4, we defined such a test, the Translatable Test, inspired by the Substitutability Test of Knott (1996). Cartoni et al. (2013) proposed a novel approach to generate more reliable annotations of discourse connectives by using the translation of discourse connectives. A combination of this approach and the Translatable Test can lead to a novel method to annotate the relation of discourse connectives using crowd-sourcing.

Another approach to evaluate our resources using crowd-sourcing is to develop a set of linguistic tests for discourse connectives that a native speaker can perform, while the answers to these tests give enough information to assign a relation to discourse connectives. For example, Zufferey and Degand (2014) suggested two simple linguistic tests to differentiate *COMPARISON.Concession* and *COMPARISON.Contrast* and to disambiguate *pragmatic* discourse relations from *non-pragmatic* discourse relations. Another example is the Substitutability Test proposed by Knott (1996). We believe that these tests can also be run by crowd-sourcing.

7.3.3 Exploring the Use of the the Europarl ConcoDisco Corpora in Other Domains

In this thesis, we mainly used the Europarl ConcoDisco corpora to induce the FrConcoDisco corpora and the *ConcoLeDisCo*lexicon. We also used the FrConcoDisco corpora to train the *CLaC DC Disambiguator* for French discourse connectives. However, the Europarl ConcoDisco corpora can be used in investigate cross-lingual discourse studies, such as machine translation and cognitive studies (see Section 2.2.2 and Section 2.2.3 for more detail).

Our approach presented in Chapter 4 can be also used to automatically identify and annotate implicit discourse relations within English texts. More specifically, our approach is able to find 65% of parallel sentences where French candidate discourse connectives are dropped in the English translation. If we were able to annotate these candidate discourse connectives (for example, see Chapter 3 for how the usage of French discourse connectives can be disambiguated), then it would be possible to build a dataset of implicit discourse relations by extracting parallel sentences where French discourse connectives are dropped during the translation process, hence, an explicit discourse relation is expressed implicitly in the English sentence (for example, see (Ex. 30) or (Ex. 33)). Extracting these implicit relations would allow us to automatically build a large-scale corpus for implicit discourse relations.

This thesis is an exploration towards the development of low-cost approaches to build two types of discourse resources: 1) discourse annotated corpora and 2) lexicons of discourse connectives.

We hope that our work has shown the effectiveness of annotation projection as an approach to build these two resources using parallel texts.

Bibliography

Anne Abeillé, Lionel Clément, and Francois Toussenel. Building a treebank for French. In *Proceedings of 2nd International Conference on Language Resources and Evaluation (LREC 2000)*, pages 165–187, Athens, Greece, May 2000.

Stergos D. Afantenos, Nicholas Asher, Farah Benamara, Myriam Bras, Cécile Fabre, Mai Ho-Dac, Anne Le Draoulec, Philippe Muller, Marie-Paule Péry-Woodley, and Laurent Prévot. An empirical resource for discovering cognitive principles of discourse organisation: The ANNODIS corpus. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 2727–2734, Istanbul, Turkey, May 2012.

Amal Al-Saif and Katja Markert. The Leeds Arabic Discourse Treebank: Annotating discourse connectives for arabic. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2010)*, pages 2046–2053, Valletta, Malta, May 2010.

Laura Alonso Alemany, Irene Castellón Masalles, and Llu´s Padró Cirera. Lexicón computacional de marcadores del discurso. *Procesamiento del Lenguaje Natural*, 29:239–246, 2002.

Amal Alsaif and Katja Markert. Modelling discourse relations for Arabic. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, pages 736–747, Edinburgh, Scotland, UK, July 2011.

Nicholas Asher. *Reference to abstract objects in discourse*. Springer, 1993.

Nicholas Asher and Alex Lascarides. *Logics of conversation*. Cambridge University Press, 2003.

Fatemeh Torabi Asr and Vera Demberg. Implicitness of discourse relations. In *Proceedings of the*

24th International Conference on Computational Linguistics: Technical Papers (COLING 2012), pages 2669–2684, Mumbai, December 2012.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Third International Conference on Learning Representations (ICLR 2015)*, San Diego, California, May 2015.

Luisa Bentivogli and Emanuele Pianta. Exploiting parallel texts in the creation of multilingual semantically annotated resources: The MultiSemCor Corpus. *Natural Language Engineering*, 11 (3):247–261, 2005.

Steven Bethard, Philip Ogren, and Lee Becker. ClearTK 2.0: Design patterns for machine learning in UIMA. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 3289–3293, Reykjavik, Iceland, May 2014.

Chloé Braud and Pascal Denis. Comparing word representations for implicit Discourse relation classification. In *Empirical Methods in Natural Language Processing (EMNLP 2015)*, 2015.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19 (2):263–311, 1993.

Harry Bunt and Rashmi Prasad. ISO DR-Core (ISO 24617-8): Core concepts for the annotation of discourse relations. In *Proceedings 12th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-12)*, pages 45–54, 2016.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar F. Zaidan. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 17–53, Uppsala, Sweden, July 2010.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. Findings of the 2012 Workshop on Statistical Machine Translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada, June 2012.

Marie Candito and Djamé Seddah. Effectively long-distance dependencies in French: annotation and parsing evaluation. In *Proceedings of the 11th International Workshop on Treebanks and Linguistic Theories (TLT 11)*, pages 61–72, Lisbon, Portugal, November 2012.

Lynn Carlson and Daniel Marcu. Discourse tagging reference manual. *ISI Technical Report ISI-TR-545*, 54, 2001.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of Second SIGdial Workshop on Discourse and Dialogue (SIGDIAL 2001)*, pages 1–10, Aalborg, Denmark, September 2001.

Bruno Cartoni and Thomas Meyer. Extracting directional and comparable corpora from a multilingual corpus for translation studies. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*, pages 2132–2137, Istanbul, Turkey, May 2012.

Bruno Cartoni, Sandrine Zufferey, and Thomas Meyer. Annotating the meaning of discourse connectives by looking at their translation: The translation-spotting technique. *Dialogue & Discourse*, 4(2):65–86, 2013.

Junyoung Chung, Kyunghyun Cho, and Yoshua Bengio. A character-level decoder without explicit segmentation for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pages 1693–1703, Berlin, Germany, August 2016.

Iria Da Cunha, Juan-Manuel Torres-Moreno, and Gerardo Sierra. On the development of the RST Spanish Treebank. In *Proceedings of the 5th Linguistic Annotation Workshop (ACL HLT 2011)*, pages 1–10, Oregon, USA, June 2011.

L. Danlos, M. Colinet, and J. Steinlin. FDTB1: repérage des connecteurs de discours en corpus. In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2015)*, pages 350–356, Caen, France, June 2015.

Laurence Danlos, Diégo Antolinos-Basso, Chloé Braud, and Charlotte Roze. Vers le FDTB: French Discourse Tree Bank. In *Actes de la 19ème conférence sur le Traitement Automatique des Langues Naturelles (TALN 2012)*, volume 2, pages 471–478, Grenoble, France, June 2012.

Debopam Das and Maite Taboada. Explicit and implicit coherence relations: A corpus study. In *Proceedings of the 2013 Annual Conference of the Canadian Linguistic Association*, 2013.

Vera Demberg, Fatemeh Torabi Asr, and Merel Scholman. How consistent are our discourse annotations? Insights from mapping RST-DT and PDTB annotations. *arXiv preprint arXiv:1704.08893*, 2017.

Heiner Drenhaus, Vera Demberg, Judith Köhne, and Francesca Delogu. Incremental and predictive discourse processing based on causal and concessive discourse markers: ERP studies on German and English. In *Proceedings of the 36th Annual Cognitive Science Conference (COGSCI 2014)*, Quebec City, Canada, July 2014.

Stefan Evert. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. PhD dissertation, Institut für Maschinelle Sprachverarbeitung, University of Stuttgart, 2004.

Syeed Ibn Faiz and Robert E. Mercer. Identifying explicit discourse connectives in Text. In *Proceedings of the 26th Canadian Conference on Artificial Intelligence (AI 2013)*, pages 64–76. LNAI 7884, Springer, May 2013.

Benamara Zitoune Farah, Nicholas Asher, Yvette Yannick Mathieu, Vladimir Popescu, and Baptiste Chardon. Evaluation in discourse: A corpus-based study. *Dialogue & Discourse*, 7(1), 2016.

David Ferrucci and Adam Lally. UIMA: An architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3-4):327–348, 2004.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2016)*, pages 866–875, San Diego, California, June 2016.

Deen G. Freelon. ReCal: Intercoder reliability calculation as a web service. *International Journal of Internet Science*, 5(1):20–33, 2010.

Qin Gao and Stephan Vogel. Parallel Implementations of Word Alignment Tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57, Columbus, OH, USA, 2008.

Spence Green, Marie-Catherine de Marneffe, John Bauer, and Christopher D. Manning. Multiword Expression Identification with Tree Substitution Grammars: A Parsing *tour de force* with French. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, pages 725–735, Edinburgh, Scotland, UK, 2011.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

Michael Halliday. *An introduction to functional grammar*. Routledge, 1985.

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 2016)*, pages 1424–1433, Berlin, Germany, August 2016.

Jerry R Hobbs. *Literature and cognition*. Number 21 in CSLI Lecture Notes. Center for the Study of Language (CSLI), 1990.

Jet Hoek and Sandrine Zufferey. Factors influencing the implicitation of discourse relations across languages. In *Proceedings 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-11)*, pages 39–45. TiCC, Tilburg Center for Cognition and Communication, 2015.

Eduard H Hovy. Parsimonious and profligate approaches to the question of discourse structure relations. In *Proceedings of the Fifth International Workshop on Natural Language Generation*, Dawson, Pennsylvania, June 1990.

Eduard H. Hovy. The multifunctionality of discourse markers. In *Proceedings of the Workshop on Discourse Markers*, pages 1–12, Egmond-aan-Zee, The Netherlands, January 1995.

Stig Johansson. Contrastive linguistics and corpora. Reports from the project languages in contrast, University of Oslo, 2000.

Stig Johansson. *Seeing through Multilingual Corpora: On the Use of Corpora in Contrastive Studies*. John Benjamins Publishing, 2007.

A. K. Joshi and Y. Schabes. Tree-adjoining grammars. *Handbook of Formal Languages*, 3:69–124, 1997.

Iskandar Keskes. *Discourse analysis of arabic documents and application to automatic summarization*. PhD dissertation, Université Toulouse III-Paul Sabatier, 2015.

Adam Kilgarriff, Vojtěch Kovář, Simon Krek, Irena Srdanović, and Carole Tiberius. A quantitative evaluation of word sketches. In *Proceedings of the 14th EURALEX International Congress*, pages 372–379, Leeuwarden, The Netherlands, July 2010.

Richard Kittredge, Tanya Korelsky, and Owen Rambow. On the need for domain communication knowledge. *Computational Intelligence*, 7(4):305–314, 1991.

Alistair Knott. *A data-driven methodology for motivating a set of coherence relations*. PhD dissertation, University of Edinburgh, Computer Science Department, 1996.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit*, volume 5, pages 79–86, Phuket, Thailand, September 2005.

Philipp Koehn. *Statistical Machine Translation*. Cambridge University Press, 2009. ISBN 0-521-87415-7.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions (ACL 2007)*, pages 177–180, Prague, June 2007.

Klaus Krippendorff. *Content analysis: An introduction to its methodology*. Sage, 2004.

Majid Laali and Leila Kosseim. Inducing discourse connectives from parallel texts. In *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers (COLING 2014)*, pages 610–619, Dublin, Ireland, August 2014.

Majid Laali and Leila Kosseim. Automatic disambiguation of French discourse connectives. *International Journal of Computational Linguistics and Applications (IJCLA)*, 7(1):11–30, 2016.

Majid Laali and Leila Kosseim. Automatic mapping of French discourse connectives to PDTB discourse relations. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL 2017)*, pages 1–6, Saarbrücken, Germany, August 2017a.

Majid Laali and Leila Kosseim. Improving discourse relation projection to build discourse annotated corpora. In *Proceedings of the 11th biennial Recent Advances in Natural Language Processing (RANLP 2017)*, pages 407–416, Varna, Bulgaria, September 2017b.

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

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 (CoNLL 2016)*, pages 92–99, Berlin, Germany, August 2016.

Alex Lascarides and Nicholas Asher. Segmented discourse representation theory: Dynamic semantics with discourse structure. In *Computing Meaning: Volume 3*, pages 87–124. Kluwer Academic Publishers, 2007.

Jiwei Li, Rumeng Li, and Eduard H. Hovy. Recursive deep models for discourse parsing. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pages 2061–2069, Doha, Qatar, October 2014a.

Junyi Jessy Li, Marine Carpuat, and Ani Nenkova. Assessing the discourse factors that influence the quality of machine translation. In *Proceedings of the 52nd Annual Meeting of the Association* *for Computational Linguistics (Short Papers) (ACL 2014)*, pages 283–288, Baltimore, Maryland, USA, June 2014b.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. A PDTB-styled end-to-end discourse parser. Technical Report TRB8/10, School of Computing, National University of Singapore, August 2010.

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. doi: 10.1017/S1351324912000307.

Yang Liu, Sujian Li, Xiaodong Zhang, and Zhifang Sui. Implicit discourse relation classification via multi-task neural networks. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

Minh-Thang Luong and Christopher D. Manning. Achieving open vocabulary neural machine translation with hybrid word-character models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pages 1054–1063, Berlin, Germany, August 2016.

Elisabeth Maier and Eduard Hovy. Organising discourse structure relations using metafunctions. *New concepts in natural language generation*, pages 69–86, 1993.

William C. Mann and Sandra A. Thompson. Rhetorical structure theory: A framework for the analysis of texts. *IPrA Papers in Pragmatics*, 1:79–105, 1987.

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

William C. Mann, Christian M.I.M. Matthiessen, and Sandra A. Thompson. Rhetorical structure theory and text analysis. Technical report, University of Southern California, Amsterdam and Philadelphia: John Benjamins, 1992.

Raghavan-Prabhakar Manning, Christopher D. and Hinrich Schutze. *Introduction to information retrieval*, volume 1. Cambridge University Press Cambridge, 2008.

Daniel Marcu. The rhetorical parsing of unrestricted texts: A surface-based approach. *Computational Linguistics*, 26(3):395–448, 2000.

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.

Thomas Meyer. Disambiguating temporal–contrastive discourse connectives for machine translation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2011)*, pages 46–51, Portland, OR, USA, June 2011.

Thomas Meyer and Lucie Poláková. Machine translation with many manually labeled discourse connectives. In *Proceedings of the 1st DiscoMT Workshop at the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, August 2013.

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

Keith K. Millis, Jonathan M. Golding, and Gregory Barker. Causal connectives increase inference generation. *Discourse Processes*, 20(1):29–49, 1995.

Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. Annotating discourse connectives and their arguments. In *Proceedings of the HLT/NAACL Workshop on Frontiers in Corpus Annotation*, pages 9–16, 2004.

Marcus Mitchell, Beatrice Santorini, and Mary Ann Marcinkiewicz. Treebank-2 (LDC95T7). *Web Download. Linguistic Data Consortium*, 1995.

Lucie Mladová, Sarka Zikanova, and Eva Hajicová. From sentence to discourse: Building an annotation scheme for discourse based on prague dependency treebank. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, pages 28–30, Morocco, Marrakech, May 2008.

Philippe Muller, Marianne Vergez-Couret, Laurent Prévot, Nicholas Asher, Benamara Farah, Myriam Bras, Anne Le Draoulec, and Laure Vieu. Manuel d'annotation en relations de discours du projet annodis. *Carnets de grammaire*, 21:34, 2012.

John D. Murray. Logical connectives and local coherence. In Robert F. Lorch and Edward J O'Brien, editors, *Sources of Coherence in Reading*, pages 107–126. Hillsdale, N.J. : L. Erlbaum Associates, 1995.

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

Jiř´ M´rovsky, Pavl´na Synková, Magdaléna Rysová, and Lucie Poláková. Designing CzeDLex–A lexicon of Czech discourse connectives. In *Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation*, pages 449–457, Seoul, Korea, October 2016.

Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

Stephan Oepen, Jonathon Read, Tatjana Scheffler, Uladzimir Sidarenka, Manfed Stede, Erik Velldal, and Lilja Ovrelid. OPT: Oslo—Potsdam—Teesside pipelining rules, rankers, and classifier ensembles for Shallow Discourse Parsing. In *Proceedings of the 20th Conference on Computational Natural Language Learning (CoNLL 2016)*, pages 20–26, Berlin, Germany, August 2016.

Umangi Oza, Rashmi Prasad, Sudheer Kolachina, Dipti Misra Sharma, and Aravind Joshi. The Hindi discourse relation bank. In *Proceedings of the Third Linguistic Annotation Workshop*, pages 158–161, Suntec, Singapore, 2009.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL 2002)*, pages 311–318, Philadelphia, Pennsylvania, July 2002.

Thiago Alexandre Salgueiro Pardo and Eloize Rossi Marques Seno. Rhetalho: um corpus de referência anotado retoricamente. In *Anais do V Encontro de Corpora*, pages 24–25, São Carlos, November 2005.

Thiago Alexandre Salgueiro Pardo, Maria das Gracas Volpe Nunes, and Lucia Helena Machado Rino. Dizer: An automatic discourse analyzer for brazilian portuguese. *Lecture Notes in Artificial Intelligence*, 3171:224–234, 2008.

P. Pecina. Lexical association measures and collocation extraction. *Language Resources and Evaluation*, 44(1):137–158, 2010.

Ted Pedersen, Satanjeev Banerjee, Bridget McInnes, Saiyam Kohli, Mahesh Joshi, and Ying Liu. The Ngram Statistics Package (Text:: NSP)-A flexible tool for identifying ngrams, collocations, and word associations. In *Workshop on Multiword Expression: From Parsing and Generation to the Real World (MWE 2011)*, pages 131–133, Portland, OR, USA, June 2011.

Slav Petrov and Dan Klein. Improved Inference for Unlexicalized Parsing. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2007)*, pages 404–411, Rochester, NY, April 2007.

Emily Pitler and Ani Nenkova. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP 2009)*, pages 13–16, Suntec, Singapore, August 2009.

Emily Pitler, Annie Louis, and Ani Nenkova. Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP 2009)*, pages 683–691, Suntec, Singapore, August 2009.

Lucie Poláková, Jiř´ M´rovsky, Anna Nedoluzhko, Pavl´na J´nová, Šárka Zikánová, and Eva Hajičová. Introducing the Prague Discourse Treebank 1.0. In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJNLP 2013)*, pages 91–99, Nagoya, Japan, October 2013.

Andrei Popescu-Belis, Thomas Meyer, Jeevanthi Liyanapathirana, Bruno Cartoni, and Sandrine Zufferey. Discourse-level annotation over europarl for machine translation: Connectives and pronouns. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 23–25, Istanbul, Turkey, May 2012.

Rashmi Prasad, Eleni Miltsakaki, Aravind Joshi, and Bonnie Webber. Annotation and data mining of the Penn Discourse Treebank. In *Proceedings of the 2004 ACL Workshop on Discourse Annotation*, pages 88–97, Barcelona, Spain, July 2004.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K. Joshi, and Bonnie L. Webber. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, pages 28–30, Marrakech, Morocco, May 2008a.

Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo, and Bonnie L. Webber. The Penn Discourse Treebank 2.0 annotation manual. Technical Report IRCS-08-01, Institute for Research in Cognitive Science, University of Pennsylvania, Philadelphia, USA, 2008b.

Rashmi Prasad, Aravind Joshi, and Bonnie Webber. Realization of discourse relations by other means: alternative lexicalizations. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters (COLING 2010)*, pages 1023–1031, Beijing, China, August 2010.

J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993. ISBN 1-55860-238-0.

Brian Reese, Julie Hunter, Nicholas Asher, Pascal Denis, and Jason Baldridge. Reference manual for the analysis and annotation of rhetorical structure (version 1.0). Technical report, Technical report. Austin: University of Texas, Departments of Linguistics and Philosophy. Available online: http://timeml. org/jamesp/annotation_manual.pdf, 2007.

Charlotte Roze, Laurence Danlos, and Philippe Muller. LEXCONN: A French lexicon of discourse connectives. *Discours [En ligne]*, 10, 2012. doi: 10.4000/discours.8645.

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

Carolina Scarton. *Document-Level Machine Translation Quality Estimation*. PhD thesis, University of Sheffield, 2016.

Tatjana Scheffler and Manfred Stede. Mapping PDTB-style connective annotation to RST-style discourse annotation. In *the 13th Conference on Natural Language Processing (KONVENS 2016)*, pages 242–247, Bochum, Germany, September 2016.

Violeta Seretan. *Syntax-Based Collocation Extraction*, volume 44. Springer-Verlag New York Inc, 2010.

Manfred Stede. The POTSDAM commentary corpus. In *Proceedings of the 2004 ACL Workshop on Discourse Annotation*, pages 96–102, Barcelona, Spain, July 2004.

Manfred Stede and Yulia Grishina. Anaphoricity in connectives: A case study on German. In *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016*, pages 41–46, San Diego, California, June 2016.

Manfred Stede and Arne Neumann. Potsdam Commentary Corpus 2.0: Annotation for discourse research. In *Proceedings of 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 925–929, Reykjavik, Iceland, May 2014.

Manfred Stede and Carla Umbach. DiMLex: A lexicon of discourse markers for text generation and understanding. In *Proceedings of the 17th International Conference on Computational Linguistics (COLING 1998)*, pages 1238–1242, Montreal, Canada, August 1998.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th International Conference on Machine Learning (ICML 2013)*, pages 1139–1147, Atlanta, Georgia, USA, June 2013.

Maite Taboada. Implicit and explicit coherence relations. In *Discourse, of Course*, pages 127–140. Amsterdam/Philadelphia: John Benjamins, 2009.

Maite Taboada and Maria de los Ángeles Gómez-González. Discourse markers and coherence relations: Comparison across markers, languages and modalities. *Linguistics and the Human Sciences*, 6(1-3):17–41, 2012.

Maite Taboada and William C. Mann. Rhetorical structure theory: Looking back and moving ahead. *Discourse studies*, 8(3):423–459, 2006.

Maite Taboada and Jan Renkema. *Discourse Relations Reference Corpus [Corpus]*. Simon Fraser University and Tilburg University, 2008.

Heike Telljohann, Erhard W. Hinrichs, Sandra Kübler, Heike Zinsmeister, and Kathrin Beck. Stylebook for the Tübingen treebank of written German (TüBa-D/Z). Technical report, Universität Tübingen, Tübingen, August 2006.

Jörg Tiedemann. News from OPUS-A collection of multilingual parallel corpora with tools and interfaces. In *Proceedings of Recent Advances in Natural Language Processing (RANLP 2009)*, volume 5, pages 237–248, Borovets, Bulgaria, September 2009.

Jörg Tiedemann. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*, pages 2214–2218, Istanbul, Turkey, May 2012.

Jörg Tiedemann. Improving the cross-lingual projection of syntactic dependencies. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 191–199, Vilnius, Lithuania, May 2015.

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-Volume 1 (NAACL/HLT)*, pages 173–180, 2003.

Barak Turovsky. Found in translation: More accurate, fluent sentences in Google Translate, November 2016.

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

Yizhong Wang, Sujian Li, and Houfeng Wang. A two-stage parsing method for text-level discourse analysis. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017) (Volume 2: Short Papers)*, volume 2, pages 184–188, 2017.

Bonnie Webber and Aravind Joshi. Anchoring a lexicalized tree-adjoining grammar for discourse. In *COLING/ACL Workshop on Discourse Relations and Discourse Markers*, pages 86–92, 1998.

Bonnie Webber and Aravind Joshi. Discourse structure and computation: past, present and future. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012) Special Workshop on Rediscovering 50 Years of Discoveries*, pages 42–54, 2012.

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

Florian Wolf and Edward Gibson. Representing discourse coherence: A corpus-based study. *Computational Linguistics*, 31(2):249–287, 2005.

Nianwen Xue. Annotating discourse connectives in the Chinese Treebank. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 84–91, 2005.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Rashmi PrasadO Christopher Bryant, and Attapol T. Rutherford. The CoNLL-2015 shared task on shallow discourse parsing. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task (CoNLL 2015)*, pages 1–16, Beijing, China, July 2015.

Nianwen Xue, Hwee Tou Ng, Attapol Rutherford, Bonnie Webber, Chuan Wang, and Hongmin Wang. CoNLL 2016 Shared Task on multilingual shallow discourse parsing. In *Proceedings of the 20th Conference on Computational Natural Language Learning (CoNLL 2016)*, pages 1–19, Berlin, Germany, August 2016.

David Yarowsky, Grace Ngai, and Richard Wicentowski. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research (HLT 2001)*, pages 1–8, San Diego, California, March 2001.

Frances Yung, Kevin Duh, Taku Komura, and Yuji Matsumoto. A psycholinguistic model for the marking of discourse relations. *Dialogue & Discourse*, 8(1):106–131, 2017.

Deniz Zeyrek, Isn Demirsahin, Aysğ Sevdik-Call, Hale Ögel Balaban, İhsan Yalcnkaya, and Ümit Deniz Turan. The annotation scheme of the Turkish Discourse Bank and an evaluation of inconsistent annotations. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 282–289, Uppsala, Sweden, July 2010.

Biao Zhang, Jinsong Su, Deyi Xiong, Yaojie Lu, Hong Duan, and Junfeng Yao. Shallow convolutional neural network for implicit discourse relation recognition. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*, pages 2230–2235, Lisbon, Portugal, September 2015.

Lanjun Zhou, Wei Gao, Bin Li, Zhong Wei, and Kam-fai Wong. Cross-lingual identification of Ambiguous discourse Connectives for resource-poor Language. In *Proceedings of the 24th International Conference on Computational Linguistics: Technical Papers (COLING 2012)*, pages 1409–1418, Mumbai, December 2012.

Yuping Zhou and Nianwen Xue. PDTB-style discourse annotation of Chinese text. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, pages 69–77, Jeju, Republic of Korea, 2012.

Farah Benamara Zitoune and Maite Taboada. Mapping different rhetorical relation annotations: A proposal. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics (*SEM 2015)*, pages 147–152, Denver, Colorado, June 2015.

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

Sandrine Zufferey and Bruno Cartoni. English and French causal connectives in contrast. *Languages in Contrast*, 12(2):232–250, 2012.

Sandrine Zufferey and Liesbeth Degand. Representing the meaning of discourse connectives for multilingual purposes. *Corpus Linguistics and Linguistic Theory*, 10, 2014. doi: 10.1515/cllt-2013-0022.

Sandrine Zufferey and Pascal M. Gygax. The role of perspective shifts for processing and translating discourse relations. *Discourse Processes*, 53(7):532–555, 2015. doi: 10.1080/0163853X.2015.1062839.

Appendix A

Mapping PDTB Relations to RST Relations

In the PDTB, only surface discourse relations were annotated and nested discourse relations were not considered (see Chapter 2). This raises the question of how many relations have been ignored in this framework. The goal of this appendix is to address this question. Recall from Chapter 2, that the RST-DT (Carlson et al., 2001) annotates a portion of the corpus annotated by the PDTB (Prasad et al., 2008a). Hence, we can use this common corpus to address this question. We used Rhetorical Structure Theory (Mann and Thompson, 1987) as a reference framework and compared the PDTB (Prasad et al., 2008a) relations and RST relations annotated in the RST-DT (Carlson et al., 2001).

Before comparing their annotations of the PDTB and the RST-DT, we review the annotation schemas in these two corpora.

A.1 RST Annotation Schema

In Rhetorical Structure Theory (RST), to annotated discourse relations, the text is first segmented to non-overlapping clauses which are referred to Elementary Discourse Units (EDUs) Mann and Thompson (1988). To make this notion more precise for annotating the boundaries of EDUs, Carlson et al. (2001) excluded some clauses from EDUs. Specifically, he excluded:

(1) Clauses that are subjects or objects of a main verb.

(2) Clauses that are complements of a main verb.

See (Carlson et al., 2001) for more detail on how EDUs are formally defined.

In the next step, EDUs are connected to each other using discourse relations to build Complex Discourse Units (CDUs). This process continues by connecting EDUs and CDUs until all EDUs of the text are connected to each other and create a tree structure over the text.

Because of the tree-structure of RST, it is difficult to annotated discourse relations between embedded clauses and matrix clauses. To annotate these relations, in RST-DT, the *Same-Unit* relation has been defined which connects two text spans of a matrix clause. This allows the embedded clause to be connected to one of the text spans of the matrix clause while maintaining a tree structure for the discourse annotations. For example, in (Ex. 38), EDU1 and EDU3 can be considered as one clause that has been broken with the embedded structure (i.e. EDU2).

(Ex. 38) [But maintaining the key components of his strategy]$_{\text{EDU1}}$ [– a stable exchange rate and high levels of imports –]$_{\text{EDU2}}$ [will consume enormous amounts of foreign exchange.]$_{\text{EDU3}}$(wsj_0300)

As discussed in Section 2.1.1.1, RST (Mann and Thompson, 1987) also proposed the notion of a nucleus-satellite view on rhetorical relations, in which the span of the satellite text plays a subordinate role to the main nucleus text. The left hand side of Figure A.1 shows the RST tree of (Ex. 39), where the arrows are labelled with the name of the rhetorical relation and point to the nucleus span.

(Ex. 39) Kidder competitors aren't outwardly hostile to the firm, as many are to a tough competitor like Drexel Burnham Lambert Inc. that doesn't have Kidder's long history.
However, competitors say that Kidder's hiring binge involving executive-level staffers, some with multiple-year contract guarantees, could backfire unless there are results.

Using this annotation schema, 380 newspaper articles of the Wall Street Journal corpus (Mitchell et al., 1995) have been annotated in the RST Discourse Treebank (RST-DT; Carlson et al., 2001). In this corpus, a set of 78 discourse relations is used. The inter-annotator Kappa agreement on span

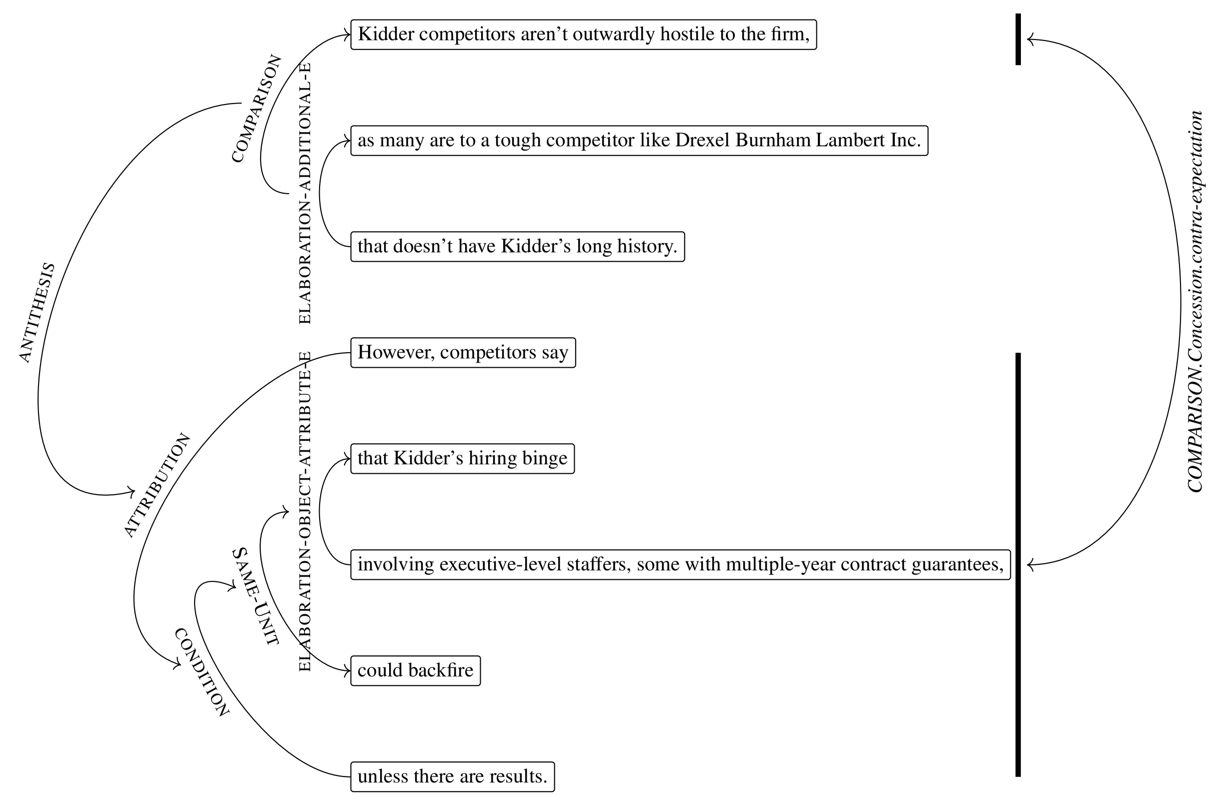


Figure A.1: Annotations of the RST-DT and the PDTB on the same text (taken from WSJ_0604). On the left, the RST annotations are shown. Each arrow points to the nucleus span and marked with an RST relation. On the right a PDTB discourse relation is shown. The two arguments of the relations are marked with the bold line and the relation is labeled withing the arrow detection of EDUs, detecting nuclear EDUs and assigning discourse relation was 90.0%, 85.6%, and 75.6%, respectively (Carlson et al., 2001). Note that, according to Krippendorff (2004), values of kappa > 0.8 reflect very high agreement, while values between 0.6 and 0.8 reflect good agreement.

A.2 PDTB Annotation Schema

In the PDTB, a different approach is used to annotate discourse relations. While, in RST, first texts are segmented (i.e. EDUs) and then discourse relations between these segments are annotated, in the PDTB, this process is done in the other direction: first the presence of discourse relations are identified and then, the texts are segmented.

As indicated in Section 2.1.1.3, in the PDTB, the presence of discourse relations were identified based on a set of 100 discourse connectives. Moreover, it is also assumed that there is an implicit discourse relation between each two consecutive sentences even if there is no explicit discourse connectives between them. Note that in the PDTB, implicit relations within sentences were not annotated. For example, the *Purpose* discourse relation implicitly signalled within (Ex. 40) between the *italic text* and <u>underlined text</u> were ignored in the PDTB.

(Ex. 40) *The Court of Appeals for the Federal Circuit was created in 1982* <u>to serve, among other things, as the court of last resort for most patent disputes.</u>

In the PDTB all discourse relations are binary relations between two text spans referred to as Arg1 and Arg2. To identify the text spans of Arg1 and Arg2, the PDTB follows the Minimality Principle (Prasad et al., 2008b, p. 14). According to this principle, the PDTB annotators should select only the required and sufficient clauses that are necessary for the interpretation of the discourse relations.

Now that we have summarized the annotation schema of both the PDTB and the RST-DT, let us now see how discourse relations of these two corpora can be mapped to each other.

A.3 Experiment

Three hundred fifty-nine (359) articles of the *Wall Street Journal* corpus (Mitchell et al., 1995) have been annotated in both the RST-DT and the PDTB[1].

A.3.1 Counting Relations

Table A.1 provides statistics of these articles. As discussed in Section A.1, the *Same-Unit* relations are not a discourse relation per se and were only defined to guarantee the tree structure of the annotations. Therefore, we excluded the 2,640 *Same-Unit* relations from the annotated relations in the RST-DT, which resulted in 17,861 ($20,501 - 2,640$) valid discourse relations.

Raw Statistics		**RST-DT Statistics**		**PDTB Statistics**	
# Words	166,047	# EDUs	20,860	# PDTB relations	6,781
# Paragraphs	4,103	# RST relations	20,501	# Explicit relations	3,031
		# Valid RST relations	17,861	# Non-explicit relations	3,750

Table A.1: Statistics of the annotations of the RST-DT and the PDTB on the 359 common articles of the *Wall Street Journal* corpus.

Based on these statistics, the proportion of the relations in the PDTB is 38.0% ($6,781/20,501$) of the number of relations in the RST-DT. Explicit relations consist of 44.7% ($3,031/6,781$) of the relations annotated in the PDTB. This shows that a large portion of discourse relations in the PDTB are explicit. If the explicit relations are compared against all valid RST-DT relations, the proportion of explicit relations is 17.0% ($3,031/17,861$).

A.3.2 Aligning PDTB to RST Discourse Relations

Counting relations, as done in Section A.3.1, assumes that PDTB relations are equivalent to RST relations. This is not the case. The PDTB and the RST-DT use different annotation schemas. In particular, the definition of the building block of discourse relations are different in these two frameworks. In RST-DT, the relations are annotated in a hierarchical tree structure; therefore, the

[1]The RST-DT contains 380 of articles of the *Wall Street Journal* corpus. However, because 21 of these articles were not annotated in the Penn Tree Bank (PTB) or they could not be converted to the format required in the PDTB (Prasad et al., 2008b, p.8), they were excluded from the PDTB. Hence the common corpus between the RST-DT and the PDTB includes 359 (380 - 21) articles.

relations in the higher levels of the tree structure cover larger text spans. This is not the case in the PDTB because of the Minimality Principle. Hence, even if the annotators of both schemas had the same interpretation of the text and the same relation in mind, they might select different text spans for the relation. Ideally, one should align the two resources and compare each relation one by one.

A.3.2.1 Alignment Method

To compare discourse relations between the PDTB and RST, we mapped each PDTB discourse relations to an RST discourse relation, provided that:

(1) The mapped RST relation should cover both Arg1 and Arg2 of the PDTB discourse relations. As discussed before, as a result of the Minimality Principle, PDTB annotators select the required and sufficient clauses. That means that if the same relation is also annotated in RST, it has to include at least the same text spans (i.e. Arg1 and Arg2).

(2) If Arg1 and/or Arg2 of the PDTB relation is covered by a descendant of the mapped RST relation, then all nodes in the path to the descendant child should be a Nucleus of a relation. In other words, by applying this constraint, we enforce the Strong Nuclearity hypothesis (Marcu, 2000), which states that if there is a relation between two text spans, the same relation should also hold between the nucleus of these two spans.

We used Algorithm 5 to create mappings with the above two constraints. This algorithm takes as input a list of discourse relations annotated in the PDTB and the RST-DT and returns mappings between the PDTB discourse relations and the RST relations. In this algorithm, for each PDTB relations and for each RST discourse unit (i.e. EDU and CDU), we find the smallest unit that covers Arg1 or Arg2 (Lines 3-5). Then, we compute the path from these two units to the root of the tree annotated in the RST-DT (Lines 6-7). Using these two paths, we compute the lowest common ancestor (lca) in the tree that covers both Arg1 and Arg2 (Line 8). Then, we check that the nodes after the immediate descendants of lca are all nuclei to ensure the Strong Nuclearity hypothesis

(Line 9). If this constraint holds, then we map the PDTB relation to lca.

Algorithm 5: Map-PDTB-RST-Relations

Input: $pdtbRelations$: PDTB relations.

Input: $rstUnits$: RST discourse units that connected to each other using RST relations.

Output: $mapping$: a mapping between PDTB relations and RST relations.

1 $mapping = \{\}$;

2 **foreach** $relation_{pdtb} \in pdtbRelations$ **do**

3 $\quad \{arg_1, arg_2\} \leftarrow relation_{pdtb}$;

4 $\quad arg_1^{rst} = GetSmallestRstUnitCovering(arg_1, rstUnits)$;

5 $\quad arg_2^{rst} = GetSmallestRstUnitCovering(arg_2, rstUnits)$;

6 $\quad path_1 = GetPathToRoot(arg_1^{rst}, rstUnits)$;

7 $\quad path_2 = GetPathToRoot(arg_2^{rst}, rstUnits)$;

8 $\quad lca = LowestCommonAncestor(path_1, path_2)$;

9 $\quad$ **if** $\Big(AllNucleus(path_1, lca)$ *And* $AllNucleus(path_2, lca + 1)\Big)$ *Or* $\Big(AllNucleus(path_1, lca + 1)$ *And* $AllNucleus(path_2, lca)\Big)$ **then**

10 $\quad\quad \{relation_{pdtb}, lca\} \rightarrow mapping$

11 $\quad$ **end**

12 **end**

A.3.2.2 Results and Analysis

Using Algorithm 5, we were able to map 77.4% of the PDTB relations to a relation in RST-DT. Figure A.1 shows a mapping that this algorithm has found between the *COMPARISON.Concession.contraexpectation* relation annotated in the PDTB and the ANTITHESIS relation annotated in the RST-DT.

To understand why some of the PDTB relations are not mapped to RST relations, we have manually analyzed a subset of these. In most cases, it seems the PDTB annotators did not interpret the same discourse structure as the RST annotators. For example, consider part of discourse structure of (Ex. 41) shown in Figure A.2.

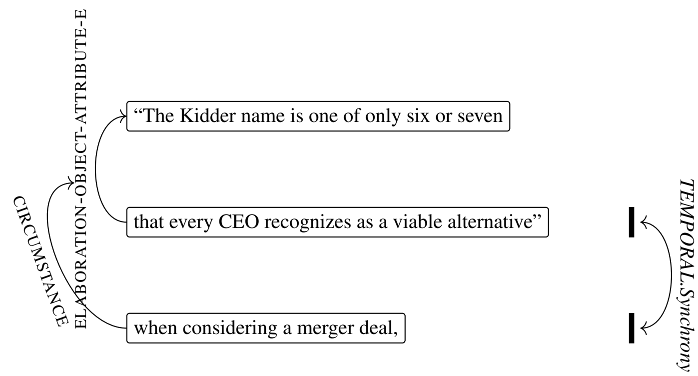


Figure A.2: The example shows that the PDTB annotation (right) is not consistent with the RST annotation (left).

(Ex. 41) *The firm's new head of mergers and acquisitions under Mr. Newquist, B.J. Megargel, talks of the opportunity to "rebuild a franchise" at Kidder. "The Kidder name is one of only six or seven that every CEO recognizes as a viable alternative" when considering a merger deal, he says.* (WSJ_0604)

As Figure A.2 shows, the PDTB annotation connects the *when* clause to the time that *CEO recognizes* but the RST annotation connects this clause to the *Kidder name*.

To understand why the number of relations in RST is higher than in the PDTB, we manually analyzed a random sample of RST relations that have not been mapped to PDTB relations. The most frequent RST relations that are not mapped to PDTB relations are ATTRIBUTION, ELABORATION-ADDITIONAL and LIST relations. These three relations make up 47.9% of the RST relations that are not mapped to PDTB relations. For example, Figure A.3 shows two RST relations that have not been annotated in the PDTB.

PDTB does not consider ATTRIBUTION relations as a discourse relation. Regarding ELABORATION-ADDITIONAL, according to our error analysis, most of the instances of this relation provide information to a named entity. Recall that the PDTB only annotates entity-based information that appears in two adjacent sentences, not within sentences. Hence these RST relations cannot find an equivalent the the PDTB.

Finally, recall from Section A.2 that implicit relations within sentences are not marked in the PDTB either. For example, Figure A.4 shows an ATTRIBUTION and a PURPOSE relations that have

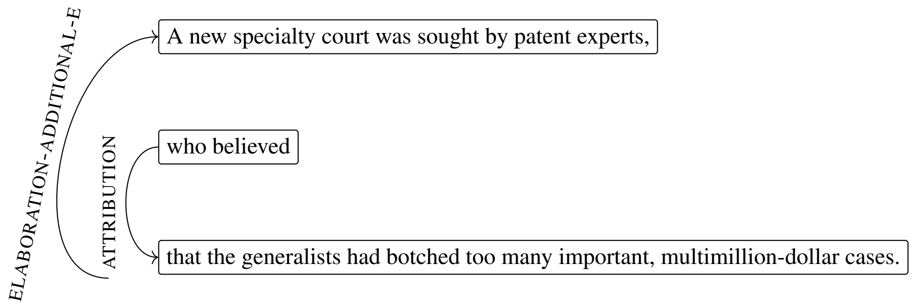


Figure A.3: An example of ATTRIBUTION and ELABORATION-ADDITIONAL in RST (taken from WSJ_0601).

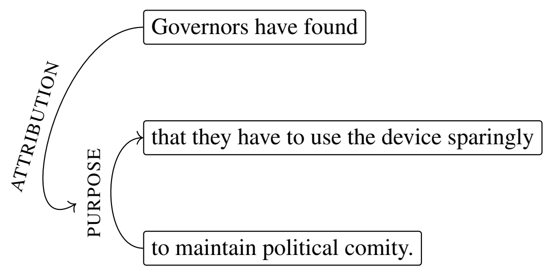


Figure A.4: An example of an implicit relation within a sentence that has not been annotated in the PDTB (taken from WSJ_0609).

not been annotated in the PDTB, but was annotated in the RST-DT.

A.4 Conclusion

The purpose of this appendix was to quantify and analyze the discourse relations that the PDTB does not consider compared to RST. To do this, we compared the annotations of the 359 common articles from the Wall Street Journal corpus that are annotated in both frameworks.

In Section A.3.1, we used a naive approach that considers that a PDTB relation is equivalent to an RST relation. By doing this, we determined that the PDTB relations account for 38.0% of the relations in the RST-DT.

On the other hand, in Section A.3.2 we tried to take into account the differences in the two frameworks and annotation schemes. By using the method presented in Algorithm 5, we attempted to align PDTB relations to their RST-DT counterpart. Using this method, we were able to map 77.4% of the PDTB relations to a relation in RST-DT. 47.9% of the RST relations that are not mapped to PDTB relations are ATTRIBUTION, ELABORATION-ADDITIONAL and LIST. Unlike

RST, the PDTB does not consider ATTRIBUTION relations, ELABORATION-ADDITIONAL related to named entities and implicit relations within sentences. Hence these RST relations do not have an equivalent in the PDTB.

Appendix B

Entropy of English Discourse Connectives Computed from the PDTB

English Connective	Entropy	Frequency
in contrast	1.00	22
besides	1.00	30
as a result	1.00	133
otherwise	1.00	41
instead	1.00	176
in particular	0.99	22
in the end	0.99	20
until	0.98	302
because	0.98	1062
as soon as	0.98	27
before	0.97	557
finally	0.97	73
nor	0.96	65
when	0.95	1215
once	0.94	199

English Connective	Entropy	Frequency
though	0.94	288
after	0.94	1080
ultimately	0.94	45
then	0.93	404
as long as	0.93	29
later	0.93	221
in addition	0.88	183
specifically	0.87	24
so	0.85	760
now that	0.84	26
previously	0.83	141
still	0.81	598
as if	0.81	20
since	0.80	563
if	0.80	1164
separately	0.77	80
but	0.74	3359
indeed	0.73	103
except	0.65	54
and	0.59	16386
as	0.57	3916
in fact	0.55	78
overall	0.52	78
for example	0.49	171
while	0.45	693
rather	0.44	154
so that	0.43	23

English Connective	Entropy	Frequency
therefore	0.43	23
nonetheless	0.41	24
thus	0.41	96
also	0.33	1503
for instance	0.33	81
unless	0.32	86
however	0.24	396
by contrast	0.24	26
nevertheless	0.21	30
as well	0.19	206
or	0.19	2486
further	0.16	257
meanwhile	0.14	158
plus	0.14	53
earlier	0.12	599
else	0.10	74
much as	0.10	148
moreover	0.09	84
next	0.07	629
although	0.04	268
for	0.00	8017
in turn	0.00	27
on the other hand	0.00	28
particularly	0.00	124
upon	0.00	40

Appendix C

Entropy of French Discourse Connectives Computed from the FDTB

French Connective	Entropy	Frequency
effectivement	1.00	27
sinon	1.00	27
d’ une part	1.00	28
alors	0.99	186
de même	0.99	52
auparavant	0.99	21
tout de même	0.99	21
aussi	0.97	533
surtout	0.97	167
d’ abord	0.97	102
tant que	0.96	21
par exemple	0.95	97
en attendant	0.95	30
de fait	0.95	22
maintenant	0.93	81

French Connective	Entropy	Frequency
bien qu’	0.89	23
puis	0.89	112
au lieu de	0.88	37
or	0.87	109
ensuite	0.87	75
bien que	0.87	38
ainsi	0.87	406
en particulier	0.84	56
finalement	0.82	58
et	0.81	8595
sans	0.80	599
si	0.77	502
bref	0.76	27
globalement	0.76	27
plutôt	0.75	83
comme	0.74	803
mais	0.73	1183
en fait	0.68	73
afin d’	0.67	34
après	0.67	584
faute de	0.66	29
du moins	0.65	24
au total	0.64	56
au contraire	0.63	44
de l’ autre	0.61	20
d’ autre part	0.60	82
pour que	0.59	42

French Connective	Entropy	Frequency
pour	0.58	3693
résultat	0.57	110
également	0.56	174
parce que	0.55	47
en tout cas	0.50	36
mais aussi	0.48	86
ou	0.48	1082
d' ailleurs	0.48	88
par ailleurs	0.47	50
de plus	0.46	233
du coup	0.44	22
quand	0.43	124
donc	0.41	293
a en	0.40	25
afin de	0.40	89
soit	0.39	409
enfin	0.39	172
autrement	0.38	27
bientôt	0.37	28
déjà	0.37	337
au moins	0.36	74
autant	0.36	104
en	0.35	6979
tout en	0.33	49
parce qu'	0.31	54
puisqu'	0.31	55
pourtant	0.30	130

French Connective	Entropy	Frequency
au moment où	0.29	40
comme pour	0.29	20
en réalité	0.29	20
parallèlement	0.26	23
puisque	0.25	72
jusqu’ à	0.24	153
ainsi qu’	0.24	26
encore	0.23	446
qu’ en	0.23	105
s’	0.22	1987
alors que	0.22	194
et qu’	0.22	57
et même	0.21	31
plus qu’	0.18	37
cependant	0.16	127
lorsqu’	0.14	49
depuis	0.14	733
quant à	0.13	54
avant de	0.13	55
en plus	0.13	111
en outre	0.12	59
même si	0.10	77
car	0.05	176
avant	0.04	221
que	0.02	2787
alors qu’	0.00	48
avant d’	0.00	23

French Connective	Entropy	Frequency
certes	0.00	81
d’ autant que	0.00	22
en effet	0.00	152
en revanche	0.00	124
lorsque	0.00	74
même	0.00	531
non plus	0.00	41
non seulement	0.00	47
notamment	0.00	299
néanmoins	0.00	40
pour autant	0.00	43
précisément	0.00	28
simplement	0.00	32
tandis que	0.00	84
toutefois	0.00	135
à	0	9880
à propos	0	35

Appendix D

ConcoLeDisCo Lexicon

French Connective	Discourse Relations/Probability	Freq
au même titre qu' au même titre que	TEMPORAL.Synchrony=0.0804	721
au moment où surtout au moment où	COMPARISON.Contrast=0.0066 CONTINGENCY.Cause.reason=0.0157 TEMPORAL.Asynchronous.succession=0.0054 TEMPORAL.Synchrony=0.2495	1655
au point qu' au point que	CONTINGENCY.Cause.reason=0.0052 CONTINGENCY.Cause.result=0.0052 EXPANSION.Conjunction=0.0104	192
à condition d' à condition de	CONTINGENCY.Condition=0.0782 TEMPORAL.Synchrony=0.0041	243
au point d' au point de	EXPANSION.Conjunction=0.0011	938
auparavant quelques heures auparavant	TEMPORAL.Asynchronous.predecence=0.0135	2365
au total	EXPANSION.Restatement=0.0049	609

French Connective	Discourse Relations/Probability	Freq
aussi longtemps qu' aussi longtemps que	COMPARISON.Contrast=0.0069 CONTINGENCY.Condition=0.2841 TEMPORAL.Asynchronous.predecence=0.0531	433
aussitôt qu' aussitôt que	TEMPORAL.Asynchronous.succession=0.1127 TEMPORAL.Synchrony=0.0282	71
aussitôt	EXPANSION.Conjunction=0.0107 TEMPORAL.Asynchronous.succession=0.0160 TEMPORAL.Synchrony=0.0053	187
pour une fois qu' pour une fois que	CONTINGENCY.Condition=0.1000	10
pour peu qu' pour peu que	CONTINGENCY.Condition=0.0808 TEMPORAL.Synchrony=0.0202	99
à moins d' à moins de	EXPANSION.Alternative=0.0935 TEMPORAL.Asynchronous.predecence=0.0026	385
pour terminer	EXPANSION.Conjunction=0.0021	1881
à mesure qu' à mesure que	CONTINGENCY.Cause.result=0.0022 EXPANSION.Conjunction=0.0022 TEMPORAL.Synchrony=0.4219	448
pour résumer	EXPANSION.Restatement=0.1502	233
à moins qu' à moins que	CONTINGENCY.Condition=0.0096 EXPANSION.Alternative=0.4215 TEMPORAL.Asynchronous.predecence=0.0041	726
pour commencer	TEMPORAL.Asynchronous.predecence=0.0012	859
à la place	EXPANSION.Alternative.choosen alternative=0.1739 EXPANSION.Restatement=0.0020	506

French Connective	Discourse Relations/Probability	Freq
à condition qu' à condition que	COMPARISON.Contrast=0.0012 CONTINGENCY.Condition=0.0683 TEMPORAL.Asynchronous.succession=0.0035 TEMPORAL.Synchrony=0.0035	864
pour autant qu' pour autant que	COMPARISON.Contrast=0.0018 CONTINGENCY.Cause.reason=0.0160 CONTINGENCY.Condition=0.0983 EXPANSION.Conjunction=0.0006 TEMPORAL.Synchrony=0.0055	1628
pour finir	CONTINGENCY.Condition=0.0015 EXPANSION.Conjunction=0.0044	689
autant	COMPARISON.Concession=0.0002 COMPARISON.Contrast=0.0170 CONTINGENCY.Cause.reason=0.0019 CONTINGENCY.Cause.result=0.0004 CONTINGENCY.Condition=0.0009 EXPANSION.Conjunction=0.0096 TEMPORAL.Synchrony=0.0034	8450
pour conclure	EXPANSION.Conjunction=0.0018 EXPANSION.Restatement=0.0007	2757
autrement	CONTINGENCY.Cause.result=0.0077 CONTINGENCY.Condition=0.0058 EXPANSION.Alternative=0.2426 EXPANSION.Conjunction=0.0019 TEMPORAL.Asynchronous.predecence=0.0013	1554
autant dire qu' autant dire que	EXPANSION.Restatement=0.0417	24

French Connective	Discourse Relations/Probability	Freq
avant peu avant	COMPARISON.Contrast=0.0003 CONTINGENCY.Condition=0.0005 EXPANSION.Conjunction=0.0002 EXPANSION.Restatement=0.0003 TEMPORAL.Asynchronous.predecence=0.1563 TEMPORAL.Asynchronous.succession=0.0002 TEMPORAL.Synchrony=0.0015	26208
autrement dit	CONTINGENCY.Cause.result=0.0077 EXPANSION.Restatement=0.4113 TEMPORAL.Synchrony=0.0009	2205
avant même d' avant même de	TEMPORAL.Asynchronous.predecence=0.1368	117
avant d' avant de deux jours avant d' deux jours avant de peu avant d' peu avant de	COMPARISON.Contrast=0.0004 TEMPORAL.Asynchronous.predecence=0.3202 TEMPORAL.Synchrony=0.0008	5053
avant même qu' avant même que	TEMPORAL.Asynchronous.predecence=0.2480	250
avant qu' avant que	EXPANSION.Alternative=0.0012 EXPANSION.Conjunction=0.0008 TEMPORAL.Asynchronous.predecence=0.5558 TEMPORAL.Asynchronous.succession=0.0008 TEMPORAL.Synchrony=0.0027	2571

French Connective	Discourse Relations/Probability	Freq
afin d' afin de	CONTINGENCY.Cause.result=0.0233 CONTINGENCY.Condition=0.0005 TEMPORAL.Synchrony=0.0001	33375
afin qu' afin que	CONTINGENCY.Cause.reason=0.0002 CONTINGENCY.Cause.result=0.4456 CONTINGENCY.Condition=0.0072 EXPANSION.Conjunction=0.0009 TEMPORAL.Synchrony=0.0007	8901
ainsi c'est ainsi qu' c'est ainsi que	CONTINGENCY.Cause.result=0.0656 CONTINGENCY.Condition=0.0002 EXPANSION.Conjunction=0.0206 EXPANSION.Instantiation=0.0111 EXPANSION.Restatement=0.0004 TEMPORAL.Asynchronous.predecence=0.0032 TEMPORAL.Synchrony=0.0049	56126
c'est alors qu' c'est alors que	TEMPORAL.Asynchronous.predecence=0.1552	58
alors	COMPARISON.Contrast=0.0329 CONTINGENCY.Cause.reason=0.0013 CONTINGENCY.Cause.result=0.0516 CONTINGENCY.Condition=0.0036 EXPANSION.Conjunction=0.0246 TEMPORAL.Asynchronous.predecence=0.1966 TEMPORAL.Asynchronous.succession=0.0002 TEMPORAL.Synchrony=0.0739	12124

French Connective	Discourse Relations/Probability	Freq
alors même qu' alors même que	COMPARISON.Concession=0.0070 COMPARISON.Contrast=0.0831 CONTINGENCY.Cause.reason=0.0056 TEMPORAL.Asynchronous.succession=0.0014 TEMPORAL.Synchrony=0.0732	710
à ce moment-là	CONTINGENCY.Cause.result=0.0030 TEMPORAL.Asynchronous.predecence=0.1020 TEMPORAL.Synchrony=0.0015	657
alors qu' alors que	COMPARISON.Concession=0.0083 COMPARISON.Contrast=0.2487 CONTINGENCY.Cause.reason=0.0098 CONTINGENCY.Cause.result=0.0002 CONTINGENCY.Condition=0.0038 EXPANSION.Alternative.choosen alternative=0.0001 EXPANSION.Conjunction=0.0086 TEMPORAL.Asynchronous.predecence=0.0030 TEMPORAL.Asynchronous.succession=0.0022 TEMPORAL.Synchrony=0.1620	13695
à l'inverse	COMPARISON.Contrast=0.2134 TEMPORAL.Asynchronous.predecence=0.0079 TEMPORAL.Synchrony=0.0040	253
à l'instant où	TEMPORAL.Synchrony=0.1000	10
à l'époque où	TEMPORAL.Synchrony=0.1624	117
à force d' à force de	TEMPORAL.Asynchronous.succession=0.0088	114
à force	CONTINGENCY.Condition=0.3333 TEMPORAL.Asynchronous.succession=0.1667	6

French Connective	Discourse Relations/Probability	Freq
à l’heure où	COMPARISON.Contrast=0.0180 CONTINGENCY.Cause.reason=0.0252 TEMPORAL.Synchrony=0.1385	556
bien avant qu’ bien avant que	TEMPORAL.Asynchronous.predecence=0.1045	67
puisqu’ puisque	COMPARISON.Contrast=0.0041 CONTINGENCY.Cause.reason=0.4003 CONTINGENCY.Condition=0.0073 EXPANSION.Alternative.choosen alternative=0.0003 EXPANSION.Conjunction=0.0045 TEMPORAL.Asynchronous.predecence=0.0002 TEMPORAL.Asynchronous.succession=0.0007 TEMPORAL.Synchrony=0.1780	10603
à vrai dire	COMPARISON.Concession=0.0038 EXPANSION.Conjunction=0.0585 EXPANSION.Restatement=0.0208	530
bien qu’ bien que	COMPARISON.Concession=0.2013 COMPARISON.Contrast=0.2342 CONTINGENCY.Cause.result=0.0006 CONTINGENCY.Condition=0.0009 EXPANSION.Conjunction=0.0002 TEMPORAL.Synchrony=0.0005	12526
en tout état de cause	COMPARISON.Contrast=0.0055 EXPANSION.Instantiation=0.0018	550
en particulier	EXPANSION.Conjunction=0.0055 EXPANSION.Instantiation=0.0003 EXPANSION.Restatement=0.1409	25606

French Connective	Discourse Relations/Probability	Freq
d’où qu’ d’où que	EXPANSION.Conjunction=0.0192	52
en même temps	COMPARISON.Contrast=0.0013 EXPANSION.Conjunction=0.0140 TEMPORAL.Synchrony=0.0045	3780
en même temps qu’ en même temps que	COMPARISON.Contrast=0.0168 CONTINGENCY.Cause.result=0.0015 CONTINGENCY.Condition=0.0015 TEMPORAL.Synchrony=0.0061	656
quand bien même	COMPARISON.Concession=0.0299 COMPARISON.Contrast=0.0479 EXPANSION.Conjunction=0.0120	167
en outre	COMPARISON.Contrast=0.0035 CONTINGENCY.Cause.result=0.0002 EXPANSION.Conjunction=0.6156 TEMPORAL.Asynchronous.predecence=0.0011	13306
quant à	COMPARISON.Contrast=0.0022 CONTINGENCY.Condition=0.0002 EXPANSION.Conjunction=0.0024 TEMPORAL.Synchrony=0.0016	9101
également	COMPARISON.Contrast=0.0002 CONTINGENCY.Cause.result=0.0004 EXPANSION.Conjunction=0.6575 TEMPORAL.Asynchronous.predecence=0.0007	100476

French Connective	Discourse Relations/Probability	Freq
bref	COMPARISON.Contrast=0.0010 CONTINGENCY.Cause.result=0.0024 EXPANSION.Conjunction=0.0014 EXPANSION.Restatement=0.2324	2074
surtout quand	TEMPORAL.Synchrony=0.0066	151
en fait	COMPARISON.Contrast=0.0227 CONTINGENCY.Cause.reason=0.0008 CONTINGENCY.Condition=0.0002 EXPANSION.Conjunction=0.2480 EXPANSION.Restatement=0.0555	9050
quand	COMPARISON.Contrast=0.0077 CONTINGENCY.Cause.reason=0.0039 CONTINGENCY.Condition=0.0417 EXPANSION.Alternative=0.0005 EXPANSION.Conjunction=0.0016 TEMPORAL.Asynchronous.succession=0.0087 TEMPORAL.Synchrony=0.5768	11928
étant donné qu' étant donné que	CONTINGENCY.Cause.reason=0.2534 CONTINGENCY.Condition=0.0026 EXPANSION.Conjunction=0.0002 TEMPORAL.Synchrony=0.1790	6157
quand même	COMPARISON.Concession=0.0292 COMPARISON.Contrast=0.0699 CONTINGENCY.Condition=0.0006 EXPANSION.Conjunction=0.0032 TEMPORAL.Synchrony=0.0013	1574
en gros	EXPANSION.Restatement=0.0055	183

French Connective	Discourse Relations/Probability	Freq
c’est pourquoi	CONTINGENCY.Cause.result=0.2402 EXPANSION.Conjunction=0.0007	17350
à partir du moment où	CONTINGENCY.Cause.reason=0.0312 CONTINGENCY.Condition=0.0748 TEMPORAL.Asynchronous.succession=0.0903 TEMPORAL.Synchrony=0.0685	321
pourtant	COMPARISON.Concession=0.0182 COMPARISON.Contrast=0.2266 CONTINGENCY.Cause.result=0.0003 EXPANSION.Alternative=0.0005 EXPANSION.Conjunction=0.0056 TEMPORAL.Synchrony=0.0004	7710
car	COMPARISON.Contrast=0.0017 CONTINGENCY.Cause.reason=0.4306 CONTINGENCY.Cause.result=0.0008 CONTINGENCY.Condition=0.0014 EXPANSION.Alternative=0.0004 EXPANSION.Conjunction=0.0066 TEMPORAL.Synchrony=0.1077	43107
à part ca	EXPANSION.Alternative=0.5000	4
cependant qu’ cependant que	COMPARISON.Contrast=0.0145 EXPANSION.Conjunction=0.0029	1383

French Connective	Discourse Relations/Probability	Freq
cependant	COMPARISON.Concession=0.0116 COMPARISON.Contrast=0.4932 CONTINGENCY.Cause.result=0.0007 CONTINGENCY.Condition=0.0008 EXPANSION.Conjunction=0.0036 TEMPORAL.Synchrony=0.0006	21413
même quand	COMPARISON.Contrast=0.0083 TEMPORAL.Synchrony=0.0250	120
puis	COMPARISON.Contrast=0.0010 EXPANSION.Conjunction=0.0078 TEMPORAL.Asynchronous.predecence=0.0777	6181
à supposer qu' à supposer que	CONTINGENCY.Condition=0.0984	61
pour preuve	EXPANSION.Instantiation=0.0092	218
à tel point qu' à tel point que	CONTINGENCY.Cause.result=0.0088 TEMPORAL.Asynchronous.predecence=0.0088	113
premièrement	EXPANSION.Instantiation=0.0004	7560
à ce propos	EXPANSION.Conjunction=0.0008	2473
pourvu qu' pourvu que	CONTINGENCY.Cause.reason=0.0064 CONTINGENCY.Condition=0.1026	156
à propos	EXPANSION.Conjunction=0.0004	7704

French Connective	**Discourse Relations/Probability**	**Freq**
comme	COMPARISON.Contrast=0.0004 CONTINGENCY.Cause.reason=0.0061 CONTINGENCY.Cause.result=0.0002 CONTINGENCY.Condition=0.0003 EXPANSION.Conjunction=0.0056 EXPANSION.Instantiation=0.0018 TEMPORAL.Synchrony=0.2522	116924
comme quoi	CONTINGENCY.Cause.reason=0.0455	22
à présent qu' à présent que	CONTINGENCY.Cause.reason=0.2096 TEMPORAL.Synchrony=0.0014	711
la preuve qu' la preuve que preuve qu' preuve que	TEMPORAL.Synchrony=0.0021	486
résultat	CONTINGENCY.Cause.result=0.0003 EXPANSION.Conjunction=0.0002	10140
comparativement	COMPARISON.Contrast=0.0154	65
encore	COMPARISON.Concession=0.0008 COMPARISON.Contrast=0.0030 EXPANSION.Conjunction=0.0138 TEMPORAL.Asynchronous.predecence=0.0005 TEMPORAL.Synchrony=0.0002	50861

French Connective	Discourse Relations/Probability	Freq
comme s' comme si presque comme s' presque comme si un peu comme s' un peu comme si	COMPARISON.Concession=0.0006 CONTINGENCY.Condition=0.0029 EXPANSION.Conjunction=0.4689 TEMPORAL.Synchrony=0.0076	1702
faute de quoi	CONTINGENCY.Cause.reason=0.0016 EXPANSION.Alternative=0.2752 TEMPORAL.Asynchronous.predecence=0.0033	614
non sans sans sans même	COMPARISON.Concession=0.0002 COMPARISON.Contrast=0.0039 CONTINGENCY.Cause.reason=0.0008 CONTINGENCY.Condition=0.0047 EXPANSION.Alternative=0.0050 EXPANSION.Conjunction=0.0037 TEMPORAL.Asynchronous.predecence=0.0004 TEMPORAL.Synchrony=0.0006	54679
enfin	COMPARISON.Contrast=0.0007 EXPANSION.Conjunction=0.0160 EXPANSION.Restatement=0.0009 TEMPORAL.Asynchronous.predecence=0.0017	22763
cela dit	COMPARISON.Concession=0.0049 COMPARISON.Contrast=0.1852 TEMPORAL.Synchrony=0.0016	1220
encore qu' encore que	COMPARISON.Concession=0.0081 COMPARISON.Contrast=0.0509	983

French Connective	Discourse Relations/Probability	Freq
considérant qu' considérant que	COMPARISON.Contrast=0.0264 CONTINGENCY.Cause.reason=0.0422 TEMPORAL.Synchrony=0.0185	379
sachant qu' sachant que	COMPARISON.Contrast=0.0037 CONTINGENCY.Cause.reason=0.0560 EXPANSION.Conjunction=0.0037 TEMPORAL.Asynchronous.succession=0.0012 TEMPORAL.Synchrony=0.0451	821
ceci étant dit	COMPARISON.Contrast=0.0485	309
ceci dit	COMPARISON.Contrast=0.1186 EXPANSION.Conjunction=0.0020	506
néanmoins	COMPARISON.Concession=0.0744 COMPARISON.Contrast=0.4452 CONTINGENCY.Condition=0.0005 EXPANSION.Alternative.choosen alternative=0.0002 EXPANSION.Conjunction=0.0036	10309
et puis	COMPARISON.Contrast=0.0014 EXPANSION.Conjunction=0.0638 TEMPORAL.Asynchronous.predecence=0.0837	705
de l'autre de l'autre côté	COMPARISON.Contrast=0.0655 EXPANSION.Conjunction=0.0020 TEMPORAL.Asynchronous.predecence=0.0010	1953
et et encore et même	COMPARISON.Contrast=0.0013 CONTINGENCY.Cause.result=0.0006 EXPANSION.Alternative=0.0002 EXPANSION.Conjunction=0.1355 TEMPORAL.Synchrony=0.0007	1379284

French Connective	Discourse Relations/Probability	Freq
sauf à	EXPANSION.Alternative=0.1798	89
dans l'hypothèse où	CONTINGENCY.Condition=0.1895	95
dans ce cas dans ce cas-là en ce cas	COMPARISON.Contrast=0.0006 CONTINGENCY.Condition=0.0016 TEMPORAL.Asynchronous.predecence=0.0100	3206
sans compter qu' sans compter que	EXPANSION.Conjunction=0.0333	120
sans qu' sans que	COMPARISON.Concession=0.0014 COMPARISON.Contrast=0.0127 CONTINGENCY.Condition=0.0009 EXPANSION.Alternative=0.0095 EXPANSION.Conjunction=0.0100 TEMPORAL.Asynchronous.predecence=0.0086	2202
tout comme	COMPARISON.Contrast=0.0007 CONTINGENCY.Cause.result=0.0012 EXPANSION.Conjunction=0.0204 TEMPORAL.Synchrony=0.1237	4259
sauf qu' sauf que	COMPARISON.Contrast=0.1273	55
déjà	EXPANSION.Conjunction=0.0048 TEMPORAL.Asynchronous.predecence=0.0003 TEMPORAL.Synchrony=0.0007	42836
jusqu'au moment où	TEMPORAL.Asynchronous.predecence=0.2500 TEMPORAL.Synchrony=0.0227	44

French Connective	Discourse Relations/Probability	Freq
ensuite	COMPARISON.Contrast=0.0009 CONTINGENCY.Cause.result=0.0018 EXPANSION.Conjunction=0.0234 TEMPORAL.Asynchronous.predecence=0.2686 TEMPORAL.Asynchronous.succession=0.0011	10233
sans oublier qu' sans oublier que	COMPARISON.Contrast=0.0128	78
c'est dire qu' c'est dire que est -ce dire qu' est -ce dire que	EXPANSION.Restatement=0.0213	47
d'abord	TEMPORAL.Asynchronous.predecence=0.0023	5267
c'est parce qu' c'est parce que	CONTINGENCY.Cause.reason=0.0695 CONTINGENCY.Condition=0.0016 TEMPORAL.Synchrony=0.0032	633
en plus d' en plus de	EXPANSION.Conjunction=0.0063	2392
quitte à ce qu' quitte à ce que	TEMPORAL.Asynchronous.predecence=0.1111	9
d'ailleurs	COMPARISON.Contrast=0.0027 EXPANSION.Conjunction=0.1787 EXPANSION.Restatement=0.0014 TEMPORAL.Synchrony=0.0009	7676
en plus	EXPANSION.Conjunction=0.0223 TEMPORAL.Synchrony=0.0002	8325

French Connective	Discourse Relations/Probability	Freq
d’autant plus qu’ d’autant plus que	CONTINGENCY.Cause.reason=0.0067 EXPANSION.Conjunction=0.0054 TEMPORAL.Synchrony=0.0107	746
réciproquement	COMPARISON.Contrast=0.0137	73
en vue d’ en vue de	CONTINGENCY.Cause.result=0.0008 TEMPORAL.Synchrony=0.0003	14537
en supposant qu’ en supposant que	CONTINGENCY.Condition=0.0845	71
d’une part	COMPARISON.Contrast=0.0015 EXPANSION.Instantiation=0.0004	4579
d’autant qu’ d’autant que	CONTINGENCY.Cause.reason=0.0100 EXPANSION.Conjunction=0.0050 TEMPORAL.Synchrony=0.0149	402
en vérité	COMPARISON.Contrast=0.0026 EXPANSION.Conjunction=0.1016 EXPANSION.Restatement=0.0182	384
quoiqu’ quoique	COMPARISON.Concession=0.0556 COMPARISON.Contrast=0.3940 CONTINGENCY.Condition=0.0015 EXPANSION.Conjunction=0.0030 TEMPORAL.Synchrony=0.0030	665
d’un côté	COMPARISON.Contrast=0.0029	1034
en réalité	COMPARISON.Contrast=0.0198 EXPANSION.Conjunction=0.1818 EXPANSION.Restatement=0.0369	4857
quoi qu’il en soit	COMPARISON.Contrast=0.0056	1252

French Connective	Discourse Relations/Probability	Freq
d'autre part	COMPARISON.Contrast=0.2188 EXPANSION.Alternative=0.0004 EXPANSION.Conjunction=0.0884 TEMPORAL.Synchrony=0.0006	5360
remarque	EXPANSION.Conjunction=0.0007	4036
en somme	CONTINGENCY.Condition=0.0047 EXPANSION.Restatement=0.2651	215
ou alors	EXPANSION.Alternative=0.2426	202
en résumé	CONTINGENCY.Cause.result=0.0032 EXPANSION.Restatement=0.2366	617
soit	COMPARISON.Contrast=0.0006 EXPANSION.Alternative=0.0065 EXPANSION.Restatement=0.0020	53552
reste qu' reste que	COMPARISON.Contrast=0.0258	310
jusqu'à ce qu' jusqu'à ce que	CONTINGENCY.Cause.result=0.0013 TEMPORAL.Asynchronous.predecence=0.5655	794
soit dit en passant	EXPANSION.Conjunction=0.0034	297
le jour où	CONTINGENCY.Cause.reason=0.0036 CONTINGENCY.Condition=0.0072 TEMPORAL.Synchrony=0.0755	278
par voie de conséquence	CONTINGENCY.Cause.result=0.2833	120
le fait est qu' le fait est que	EXPANSION.Conjunction=0.0011	919
dans le cas où	CONTINGENCY.Condition=0.1429 TEMPORAL.Synchrony=0.0243	329

French Connective	Discourse Relations/Probability	Freq
simplement	COMPARISON.Contrast=0.0019 EXPANSION.Alternative.choosen alternative=0.0003	12867
mais aussi	COMPARISON.Contrast=0.0039 EXPANSION.Conjunction=0.0045	17223
jusqu'au jusqu'à	CONTINGENCY.Cause.result=0.0003 TEMPORAL.Asynchronous.predecence=0.0305	9605
non seulement	COMPARISON.Contrast=0.0002 EXPANSION.Conjunction=0.0007	15289
sitôt qu' sitôt que	TEMPORAL.Asynchronous.succession=0.1429	7
tant qu' tant que	COMPARISON.Contrast=0.0070 CONTINGENCY.Cause.reason=0.0006 CONTINGENCY.Cause.result=0.0001 CONTINGENCY.Condition=0.0261 EXPANSION.Alternative=0.0030 EXPANSION.Conjunction=0.0002 TEMPORAL.Asynchronous.predecence=0.0217 TEMPORAL.Synchrony=0.0043	25067
de ce fait	CONTINGENCY.Cause.result=0.1029 EXPANSION.Conjunction=0.0009 EXPANSION.Restatement=0.0018	1088

French Connective	Discourse Relations/Probability	Freq
lorsqu' lorsque	COMPARISON.Concession=0.0001 COMPARISON.Contrast=0.0024 CONTINGENCY.Cause.reason=0.0022 CONTINGENCY.Condition=0.0476 EXPANSION.Conjunction=0.0006 TEMPORAL.Asynchronous.succession=0.0165 TEMPORAL.Synchrony=0.5678	28507
en tous cas en tout cas	COMPARISON.Contrast=0.0112 EXPANSION.Conjunction=0.0014	3489
par la suite	CONTINGENCY.Cause.result=0.0015 TEMPORAL.Asynchronous.predecence=0.0795 TEMPORAL.Asynchronous.succession=0.0007	1358
le temps qu' le temps que	TEMPORAL.Synchrony=0.0051	196
toujours est-il qu' toujours est-il que	COMPARISON.Contrast=0.0395	76
somme toute	EXPANSION.Restatement=0.0148	270
soudain	TEMPORAL.Asynchronous.predecence=0.0084	238
en tous les cas	COMPARISON.Contrast=0.0061	165
toutefois	COMPARISON.Concession=0.0103 COMPARISON.Contrast=0.4342 CONTINGENCY.Cause.result=0.0006 CONTINGENCY.Condition=0.0004 EXPANSION.Conjunction=0.0029	28291

French Connective	Discourse Relations/Probability	Freq
finalement	EXPANSION.Alternative.choosen alternative=0.0004 EXPANSION.Conjunction=0.0024 EXPANSION.Restatement=0.0012 TEMPORAL.Asynchronous.predecence=0.0018	4910
a fortiori s' a fortiori si que s' que si surtout s' surtout si	COMPARISON.Contrast=0.0009 CONTINGENCY.Condition=0.1823 EXPANSION.Alternative=0.0169 EXPANSION.Conjunction=0.0017 TEMPORAL.Asynchronous.predecence=0.0015 TEMPORAL.Synchrony=0.0091	7801
faute d' faute de	CONTINGENCY.Cause.reason=0.0552 CONTINGENCY.Condition=0.0310 EXPANSION.Alternative=0.1000 EXPANSION.Conjunction=0.0011 TEMPORAL.Synchrony=0.0092	870
selon qu' selon que	CONTINGENCY.Cause.reason=0.0070	143
au motif qu' au motif que	CONTINGENCY.Cause.reason=0.1029 TEMPORAL.Asynchronous.succession=0.0049 TEMPORAL.Synchrony=0.0098	204
si bien qu' si bien que	CONTINGENCY.Cause.result=0.1933 CONTINGENCY.Condition=0.0022 EXPANSION.Conjunction=0.0133	450
notamment	EXPANSION.Conjunction=0.0109 EXPANSION.Instantiation=0.0314 EXPANSION.Restatement=0.0574	24460

French Connective	Discourse Relations/Probability	Freq
s' si	COMPARISON.Concession=0.0031 COMPARISON.Contrast=0.0191 CONTINGENCY.Cause.reason=0.0004 CONTINGENCY.Condition=0.2684 EXPANSION.Alternative=0.0028 EXPANSION.Conjunction=0.0010 TEMPORAL.Asynchronous.predecence=0.0003 TEMPORAL.Asynchronous.succession=0.0003 TEMPORAL.Synchrony=0.0083	225599
d'un autre côté	COMPARISON.Contrast=0.6219 EXPANSION.Alternative=0.0012 EXPANSION.Conjunction=0.0165	849
dans le but qu' dans le but que	CONTINGENCY.Cause.result=0.1429	7
inversement	COMPARISON.Contrast=0.2947 EXPANSION.Alternative=0.0048	207
si tant est qu' si tant est que	CONTINGENCY.Condition=0.1919 TEMPORAL.Synchrony=0.0101	99
plus particulièrement	EXPANSION.Conjunction=0.0010 EXPANSION.Restatement=0.1577	2891
dans la mesure où	CONTINGENCY.Cause.reason=0.1722 CONTINGENCY.Condition=0.0105 EXPANSION.Conjunction=0.0004 TEMPORAL.Synchrony=0.1147	4680

French Connective	Discourse Relations/Probability	Freq
plus précisément précisément	COMPARISON.Contrast=0.0006 EXPANSION.Conjunction=0.0059 EXPANSION.Instantiation=0.0001 EXPANSION.Restatement=0.0136	8582
dans le but d’ dans le but de	CONTINGENCY.Cause.result=0.0022 TEMPORAL.Synchrony=0.0004	2251
sinon	CONTINGENCY.Condition=0.0263 EXPANSION.Alternative=0.3129 EXPANSION.Conjunction=0.0029	2052
simultanément	COMPARISON.Contrast=0.0010 EXPANSION.Conjunction=0.0163 TEMPORAL.Synchrony=0.0020	981
en ce sens	CONTINGENCY.Cause.result=0.0148 EXPANSION.Conjunction=0.0012	2506
même qu’ même que	COMPARISON.Contrast=0.0042 EXPANSION.Conjunction=0.1086 TEMPORAL.Synchrony=0.1983	958
de manière à ce qu’ de manière à ce que	CONTINGENCY.Cause.result=0.2791	1211
de manière à	CONTINGENCY.Cause.result=0.1070 TEMPORAL.Asynchronous.predecence=0.0007	2786
tout de même	COMPARISON.Concession=0.0204 COMPARISON.Contrast=0.0270 EXPANSION.Conjunction=0.0042	1665

French Connective	Discourse Relations/Probability	Freq
de même qu' de même que	CONTINGENCY.Cause.result=0.0010 CONTINGENCY.Condition=0.0003 EXPANSION.Conjunction=0.0301 TEMPORAL.Synchrony=0.0531	2958
nonobstant	COMPARISON.Concession=0.0162 COMPARISON.Contrast=0.0757	185
de même	COMPARISON.Contrast=0.0064 CONTINGENCY.Cause.result=0.0013 EXPANSION.Conjunction=0.3122 TEMPORAL.Synchrony=0.0015	5928
même s' même si	COMPARISON.Concession=0.0820 COMPARISON.Contrast=0.2101 CONTINGENCY.Cause.reason=0.0003 CONTINGENCY.Condition=0.0038 EXPANSION.Conjunction=0.0006 TEMPORAL.Synchrony=0.0007	11583
même	COMPARISON.Concession=0.0006 COMPARISON.Contrast=0.0058 CONTINGENCY.Condition=0.0006 EXPANSION.Conjunction=0.0275 EXPANSION.Restatement=0.0006	57440
de manière qu' de manière que	CONTINGENCY.Cause.result=0.3333	21
de telle manière qu' de telle manière que	CONTINGENCY.Cause.result=0.0616 EXPANSION.Conjunction=0.0137	146

French Connective	Discourse Relations/Probability	Freq
or	COMPARISON.Concession=0.0023 COMPARISON.Contrast=0.2324 CONTINGENCY.Cause.reason=0.0006 CONTINGENCY.Cause.result=0.0028 CONTINGENCY.Condition=0.0026 EXPANSION.Alternative.choosen alternative=0.0011 EXPANSION.Conjunction=0.0659 EXPANSION.Restatement=0.0003 TEMPORAL.Synchrony=0.0028	6460
de plus	COMPARISON.Contrast=0.0002 EXPANSION.Conjunction=0.1375	23218
tandis qu' tandis que	COMPARISON.Concession=0.0031 COMPARISON.Contrast=0.4656 CONTINGENCY.Cause.reason=0.0011 CONTINGENCY.Condition=0.0011 EXPANSION.Conjunction=0.0648 TEMPORAL.Synchrony=0.0597	3565
maintenant qu' maintenant que	CONTINGENCY.Cause.reason=0.3854 EXPANSION.Conjunction=0.0013 TEMPORAL.Asynchronous.succession=0.0031 TEMPORAL.Synchrony=0.0138	1593
de facon à	CONTINGENCY.Cause.result=0.0934 EXPANSION.Conjunction=0.0009	1071
maintenant	COMPARISON.Contrast=0.0003 CONTINGENCY.Cause.reason=0.0024 EXPANSION.Conjunction=0.0007 TEMPORAL.Asynchronous.predecence=0.0009	17598

French Connective	Discourse Relations/Probability	Freq
de facon qu' de facon que	CONTINGENCY.Cause.result=0.3103	58
deuxièmement	COMPARISON.Contrast=0.0012 EXPANSION.Conjunction=0.0033 TEMPORAL.Asynchronous.predecence=0.0008	8851
malgré le fait qu' malgré le fait que	COMPARISON.Concession=0.0147 COMPARISON.Contrast=0.0118 TEMPORAL.Synchrony=0.0029	339
non plus	EXPANSION.Conjunction=0.0188	6560
de fait	COMPARISON.Contrast=0.0091 CONTINGENCY.Cause.result=0.0020 EXPANSION.Conjunction=0.1330 EXPANSION.Restatement=0.0300	1534
mais	COMPARISON.Concession=0.0014 COMPARISON.Contrast=0.5841 CONTINGENCY.Cause.reason=0.0001 CONTINGENCY.Condition=0.0003 EXPANSION.Alternative.choosen alternative=0.0005 EXPANSION.Conjunction=0.0088 EXPANSION.Restatement=0.0006 TEMPORAL.Synchrony=0.0005	149752

French Connective	Discourse Relations/Probability	Freq
qu' que	COMPARISON.Contrast=0.0019 CONTINGENCY.Cause.reason=0.0022 CONTINGENCY.Cause.result=0.0006 CONTINGENCY.Condition=0.0041 EXPANSION.Conjunction=0.0020 TEMPORAL.Asynchronous.predecence=0.0003 TEMPORAL.Synchrony=0.0070	927002
de facon à ce qu' de facon à ce que	CONTINGENCY.Cause.result=0.2829 EXPANSION.Conjunction=0.0020	509
en second lieu	TEMPORAL.Asynchronous.predecence=0.0024	414
de telle facon qu' de telle facon que	CONTINGENCY.Cause.result=0.0556	72
surtout qu' surtout que	EXPANSION.Conjunction=0.0032	312
surtout	EXPANSION.Alternative.choosen alternative=0.0002 EXPANSION.Conjunction=0.0077 EXPANSION.Restatement=0.0281	16559
pour ce faire	COMPARISON.Contrast=0.0013 CONTINGENCY.Cause.result=0.0057 EXPANSION.Conjunction=0.0006	1591
de la même manière qu' de la même manière que	CONTINGENCY.Condition=0.0025 TEMPORAL.Synchrony=0.0102	393

French Connective	Discourse Relations/Probability	Freq
par contre	COMPARISON.Concession=0.0015 COMPARISON.Contrast=0.4049 CONTINGENCY.Condition=0.0011 EXPANSION.Alternative.choosen alternative=0.0211 EXPANSION.Conjunction=0.0019 TEMPORAL.Asynchronous.predecence=0.0008	2655
malgré tout	COMPARISON.Concession=0.0461 COMPARISON.Contrast=0.0906 EXPANSION.Conjunction=0.0025	1214
de la même facon	CONTINGENCY.Cause.result=0.0021 EXPANSION.Conjunction=0.1680	482
certes	COMPARISON.Concession=0.0120 COMPARISON.Contrast=0.0246 EXPANSION.Conjunction=0.0105 TEMPORAL.Synchrony=0.0004	5325
malgré qu' malgré que	COMPARISON.Concession=0.0602 COMPARISON.Contrast=0.0843 TEMPORAL.Asynchronous.succession=0.0120	83
en fin de compte	EXPANSION.Restatement=0.0019 TEMPORAL.Asynchronous.predecence=0.0009 TEMPORAL.Synchrony=0.0005	2126
mieux	CONTINGENCY.Condition=0.0003 EXPANSION.Alternative.choosen alternative=0.0005 EXPANSION.Restatement=0.0009	14838
de la même manière	CONTINGENCY.Cause.result=0.0024 EXPANSION.Conjunction=0.1665	841

French Connective	Discourse Relations/Probability	Freq
de la même facon qu' de la même facon que	COMPARISON.Contrast=0.0049 TEMPORAL.Synchrony=0.0098	204
malheureusement	COMPARISON.Contrast=0.0093	11960
du moins	COMPARISON.Contrast=0.0007 EXPANSION.Alternative=0.0051	2754
pendant qu' pendant que	COMPARISON.Concession=0.0021 COMPARISON.Contrast=0.4163 EXPANSION.Conjunction=0.0043 TEMPORAL.Synchrony=0.1845	466
peu importe	COMPARISON.Contrast=0.0021	475
du moment qu' du moment que	CONTINGENCY.Cause.reason=0.0323 CONTINGENCY.Condition=0.1613 TEMPORAL.Asynchronous.succession=0.0161 TEMPORAL.Synchrony=0.0323	62
parallèlement	COMPARISON.Contrast=0.0065 CONTINGENCY.Cause.result=0.0007 EXPANSION.Conjunction=0.0289 TEMPORAL.Synchrony=0.0088	2942
donc	COMPARISON.Contrast=0.0005 CONTINGENCY.Cause.result=0.5124 EXPANSION.Alternative.choosen alternative=0.0001 EXPANSION.Conjunction=0.0044 EXPANSION.Restatement=0.0039 TEMPORAL.Asynchronous.predecence=0.0057	59739
et dire qu' et dire que	EXPANSION.Conjunction=0.0039	254

French Connective	**Discourse Relations/Probability**	**Freq**
du fait qu' du fait que	COMPARISON.Contrast=0.0012 CONTINGENCY.Cause.reason=0.0277 CONTINGENCY.Condition=0.0019 EXPANSION.Conjunction=0.0002 TEMPORAL.Synchrony=0.0053	5803
parce qu' parce que	COMPARISON.Contrast=0.0001 CONTINGENCY.Cause.reason=0.6484 CONTINGENCY.Cause.result=0.0003 CONTINGENCY.Condition=0.0004 EXPANSION.Conjunction=0.0007 TEMPORAL.Synchrony=0.0292	31899
du coup	CONTINGENCY.Cause.result=0.0085	118
par suite	CONTINGENCY.Cause.result=0.0175	114
en revanche	COMPARISON.Concession=0.0023 COMPARISON.Contrast=0.4968 EXPANSION.Alternative.choosen alternative=0.0248 EXPANSION.Conjunction=0.0034 EXPANSION.Restatement=0.0046	2623
comme ca	CONTINGENCY.Cause.result=0.0057	175
du reste	COMPARISON.Contrast=0.0035 EXPANSION.Conjunction=0.1161	1430
du temps où	TEMPORAL.Synchrony=0.2308	13
vu qu' vu que	CONTINGENCY.Cause.reason=0.2249 CONTINGENCY.Condition=0.0020 TEMPORAL.Asynchronous.succession=0.0007 TEMPORAL.Synchrony=0.1317	1503

French Connective	**Discourse Relations/Probability**	**Freq**
une fois qu' une fois que	TEMPORAL.Asynchronous.succession=0.4290 TEMPORAL.Synchrony=0.0463	951
de toute manière de toutes manières	COMPARISON.Contrast=0.0077 EXPANSION.Conjunction=0.0039	259
par conséquent	COMPARISON.Contrast=0.0015 CONTINGENCY.Cause.reason=0.0002 CONTINGENCY.Cause.result=0.5503 EXPANSION.Conjunction=0.0009 EXPANSION.Restatement=0.0010 TEMPORAL.Asynchronous.predecence=0.0016	15720
depuis qu' depuis que	CONTINGENCY.Cause.reason=0.5249 CONTINGENCY.Condition=0.0020 TEMPORAL.Asynchronous.succession=0.0081 TEMPORAL.Synchrony=0.0010	985
à	CONTINGENCY.Condition=0.0004 EXPANSION.Conjunction=0.0002 TEMPORAL.Synchrony=0.0016	1110272
par exemple	EXPANSION.Instantiation=0.6027 EXPANSION.Restatement=0.0005	22029
depuis	CONTINGENCY.Cause.reason=0.0324 TEMPORAL.Asynchronous.succession=0.0004 TEMPORAL.Synchrony=0.0006	25545
ou bien ou bien encore	COMPARISON.Contrast=0.0010 EXPANSION.Alternative=0.2586 EXPANSION.Conjunction=0.0021	955

French Connective	Discourse Relations/Probability	Freq
de sorte qu' de sorte que	CONTINGENCY.Cause.result=0.3910 EXPANSION.Conjunction=0.0132 TEMPORAL.Synchrony=0.0004	2642
ou ou encore	CONTINGENCY.Condition=0.0003 EXPANSION.Alternative=0.0681 EXPANSION.Conjunction=0.0045 TEMPORAL.Synchrony=0.0001	99957
par ailleurs	COMPARISON.Contrast=0.0687 EXPANSION.Alternative=0.0057 EXPANSION.Conjunction=0.4697 EXPANSION.Restatement=0.0007 TEMPORAL.Asynchronous.predecence=0.0012 TEMPORAL.Synchrony=0.0015	9610
un peu plus tard	TEMPORAL.Asynchronous.predecence=0.0127	79
outre le fait qu' outre le fait que outre qu' outre que	CONTINGENCY.Cause.reason=0.0055 EXPANSION.Conjunction=0.0874	183
de toute facon de toutes facons	COMPARISON.Contrast=0.0141 EXPANSION.Conjunction=0.0022	1347
un jour	TEMPORAL.Synchrony=0.0008	2423
par le fait qu' par le fait que	CONTINGENCY.Cause.reason=0.0227 TEMPORAL.Synchrony=0.0009	1056
à en	CONTINGENCY.Cause.result=0.0007 CONTINGENCY.Condition=0.0013 TEMPORAL.Synchrony=0.0027	1489

French Connective	Discourse Relations/Probability	Freq
dire encore qu' dire encore que dire qu' dire que	CONTINGENCY.Cause.reason=0.0003 EXPANSION.Conjunction=0.0003 TEMPORAL.Synchrony=0.0002	18793
en attendant	COMPARISON.Contrast=0.0173 TEMPORAL.Asynchronous.predecence=0.0242 TEMPORAL.Synchrony=0.1028	866
en	COMPARISON.Contrast=0.0007 CONTINGENCY.Cause.reason=0.0002 CONTINGENCY.Cause.result=0.0003 CONTINGENCY.Condition=0.0014 EXPANSION.Conjunction=0.0016 TEMPORAL.Asynchronous.predecence=0.0001 TEMPORAL.Synchrony=0.0061	691207
même en notamment en qu'en	CONTINGENCY.Condition=0.0019 EXPANSION.Conjunction=0.0009 EXPANSION.Instantiation=0.0009 EXPANSION.Restatement=0.0016 TEMPORAL.Asynchronous.predecence=0.0002 TEMPORAL.Synchrony=0.0019	12643
tout en	COMPARISON.Concession=0.0013 COMPARISON.Contrast=0.1689 CONTINGENCY.Cause.result=0.0007 CONTINGENCY.Condition=0.0002 EXPANSION.Conjunction=0.0034 TEMPORAL.Synchrony=0.0552	8698

French Connective	Discourse Relations/Probability	Freq
en effet	COMPARISON.Contrast=0.0064 CONTINGENCY.Cause.reason=0.0414 CONTINGENCY.Cause.result=0.0002 CONTINGENCY.Condition=0.0004 EXPANSION.Alternative.choosen alternative=0.0003 EXPANSION.Conjunction=0.2532 EXPANSION.Restatement=0.0119 TEMPORAL.Synchrony=0.0044	18775
in fine	EXPANSION.Restatement=0.0068	148
comme pour pour sauf pour	COMPARISON.Contrast=0.0003 CONTINGENCY.Cause.reason=0.0003 CONTINGENCY.Cause.result=0.0027 CONTINGENCY.Condition=0.0028 EXPANSION.Conjunction=0.0005 TEMPORAL.Asynchronous.predecence=0.0006 TEMPORAL.Synchrony=0.0034	483077
dès qu' dès que	CONTINGENCY.Cause.reason=0.0017 CONTINGENCY.Condition=0.0046 TEMPORAL.Asynchronous.succession=0.2840 TEMPORAL.Synchrony=0.0608	2370
pour autant	COMPARISON.Concession=0.0071 COMPARISON.Contrast=0.0977 CONTINGENCY.Cause.result=0.0047 CONTINGENCY.Condition=0.0459 EXPANSION.Conjunction=0.0018 TEMPORAL.Synchrony=0.0024	1699

French Connective	Discourse Relations/Probability	Freq
pour qu' pour que	CONTINGENCY.Cause.reason=0.0009 CONTINGENCY.Cause.result=0.1415 CONTINGENCY.Condition=0.0242 EXPANSION.Conjunction=0.0013 TEMPORAL.Asynchronous.predecence=0.0059 TEMPORAL.Synchrony=0.0010	17469
dès lors qu' dès lors que	CONTINGENCY.Cause.reason=0.1079 CONTINGENCY.Cause.result=0.0078 CONTINGENCY.Condition=0.0152 EXPANSION.Conjunction=0.0005 TEMPORAL.Asynchronous.succession=0.0152 TEMPORAL.Synchrony=0.0836	2178
plutôt	COMPARISON.Contrast=0.0027 EXPANSION.Alternative.choosen alternative=0.1247 EXPANSION.Conjunction=0.0033 EXPANSION.Restatement=0.0793 TEMPORAL.Asynchronous.predecence=0.0003	6934
décidément	EXPANSION.Restatement=0.0066	152
dès lors	CONTINGENCY.Cause.reason=0.0045 CONTINGENCY.Cause.result=0.4923 CONTINGENCY.Condition=0.0029 EXPANSION.Conjunction=0.0020 TEMPORAL.Asynchronous.predecence=0.0099 TEMPORAL.Asynchronous.succession=0.0007 TEMPORAL.Synchrony=0.0104	10527
plutôt qu' plutôt que	EXPANSION.Alternative.choosen alternative=0.0031 EXPANSION.Conjunction=0.0003	2927

French Connective	Discourse Relations/Probability	Freq
deux mois plus tard plus tard quelques jours plus tard	TEMPORAL.Asynchronous.predecence=0.0065	3692
effectivement	COMPARISON.Contrast=0.0094 CONTINGENCY.Cause.reason=0.0003 CONTINGENCY.Cause.result=0.0007 EXPANSION.Conjunction=0.1125 EXPANSION.Restatement=0.0188 TEMPORAL.Synchrony=0.0004	6802
après	COMPARISON.Concession=0.0003 COMPARISON.Contrast=0.0001 CONTINGENCY.Cause.reason=0.0054 CONTINGENCY.Condition=0.0011 TEMPORAL.Asynchronous.predecence=0.0065 TEMPORAL.Asynchronous.succession=0.0593 TEMPORAL.Synchrony=0.0081	29173
17 ans après après plusieurs mois cinq ans après huit jours après huit mois après peu après plus de quatre-vingts ans après trois mois après un mois après	TEMPORAL.Asynchronous.succession=0.0030	332

French Connective	Discourse Relations/Probability	Freq
après qu' après que quelques mois après qu' quelques mois après que six mois après qu' six mois après que un mois après qu' un mois après que	COMPARISON.Concession=0.0012 CONTINGENCY.Cause.reason=0.0431 TEMPORAL.Asynchronous.succession=0.5401 TEMPORAL.Synchrony=0.0251	835
après tout	CONTINGENCY.Cause.reason=0.0010 TEMPORAL.Synchrony=0.0015	2046
attendu qu' attendu que	CONTINGENCY.Cause.reason=0.0783 TEMPORAL.Synchrony=0.1084	166
après quoi	TEMPORAL.Asynchronous.predecence=0.1307	176
au cas où	CONTINGENCY.Condition=0.2250 TEMPORAL.Synchrony=0.0133	600
au bout du compte	CONTINGENCY.Cause.result=0.0034	593
à défaut d' à défaut de	CONTINGENCY.Cause.reason=0.0094 CONTINGENCY.Condition=0.0281 EXPANSION.Alternative=0.0281	320
à cet égard	COMPARISON.Contrast=0.0003 CONTINGENCY.Cause.result=0.0023	11616
à dire vrai	EXPANSION.Conjunction=0.0962	52
en définitive	EXPANSION.Conjunction=0.0081 EXPANSION.Restatement=0.0749 TEMPORAL.Synchrony=0.0010	988
en d'autres termes	CONTINGENCY.Cause.result=0.0013 EXPANSION.Restatement=0.6390	2249

French Connective	Discourse Relations/Probability	Freq
au contraire	COMPARISON.Contrast=0.3395 CONTINGENCY.Cause.reason=0.0004 EXPANSION.Alternative=0.0007 EXPANSION.Alternative.choosen alternative=0.0873 EXPANSION.Conjunction=0.0088 EXPANSION.Restatement=0.0218	5694
en dépit du fait qu' en dépit du fait que	COMPARISON.Concession=0.0050 COMPARISON.Contrast=0.0099	202
en comparaison	COMPARISON.Contrast=0.0456 TEMPORAL.Synchrony=0.0038	263
au fait	CONTINGENCY.Cause.reason=0.0018	2778
dans le sens où	CONTINGENCY.Cause.reason=0.0340 EXPANSION.Conjunction=0.0031 TEMPORAL.Synchrony=0.0062	324
en conséquence	COMPARISON.Contrast=0.0028 CONTINGENCY.Cause.result=0.4447 TEMPORAL.Asynchronous.predecence=0.0008	3605
au lieu d' au lieu de	EXPANSION.Alternative.choosen alternative=0.0134 TEMPORAL.Synchrony=0.0004	4470
par comparaison	COMPARISON.Contrast=0.0306 EXPANSION.Instantiation=0.0102	98
au lieu	EXPANSION.Alternative.choosen alternative=0.1750 TEMPORAL.Synchrony=0.0052	577

French Connective	Discourse Relations/Probability	Freq
aussi	COMPARISON.Contrast=0.0016 CONTINGENCY.Cause.result=0.0125 EXPANSION.Conjunction=0.4352 TEMPORAL.Asynchronous.predecence=0.0006 TEMPORAL.Synchrony=0.0021	79669
en bref	EXPANSION.Conjunction=0.0026 EXPANSION.Restatement=0.5556	378
dans le sens qu' dans le sens que	TEMPORAL.Synchrony=0.0357	112
au moment d' au moment de	COMPARISON.Contrast=0.0013 TEMPORAL.Asynchronous.succession=0.0006 TEMPORAL.Synchrony=0.1658	1562
en ce sens qu' en ce sens que	CONTINGENCY.Cause.reason=0.0114 TEMPORAL.Synchrony=0.0016	612