Automatic extraction of causal knowledge from natural language texts

Grivaz, Cécile

Automatic extraction of causal knowledge from natural language texts

Cécile Grivaz
Supervisors: Jacques Moeschler and Martin Rajman
Jury president: Paola Merlo
Other jury members: Paul Sabatier and Caroline Sporleder

Thesis N° 747
Genève
2012

La faculté des lettres sur le préavis d'une comission composée de Mesdames et Messieurs les professeurs Paola Merlo, présidente du jury ; Jacques Moeschler, directeur de thèse ; Dr. Martin Rajman (EPFL) ; Caroline Sporleder (Saarland Universität, Saarbrücken) ; Paul Sabatier (CNRS et Université de Provence), autorise l'impression de la présente thèse, sans exprimer d'opinion sur les propositions qui y sont énoncées.

Genève, le 16 mai 2012

Le Doyen : Nicolas ZUFFEREY

Thèse N° 747

À ceux qui me manquent

Acknowledgements

First and foremost, I would like to express my deep gratitude towards my thesis supervisors Jacques Moeschler and Martin Rajman. Jacques Moeschler supervised the linguistics aspect of this thesis. He offered me a considerable amount of time, lots of advices and very interesting discussions. He was always patient with my occasionally lacking understanding of linguistics terms, and open and curious to the whole computational part of the thesis. Martin Rajman never ceased to amaze me with the sharpness of his intelligence. He was also very open and available to answer my numerous questions and need for advice. Clearly, this thesis owes incalculably to their supervision.

This thesis was financed by a Swiss National Science Foundation project (100012-113382), by a scholarship from the Ernest Boninchi Foundation and by a Swiss National Science Foundation fellowships for prospective researchers (1-129213). These institutions made this work possible, for which they have my gratitude.

My deep gratitude goes toward Caroline Sporleder. She supervised me during my stay in Saarbrücken. She was always available, and gave me the advices and evaluation that I needed. My stay was also an opportunity for very interesting discussions, and improved this work considerably. Finally, Caroline Sporleder accepted to be part of my thesis jury and to evaluate this work.

I wish to thank Paola Merlo, the jury president, and Paul Sabatier for evaluating this thesis. Their feedback and advices were extremely helpful, and I am deeply grateful for the time they dedicated to evaluating this thesis and to participate to my thesis defence.

On a more personal note, I wish to thank my closest friends. Emmanuel Eckard knows that I can't possibly sufficiently express my gratitude for his

scientific, logistical, linguistic, psychological and general help. I am proud and grateful to be part of the group that we form together with Stéphane Magenat, Elisa Laurenti and Cyrille Dunant. You were always there to discus any subject and science was not the least of our talks. You provided most of my physical and metaphorical food for thought. I am grateful in many ways, but suffice to say here that you, with my other friends, were part of the necessary conditions for this work.

Finalement, je suis profondément reconnaissante pour l'aide que j'ai reçue de ma famille. Mes parents m'ont élevée dans l'idée qu'ils me soutiendraient dans ce que je choisirai d'entreprendre, sans faire jamais peser sur moi les pressions que la société exerce, et en particulier celles liées aux apparences et au genre. Mes parents et ma sœur font preuve d'une recherche constante de ce qui est sincère et précieux, et cette atmosphère m'a permis de recevoir leur indispensable et considérable soutien avec confiance et sérénité.

Abstract

Causal knowledge is fundamental for our understanding of the world. A programme that would automatically recognise the expression of causation would, for example, allow a user to save time in finding causes or consequences in a large text corpora. This would be useful for finding possible causes of a specific symptom in a medical corpus, for example. The object of this thesis is to study the automatic recognition of relations that hold between clauses, such as in *All is white, it has been snowing for a few hours*, especially when they do not bear any causal marker such as *because*. Previous work on this task, although providing results that were more accurate than a random baseline, does not achieve sufficient accuracy for the practical use of such systems.

The first part of this thesis studies the human recognition of implicit causal relations by carrying out annotation experiments that aim for maximum reproducibility, using only naïve annotators and written guidelines. In these experiments, we validate the hypothesis that, despite the difficulty in finding a precise common-sense definition of causation, it is possible to identify a set of simpler features of causation that are statistically associated with its recognition in human reasoning. We list these features and use them as a basis for an annotation manual. We show that the task is subjective and leads to a low inter-annotator agreement, even with the use of carefully designed annotation instructions. We then propose an evaluation protocol that takes into account the intrinsic difficulty of defining the task, and its subjective nature. Our protocol makes use of the given inter-annotator agreement and compares the results of the automatic system to the normal human disagreement range.

Previously presented linguistic theories to explain how humans can recognise implicit causation, primarily make use of *world knowledge* that codes whether two eventualities are normally causally connected in the world. Almost all previous computational linguistics work represents this world knowledge in some way, typically as a lexical feature. In this thesis, we show that the most likely feature used to represent eventualities –verb pairs– is not predictive of causation in practice. To come to this conclusion, we compute

verb pairs predictability on large corpora and show that, as the reliability of the measure grows, the measure tends to show statistical independence.

We then study the exact nature of the necessary world knowledge by manually analysing implicit causation occurrences on a small corpus. We show that the eventualities that appear in causal relations call for a much more complex representation than simple verb pairs. Moreover, the current state of the art in computational linguistics is not sufficient to allow us to represent nor to acquire the world knowledge necessary for the accurate automatic recognition of causal relations.

In this thesis, we argue that the field of causal relation recognition cannot make important progress without first solving the problem of abstract eventuality representation and of clustering textual representations of eventualities into the corresponding abstract eventuality classes.

Contents

1 Introduction **7**

1.1 Motivation 7

1.2 Definition of the task 9

1.3 Research questions 11

1.4 Thesis organisation 13

2 State of the art **15**

2.1 Introduction 15

2.2 Linguistics models for the human recognition of implicit causation in natural language texts 16

2.2.1 Segmented discourse representation theory 18

2.2.2 Relevance theory 21

2.2.3 Other theories of causation 29

2.3 Computational linguistics framework: machine learning for categorisation tasks 31

2.4 Previous work on the automatic recognition of causal relations 32

2.4.1 Causal relation recognition 33

2.4.2 Discourse relation classification 35

2.5 Corpora annotated with implicit causal relations 40

2.5.1 The rhetorical structure theory corpus and the Penn discourse treebank 41

2.5.2 The ANNODIS project 41

2.5.3 Inui 42

2.5.4 Hovy and colleagues 43

2.6 Evaluation 43
2.6.1 Gold standard and the problem of subjectivity and low inter-annotator agreements 45
2.6.2 Creating a gold standard 48
2.6.3 Inter-annotator agreement measures 50
2.6.4 Evaluation that takes into account the inter-annotator agreement of the gold standard. 52
2.7 Conclusion 52

3 Human recognition of causation 54
3.1 Introduction 54
3.2 Intuitive characteristics 57
3.2.1 Methodology 57
3.2.2 Results 58
3.2.3 Discussion of the results 61
3.3 Correlation between judgement on causation and on its characteristics 66
3.3.1 Methodology 66
3.3.2 Result analysis 69
3.4 Annotation instructions 72
3.4.1 Linguistic tests 72
3.4.2 Counterfactuality 75
3.4.3 Temporal asymetry 75
3.4.4 Asymetry 76
3.4.5 Causal chains 76
3.4.6 Further annotation rules 77
3.4.7 Testing the manual 77
3.5 Taking the subjectivity of a task into account for its evaluation 79
3.6 Conclusion and perspectives 83

4 Verb pairs 87
4.1 Introduction 87

4.2 Methodology 89
4.2.1 Description of our corpora 89
4.2.2 Modelling of causation 93
4.3 Frequency curves results: a very long tail 94
4.3.1 Distribution of verb pairs in general corpora 94
4.3.2 Distribution of verb pairs in causal relations 98
4.4 Predictability 101
4.4.1 Choice of a predictability measure 101
4.4.2 Verb pair predictability for the presence of *because* . . 103
4.4.3 PDTB predictability 105
4.5 Some classification experiments 113
4.6 Conclusion 115

5 World knowledge **117**
5.1 Introduction 117
5.2 Corpus description 120
5.3 Textual order of the eventualities 124
5.4 Using causal rules to recognise implicit causation 128
5.5 Dimensions of causal rules: causal chains and eventuality hierarchies 129
5.5.1 Causal rules and causal chains: how far does a causal rule go? 129
5.5.2 Causal rules and eventuality hierarchies: how specific is a causal rule? 133
5.6 Predictable difficulties for automation 136
5.6.1 Representation of causal rules 137
5.6.2 Acquiring causal rules 137
5.6.3 Exploiting causal rules 138
5.6.4 Matching clauses to eventuality types 139
5.6.5 Matching eventualities to causal rules 140
5.6.6 Other necessary information 141
5.7 Conclusion 142

6 Conclusion **145**

6.1 Main contributions 146

6.2 Perspectives 147

A French annotation manual **157**

B Texts to analyse for the annotation experiments **160**

B.1 Eliciting intuitive characteristics of causation 160

B.1.1 Parce que without connector 160

B.1.2 With parce que 162

B.1.3 Donc without connector 162

B.1.4 With donc 163

B.1.5 Mais removed 164

B.1.6 With mais 164

B.2 Example of a file that was printed and given to a subject for this experiment 165

B.3 First annotation experiment 169

B.4 Second annotation experiment 171

C Implicit reason occurrences in children tales **177**

D A small sample of the feature extraction programme **179**

D.1 Lemmatizing 179

D.2 Getting hypernyms for content words 180

List of Figures

2.1 Example of an SDRT relation tree. 20
2.2 Directional inferences information type hierarchy. 26

3.1 Types of sentences identified as causal 63
3.2 Justification types for causal intuitions 64
3.3 Test programme CARMA. 67
3.4 Characteristics associated with causal and non-causal sentences. 69
3.5 Association of characteristics with causation. 70
3.6 New CARMA. 73

4.1 The contingency relation in the Penn discourse treebank . . . 91
4.2 Corpora used for predictability experiments with verb pairs . 93
4.3 Predictable false positives and negatives due when using the presence of *because* as a model of causation. 94
4.4 Total inter-sentential occurrences of verb pairs in linear scale on ChiC 96
4.5 Total inter-sentential occurrences of verb pairs in log scale on ChiC 96
4.6 Total inter-sentential occurrences of verb pairs in log scale in APW 97
4.7 Total inter-sentential occurrences of verb pairs in log scale in PDTB 97
4.8 Frequent general verb pairs. 98
4.9 Causal occurrences of verb pairs in log scale on ChiC 99
4.10 Causal occurrences of verb pairs in log scale in APW 99

4.11 Causal occurrences of verb pairs in log scale in PDTB 100
4.12 Top ten verb pairs in causal relations 100
4.13 Top 20 predictive verb pairs for the presence of *because* 104
4.14 Intersections and unions of several PDTB partitioning 108
4.15 Predicitve verb pairs in the Penn Discourse Treebank 109
4.16 Frequency vs Q on the *because* partition 110
4.17 Frequency vs Q on the *cause* partition 110
4.18 Frequency vs Q on the *explicit cause* partition 111
4.19 Frequency vs Q on the *implicit cause* partition 111
4.20 Frequency vs Q on the *implicit reason* partition 112
4.21 Frequency vs Q on the *reason* partition 112
4.22 Automatic classification results 114

5.1 Different types of causal relations in French tales 122
5.2 Causal markers in French tales 123

Chapter 1

Introduction

The recognition of causal relations is a fundamental part of human reasoning. Imitating this ability is thus an important task in artificial intelligence. In this thesis, we will explore the recognition of causal relations in a specific environment: natural language texts. As causation is a central part of our daily lives, it is typically common in the texts that we produce. We use it to describe relations between eventualities that happen to us or that we witness. We use it to convince the people that listen to us or read our arguments. It is our primary source of explanations.

The goal of this thesis is to study the expression of causation in natural-language texts. We aim to study relations that hold between clauses such as in *All is white, it has been snowing for a few hours*, especially when they do not bear any causal marker such as *because*. As the only source of answers to *why* questions, causation recognition would be a useful addition to many natural-language processing systems, in particular to question-answering and dialogue systems.

1.1 Motivation

The recognition of causal relations and the more general task of discourse relations classification have been the motivation for much previous computational linguistics work, as we will see in chapter 2. Such a system would indeed be very useful for the field, as a part of other natural language processing systems. It would allow researchers to develop better programmes

such as text summarisers, as causal relations are important data in a text. It would also be a central system in question answering, as causal relations are the main source of answers to *why* questions. This argument also applies to dialogue systems. Swanson and Smalheiser (1997) have even hypothesised that it is possible to find new scientific information by transitively combining causal links that are documented in scientific studies corpora. In general, a system that is able to recognise and extract causal relations in natural language texts would fundamentally ease the retrieval of information from large corpora.

Similar motivations have led many authors to attempt to build this type of systems. These can be programmes that aim at recognising causal relations only, or more general attempts in the field of discourse relation recognition. Good results have been achieved for marked relations (Khoo et al., 2000; Zhou et al., 2010). The field of discourse relation classification experienced a growing interest in the past decade, which is evident as, or maybe even partially due to, the release of two manually-annotated corpora of discourse relations (Carlson et al., 2002; Prasad et al., 2007).

Despite this recent interest growth, implicit relations recognition is still an open problem. The recent work of Pitler and her colleagues (Pitler et al., 2009), for example, achieves an F-score of 47.13% corresponding to a 67.3% accuracy for the super class of causation named *contingency*. These results tend to be better than random baselines, but such systems are not sufficiently accurate to be used in practice; for example, as a part of a larger question-answering system. Furthermore, these results are evaluations of a classifier that recognises a super-classes of causation and, as far as we know, nobody has yet succeeded in classifying the finer grained relations of the Penn Discourse Treebank (the evaluation corpus used by Pitler and her colleagues.)

In this thesis, we will explain the lack of accurate results in implicit causation recognition as a consequence of the lack of world knowledge resources and of the intrinsic difficulties in building them, whether manually or automatically. We will argue that the current state of the art in natural language processing is not sufficient to give us the necessary tools to solve this problem in a satisfactory manner. Finally, we will give directions for the necessary computational linguistics research that must be undertaken in order to give

rise to accurate implicit causation recognition systems at some future date.

1.2 Definition of the task

The task that this thesis will study can be defined using several axes. In this section, we will specify this task. We will characterise the phenomenon that we wish to automatically recognise and we will sketch out the natural language processing framework in which this study takes place.

Syntactic representation of the eventualities Causation is a link between two eventualities [1], which can be represented in the text in several different syntactic ways. First, one of the eventualities may possibly be ellipsed or only alluded to but not clearly represented in the text, such as in example (1.1). There the consequence is represented as *fall*, but we know, of the causing eventuality, only that *she* is an agent in it. Causation is not otherwise represented in the sentence at all.

Second, eventualities can be represented as noun phrases, such as in (1.2), where the cause of the headache is the lack of sleep and both the cause and the consequence are represented in the sentence as noun phrases. The eventualities can also be represented as clauses such as in (1.3), where *my kitchen was lacking something* is a clause representing the cause and *I bought a toaster* is a clause representing the consequence. Finally, the eventualities can be represented as a mix of both, such as in (1.4) where the noun phrase *your sense of humour* represents the cause of the clause *she likes you*.

(1.1) She made him fall.

(1.2) Lack of sleep gave him a headache.

(1.3) My kitchen was lacking something, so I bought a toaster.

(1.4) She likes you because of your sense of humour.

In the present study, we will focus on causal relations holding between two eventualities both represented as clauses in the text.

[1] From here on, and in this entire thesis, we will use the term *eventuality* to describe any kind of world happening. We will use this term to enclose all aspectual classes that a predicate and its arguments can take, even if the eventuality has the aspectual class of a state such as in *Stéphane has long hair*.

Explicit versus implicit causation Causation can be represented in the text using an explicit causal marker such as *because* or *so*. These instances are relatively easy to recognise, using patterns that contain the marker (Khoo et al., 2000).

It can also be represented implicitly without the use of any marker, such as in (1.5). Example (1.5) shows an implicit causal relation with an anti-chronological order: *he was hungry* represents the cause of *John ate a double-cheese pizza*. In this thesis, we will focus on implicit, also called *unmarked*, causal relations.

(1.5) John ate a double-cheese pizza. He was very hungry

Binary classification task Causation can be considered as one of several discourse relations (see, Prasad et al., 2008). Much previous work exist on classifying these relations (Marcu and Echihabi, 2001; Sporleder and Lascarides, 2007; Pitler et al., 2009). Some classify less finely grained relations and recognise only *contingency*, which is a super-relation of causation and which also contains other types of relations, such as conditional relations. In this study, we will focus on the recognition of causation only, as a binary classification task that differentiates *causal* versus *non-causal* relations.

On the contrary, some authors have sought to classify causation further into sub-classes (Garcia, 1998; Prasad et al., 2007). We will not study systems that make use of such distinctions as *cause* versus *reason*, where the second is a psychological cause while the first must be independent of human motivations.

Most of this thesis will take place within the framework of machine learning, with a slight digression into rule-based systems in chapter 5. As we will see in chapter 3, causation recognition is a very subjective task and that, in certain cases, as we will see in chapter 5, occurrences are not clear cut between causal and non-causal and a continuum exists. In these circumstances, a statistical system is more appropriate than a rule based system that requires clear-cut judgements. Moreover, previous work on the recognition of implicit causal relations makes use of machine learning. For a classification task, machine learning programmes are able to learn regularities from many examples of items pertaining to each class and then to use these regularities to classify new, never previously seen items into their corresponding classes.

We will not focus on finding the exact boundaries of the text elements representing the cause and consequence eventualities. This has also been the case in previous work, where the items for classification were predetermined either as any pair of clauses from the text or as manually annotated text segments from evaluation corpora, where researchers testing the system have given it as input segments of texts that were hand-divided in the evaluation corpus. They have then measured the system's accuracy by determining whether the system classified the given segments into the correct hand-assigned class.

1.3 Research questions

In this thesis, we will first seek a working definition of the task that we aim to do automatically. To define causation, we will present annotation experiments that result in a coherent annotation manual that will serve as a work definition of causation. In doing so, we will show that the task is a very subjective one and leads to low inter-annotator agreement. We will argue that this low level of agreement should be taken into account when evaluating systems that aim at doing the task automatically, and we will propose an appropriate evaluation metric (chapter 3). From this perspective, we will investigate the following research questions:

Question 1 What are the defining characteristics of causation? Which characteristics of causation are associated with causation in human judgement?

Question 2 How can we take the subjectivity of a task into account when evaluating systems that carry it out automatically?

Few linguistic theories explain how humans solve the task of recognising implicit causal relations. The ones that do so rely mostly on the idea of *world knowledge*, as we will see in more detail in chapter 2. The general idea is that humans have access to an internal database of eventualities that are usually causally linked to each other and that they access this database to verify if an implicit statement might plausibly represent a causal relation.

The vast majority of previous computational linguistics work makes use of some representation of world knowledge. When doing so, one needs a com-

putational way to represent the eventualities that can be linked by causation. These eventualities have been represented as pairs of verbs (Pechsiri et al., 2006; Beamer and Girju, 2009), as pairs of words (Marcu and Echihabi, 2001; Pitler et al., 2009), as n-grams (Sporleder and Lascarides, 2007; Zhou et al., 2010), or as more complex features related to individual words (Pitler et al., 2009). In the light of previous linguistic work, we claim that this computational linguistics work relies on the implicit hypothesis that eventualities can be at least somewhat represented in the form of these features and that some regularities can be learnt on the training corpus so as to form a world knowledge base. In other words, the textual representations of eventualities can appear often enough for a system to be able to learn the causal link between pairs of these eventualities. In regard to these central features, the field suffers from a evident sparseness problem: the hand-annotated corpora are not sufficiently large for statistical regularities to be observed on word pairs.

Consequently, previous natural language processing work relies, at least partly, on the hypothesis that world knowledge is an essential feature of causation, and that it can be automatically learnt. We have no doubt that world knowledge is indeed a central feature of causation, as previous linguistic work has also argued, but we maintain that this world knowledge cannot be efficiently represented nor automatically learnt in the current state of the art of computational linguistics. Our main research questions are then the following:

Question 3 Are there statistical correlations between individual verb pairs and causation?

Question 4 What is the exact nature of the world knowledge necessary to recognise implicit causation? What would be necessary to represent and acquire it?

For question 3, we chose to investigate verb pairs in particular, rather than just any word pairs, because, since they are the central part of a predicate-argument structure, verbs are most likely to carry eventuality information. If verb pairs are good representations of eventualities and if some eventualities are *a priori* more likely to enter a causal relation than others, which is consistent with intuitions about world knowledge, then there

should be some statistical associations between specific verb pairs and the presence of causation. We will show in chapter 4 that this is not the case and we will argue in chapter 5 that verb pairs are not good representations of eventualities, which are much more complex.

To answer question 4, we will study the exact nature of the necessary world knowledge in chapter 5 by manually annotating a small corpus and systematically studying the exact world knowledge necessary to recognise each implicit causation occurrence. We will show that world knowledge is much more complex than lexical features used in previous work. We will present a tentative algorithm for the recognition of implicit causation based on world knowledge. We will then show, step by step, what is missing in the state of the art for automatically running such an algorithm. We will argue that, overall, the current state of the art in natural language processing is far from enabling the development of of this type of a system.

1.4 Thesis organisation

This thesis is organised as follows: Chapter 2 gives an overview of the state of the art in several domains and relates it to the present work. We will study linguistic work that explains how humans can recognise implicit causation. We will see that the two main frameworks that do so – the Segmented Discourse Representation Theory (Lascarides and Asher, 1993) and Relevance Theory (Wilson and Sperber, 2004) – both rely heavily on world knowledge (2.2).

We will then study previous work in computational linguistics that aim either at recognising causation or at classifying discourse relations (2.4). We will show that most previous work relies at least partially on lexical features that model world knowledge. We will also see that these tasks are still open problems and that no system has yet achieved an accuracy sufficient to make them usable in practice.

We will also give an overview of inter-annotator agreement measures, which we argue are representations of the intrinsic subjectivity of the task at hand, and we will review some previous work that attempts at taking this subjectivity into account when evaluating systems that aim at automatically reproducing the corresponding annotations (2.6).

In chapter 3, we will present experiments with human annotators that aim at reaching a working definition of causation in the form of an annotation manual. We will argue that the task, which results in low inter-annotator agreements, is highly subjective. We will propose to take this subjectivity into account when evaluating automatic systems and we will give a score that might be used to this effect (3.5).

Chapter 4 will show experiments aimed at determining whether specific verb pairs are good indicators of causation. We will argue that this is not the case.

Finally, chapter 5 will propose a manual analysis of causation occurrences whereby we will determine the exact nature of the world knowledge necessary to recognise causation. We will argue that this world knowledge is extremely difficult to automatically represent and acquire, and that the current computational linguistics state of the art does not allow such complex processing.

Chapter 2

Previous work in linguistics and natural language processing

2.1 Introduction

In this chapter, we will review the previous work that will be useful for this thesis. In particular, we will summarise linguistics papers that propose modelling of the way humans recognise implicit causation, and we will review previous work in computational linguistics.

For the linguistics papers (section 2.2), our goal will be to sum up state-of-the-art models of how humans resolve the task that this thesis will study from an automation point of view. The models that we will describe are complex and were not created with the aim of automatising the process, but we will see how these models can still be used as a basis for the development of automatic systems.

In section 2.3, we will give a brief introduction to the computational linguistics framework underlying this thesis. We will overview the subject of machine learning, which encompasses computer programmes that are able to learn rules by looking at examples of how to do a specific task. We will then describe the general concept of *classification tasks*, which are a set of tasks where a computer programme aims at classifying examples into pre-determined categories.

In section 2.4, we review previous work on the automatic extraction of causal relations and on classification of discourse relations, which is a task that includes recognising causal relations. We will show to which extent the models proposed by computer scientists draw inspiration from previous linguistics theories and we will summarise the results from the computational linguistics state of the art models.

Finally, in section 2.5, we will review how these automatic systems can be evaluated. We will refer to some previous papers on the annotation of reference corpora and we will describe how these corpora are traditionally used to evaluate the quality of natural language processing programmes.

2.2 Linguistics models for the human recognition of implicit causation in natural language texts

In this section, we will give an overview of linguistics theories that explain how humans can recognise implicit causation. We will show that the main feature used in these theories is the idea of *world knowledge*. In this thesis, we will argue that since world knowledge is the central feature of previous linguistics works, it is also necessary as a feature for systems aiming at doing the same task automatically.

Natural language texts can represent causation in several manners. Causation is a link between two eventualities. These eventualities can be represented in a text as noun phrases, as in (2.1) or as clauses such as in (2.2). In this work, we focus on eventualities represented as clauses. These clauses can be connected by a causal marker such as *because*, as it is the case in (2.2), or can be implicit as in (2.3). Since it is much easier to automatically extract marked examples such as (2.2), we will focus our research on unmarked ones such as (2.3). The question is then: how can we distinguish occurrences such as (2.3) from ones such as (2.4)? In the following sections, we will study linguistics models that explain how humans can know that (2.3) is causal, whereas (2.4) is not.

(2.1) The flood caused a landslide.

(2.2) I'm happy because I'm reading a good book.

(2.3) Max fell. John pushed him. (Lascarides and Asher, 1993)

(2.4) Max stood up. John greeted him. (Lascarides and Asher, 1993)

Many linguistics theories do not attempt to explain the difference between (2.3) and (2.4) (Saussure, 2000). Saussure's paper lies in the framework of temporal ordering, wherein the difference between (2.3) and (2.4) is only important because the temporal ordering of the sentences is not the same. (2.3) is interpreted as presenting the events in the anti-chronological order. In the world, *John pushes Max* first and second *Max falls*. The events appear in reverse order in the text, with *Max fell* appearing first. Conversely, (2.4) presents the events in the chronological order. This crucial distinction is central to the linguistic theories that explain causation. Causation needs to be explained, as its presence has the power to make the temporal interpretation opposite to the textual order of the events. This is an important property, since, as we will see later in this chapter, the chronological order interpretation – usually called *narration* – is the most frequent and is often considered to be the default temporal interpretation of an utterance.

In this framework, Saussure argues that many purely semantic theories do not aim to explain the difference between (2.3) and (2.4), because they rely mostly on aspectual markers such as verb tenses that make the time go forward or not. Saussure cites the examples of (Ter Meulen, 1997) and (Bohnemeyer, 1998) as such theories. Another possible reason why theories fail to explain this difference is that causation relations such as (2.3) are not completely symmetric to narration relations such as (2.4). Causation relations, indeed, typically occur between two events: the cause and the consequence, whereas narration relations can occur between many more events, and conjoined with elaboration relations, as we will see in section 2.2.1, can give rise to quite complex relation trees spanning many events.

The main recent theories that do explain this divergence are the *segmented discourse representation theory* (SDRT) (Lascarides and Asher, 1993), and *relevance theory* (Wilson and Sperber, 2004). Both rely mainly on world knowledge that the interpreter has about the possible causal relations between the eventualities. It has to be noted that this knowledge is only necessary in ambiguous cases such as (2.3) and (2.4), but not in explicit cases such as (2.2).

2.2.1 Segmented discourse representation theory

SDRT (Lascarides and Asher, 1993, Asher and Lascarides, 2003) describes and models the way relations between eventualities in a text can be calculated by human interpreters. It relies on several types of rules known and used by the interpreter. These rules can either relate to *world knowledge* or to *language knowledge*. Language knowledge is modelled as a set of rules that explain the way relations can be expressed in natural language. SDRT provides a logic that allows the interpreter to understand discourse relations. Given an utterance that represents an eventuality in a natural language text, SDRT logic allows the interpreter to decide which other utterances are related to it. It then allows the interpreter to determine the exact nature of these relations. A particularity of this logic is that it still works in cases of incompatible rules; in these cases, the logic provides conflict resolution.

According to SDRT, any text can be associated with a relation tree. The nodes of this tree are discourse units. The elementary units –or leaves of the tree– are clauses describing eventualities. Its edges are discourse relations. There are several possible discourse relations. The ones that interest us for the present work are *narration*, which is the relation inferred from (2.4) and *explanation*, which is similar to causation and is the one expressed in (2.3).

Following SDRT, the tree is built sequentially by the interpreter. The text is processed one clause at a time and a new clause is attached to the tree already built from the preceding text. For each new clause, the task can be divided into two subtasks: finding the previously treated discourse unit that the new clause is related to, and deciding the nature of the relation that exists between these two discourse units.

A new discourse unit can be related either to the directly preceding clause in the text, or to any clause that is an ancestor of the directly preceding clause and that is related to the previous clause with edges of the type *explanation* or *elaboration* only. The next examples are from the original Lascarides and Asher paper, but the analysis is partly ours:

(2.5) Guy experienced a lovely evening last night.

(2.6) He had a fantastic meal.

(2.7) He ate salmon.

(2.8) He devoured lots of cheese.

(2.9) He won a dancing competition.

To illustrate the first step of this algorithm –the choice of an attachment site– let us assume, for the previous example, that the interpreter already knows that (2.6) is attached to (2.5) by elaboration, and that (2.7) and (2.8) are attached to (2.6), also by elaboration. Interpreters will then look for an attachment for (2.9). They can take (2.8) into account as an attachment site because it is the previous discourse unit, or (2.6) or (2.5) because they are ancestors of (2.8), linked to it by elaboration relations. In this case, (2.9) should be attached to (2.5) with a relation labelled *elaboration*. *He won a dancing competition* is indeed an event that takes places in the context of the already mentioned eventuality *Guy experienced a lovely evening last night*, and is thus an elaboration of it. Utterance (2.9) should also be attached to 2.6 with a *narration* relation, which is the default relation, as nothing prevents it here.

The previous paragraph illustrated the specific problem of the choice of the attachment site. Actually, examples (2.5) to (2.6) could be linked in SDRT to a much more complex tree-like structure, such as the one proposed in figure 2.1. Here (2.6) and (2.9) are attached to each other by narration, but both are also elaborations of (2.5). Similarly (2.7) and (2.8) are linked by narration and are both elaborations of (2.6). Globally, SDRT proposes to parse a discourse and to associate a discourse tree to it, where utterances are linked by discourse relations. This is a far more complex task, and a superset of the subject of this thesis, which is the recognition of implicit causation between a pair of clauses.

Let us now analyse the SDRT explanation of the central problem of this thesis: the difference between (2.10) and (2.11).

(2.10) Max fell. John pushed him. (Lascarides and Asher, 1993)

(2.11) Max stood up. John greeted him. (Lascarides and Asher, 1993)

Here the attachment site is trivial, since there is only one previous sentence, but we need to find the correct relation. To do this task, we need some language knowledge and some world knowledge. The needed rules for 2.10 are the following. Rule 4 pertains to world knowledge.

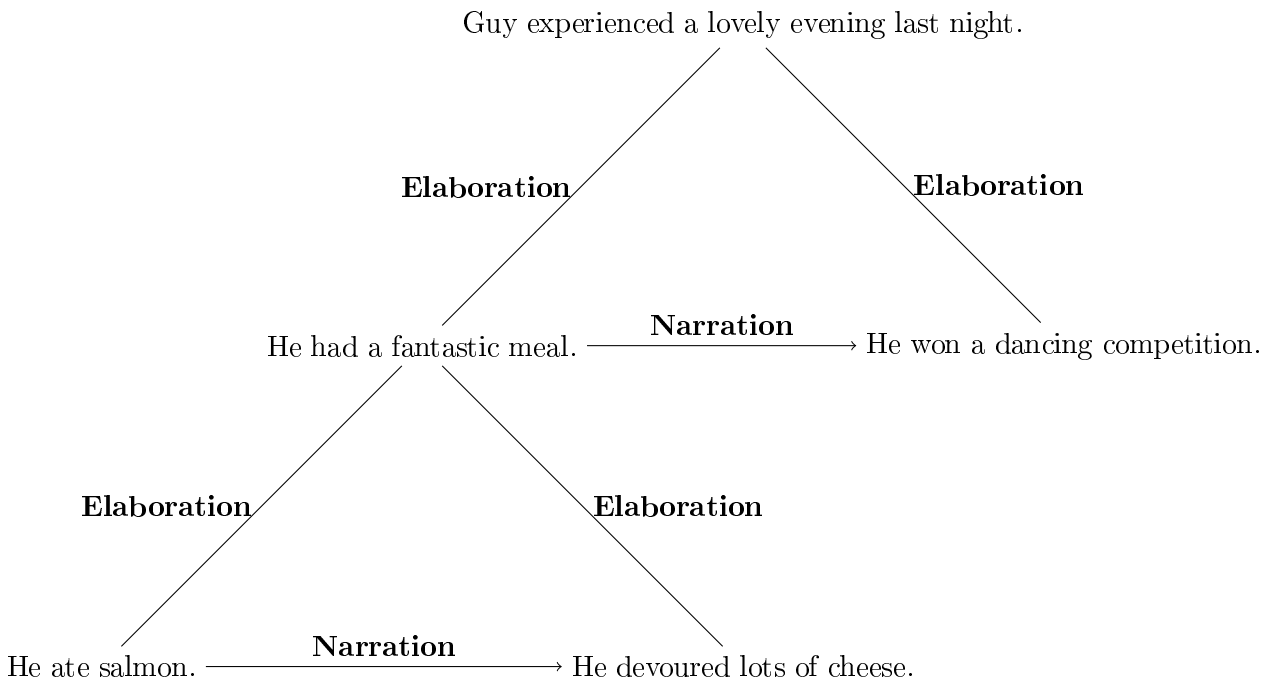


Figure 2.1: Example of an SDRT relation tree.

1. The default relation is narration.

2. In narration, the eventualities are chronologically ordered.

3. If narration holds, then the two discourse units must have a common topic.

4. When two eventualities are related by a discourse relation and one describes a falling(x) and the other describes a pushing(x, y), then the two eventualities are normally causally related, with the pushing causing the falling.

5. When one eventuality is the cause of the other, it then precedes it temporally.

6. When two eventualities have to be attached together, and the second one is the cause of the first one, then explanation is the relation between them.

Therefore, for (2.11), we must assume narration by default. The two units have a common topic: the encounter of Max and John. No other rule prevents them from being chronologically ordered, so we must link them using a *narration* relation.

For (2.10), however, narration still holds by default, but we also know that the two eventualities are normally causal and are in anti-chronological order. Since narration must present the eventualities in chronological order, we have a conflict of rules. Both rules are defeasible, so we have to choose one. In this case, SDRT asks that the most specific rule be followed. Since narration by default applies to any text, whereas the push-fall law applies only to occurrences of push and fall, the latter must be followed. We can then use the explanation rule, which states that causal eventualities in anti-chronological order must be linked to each other with an *explanation* relation.

2.2.2 Relevance theory

Relevance theory (Sperber and Wilson, 1995, Wilson and Sperber, 2004) is a general pragmatics theory. It aims at giving a modelling of pragmatics that is accurate in terms of cognitive psychology. It builds up on Grice's theory

(see, for example, Grice, 1975; Grice, 1989) but it also dissociates itself from many gricean modellings.

Basically, relevance theory relies on the notion of *relevance*, which is itself technically defined using the concepts of *effort* and *effect*. The effort is a measure of the difficulty of a cognitive task and how much cognitive effort it requires. Wilson and Sperber do not give a metric for effort, but they use this idea in order to compare different levels of effort. Although a task cannot be precisely measured for the effort it requires, it can be compared to other tasks on this aspect. It is thus possible to say that more cognitive effort is required for one task than for another.

A cognitive effect is a difference in the world representation held by somebody. For example, it can be an addition to their knowledge, a confirmation of something they already believed or an invalidation of a prior belief. A positive cognitive effect is a cognitive effect that makes the subject's world representation better than before; that is, a true information is added or a false information is removed. If the change results in a worse world representation, by the addition of false information, for example, the resulting effect is called a negative cognitive effect.

An input, in its cognitive treatment, is relevant if it maximises the positive cognitive effect it leads to, while minimising the cognitive effort it requires. The central point of Wilson and Sperber's theory is that a strong cognitive bias exists in favour of relevance. That is, when treating any input, people will treat it in a way that maximises relevance. Cognitive systems are biased in favour of seeking the maximum effect for the least effort.

Although the focus of their work is on language, Wilson and Sperber propose the maximisation of relevance as a general cognitive phenomenon, and not as a specifically linguistic system. An individual will treat any stimulus in a way that maximises relevance, but language has the specificity that it pertains to the domain of what Wilson and Sperber call *ostensive-inferential communication*. There, the source of the stimulus is another human being, and the assumption is that this communicator wants to communicate something. In this case, Wilson and Sperber state that the person making the utterance has an *informative intention*; that is, they want to communicate something. They also have a *communicative intention*; that is, an intention of informing the receiver of their informative intention. In this case, the

stimulus in itself, being intentional and having a communicative intention, can be assumed to have been made maximally relevant and can be treated in a way that optimises its relevance. To quote Wilson and Sperber :

> Every ostensive stimulus conveys a presumption of its own optimal relevance.

Relevance theory makes the hypothesis that language is used in a way that it is made maximally relevant by the communicator. Moreover, the communicator knows that the input will be treated in a way such that it maximises relevance for the interpreter. This central hypothesis makes possible the pragmatic level of language interpretation. In this framework, an utterance is treated by increasing the necessary cognitive effort until the cognitive effect is such that the interpreter's expectations of relevance are satisfied. Since communication is ostensive, the utterance can also be assumed to be the most relevant the communicator could give, given what they know and what they agree to communicate.

Relevance theory and implicit causation

Wilson and Sperber do not give a specific account of implicit causation in the relevance theory framework, but a basic account can be inferred from their treatment of other phenomena. For a more specific account of causation and temporal ordering in a relevance framework, we will present Moeschler's directional inferences model in the next section (2.2.2).

A basic account of implicit causation in the pure relevance theory framework can be deduced from (Wilson and Sperber, 1998) or (Carston, 1993). In these papers, the authors argue that, in a discourse, an utterance tends to achieve optimal relevance by answering a question raised by a previous utterance. Consequently, we can explain SDRT canonical examples (2.12) and (2.13) in the following way. In (2.12), the second utterance achieves optimal relevance by answering the question *Why did Max fall?*, whereas in (2.13), the second utterance is optimally relevant in answering the question *What happened next?* raised by the first utterance. Wilson and Sperber (1998) give example (2.14) to illustrate how relevance can be achieved when one utterance answers a question raised by the previous one. They contrast it with (2.15), where both clauses are part of one utterance and thus have to

be treated simultaneously, preventing the first one from raising a question that could be answered by the second one.

(2.12) Max fell. John pushed him.

(2.13) Max stood up. John greeted him.

(2.14) I ate somewhere nice yesterday. I ate at Macdonald's.

(2.15) I ate somewhere nice yesterday and I ate at Macdonald's.

The difference between (2.14) and (2.15) is more central to (Wilson and Sperber, 1998) and (Carston, 1993). But we claim that a similar treatment can be used to explain the difference between (2.12) and (2.13). In both cases, the first utterance can give rise to several questions, but world knowledge tells us that *John pushed him* is a good answer to the *why* question raised by *Max fell*, whereas *John greeted him* is a good answer to *what happens next?*. Carston also gives example (2.16) which gives a causal reading despite being contrary to world knowledge. Carston argues that humans have a strong tendency to seek causation, so that in the absence of appropriate world knowledge or of a more relevance achieving interpretation, causation, being highly relevant, is preferred (for more details, see also Carston, 2002).

(2.16) Max can't read; he's a linguist.

Directional inferrences

Moeschler (2000a) has given a specific model of the human evaluation of the temporal ordering of eventualities in discourse in the relevance theory framework: the *directional inferences* model. The determination of the temporal ordering of eventualities has been central in the development of SDRT, which we reviewed earlier in this chapter. The directional inferences model can thus be more explicitly compared to SDRT than the original relevance theory. It also gives a full account of the treatment of narration versus explanation cases.

The directional inferences model relies on the distinction between different types of information available to the interpreter when they process utterances. This information can pertain to linguistic knowledge or be of a more

encyclopedic or world knowledge nature. Furthermore, Moeschler introduces a distinction between *procedural* and *conceptual* information. Conceptual information is encoded in semantically full part of speech categories, whereas procedural information are encoded in closed categories and in morphological information such as verb tenses.

Some of these types of information are stronger than others, that is, if they give conflicting interpretation results with a weaker information type, they will prevail. Figure 2.2 represents the information hierarchy given in the original paper.

The directional inferences theory relies on the four leaf information type in the hierarchy contextual hypotheses, procedural clausal information, procedural morphological information and conceptual information. The theory also relies on the fact that some of these information types are stronger than others, and are the ones that will be followed for the temporal ordering of eventualities in case of a conflict. We will now review these types.

Contextual hypotheses are hypotheses that one could make regarding the eventualities, taking into account background information from the context. For example, if we know that Max and John are fighting, we could hypothesise that Max may push John to make him fall. Conversely, if we know that Max and John are friends walking on the border of a cliff, we could hypothesise that Max will push John away from the cliff to protect him after he falls. Contextual hypotheses are not linguistically encoded, contrary to the other three information types. They are a strong feature to determine the order of the eventualities, as contextual information is stronger that linguistically encoded information.

Procedural clausal information is information about the order of the eventualities that is signalled at the clause level. It typically takes the form of connectives. See, for example, the connectives in (2.20) and (2.21). Procedural clausal information is a strong feature for determining the order of the eventualities, as procedural information is stronger than conceptual information. This principle can be illustrated with example 2.17, which can be contrasted to 2.18. In these cases, we have a conceptual information associated to the push and fall events stating that pushing normally causes falling. In 2.18, nothing conflicts with this information and the sentence reads as the falling taking place before the pushing. In 2.17, however, a procedural

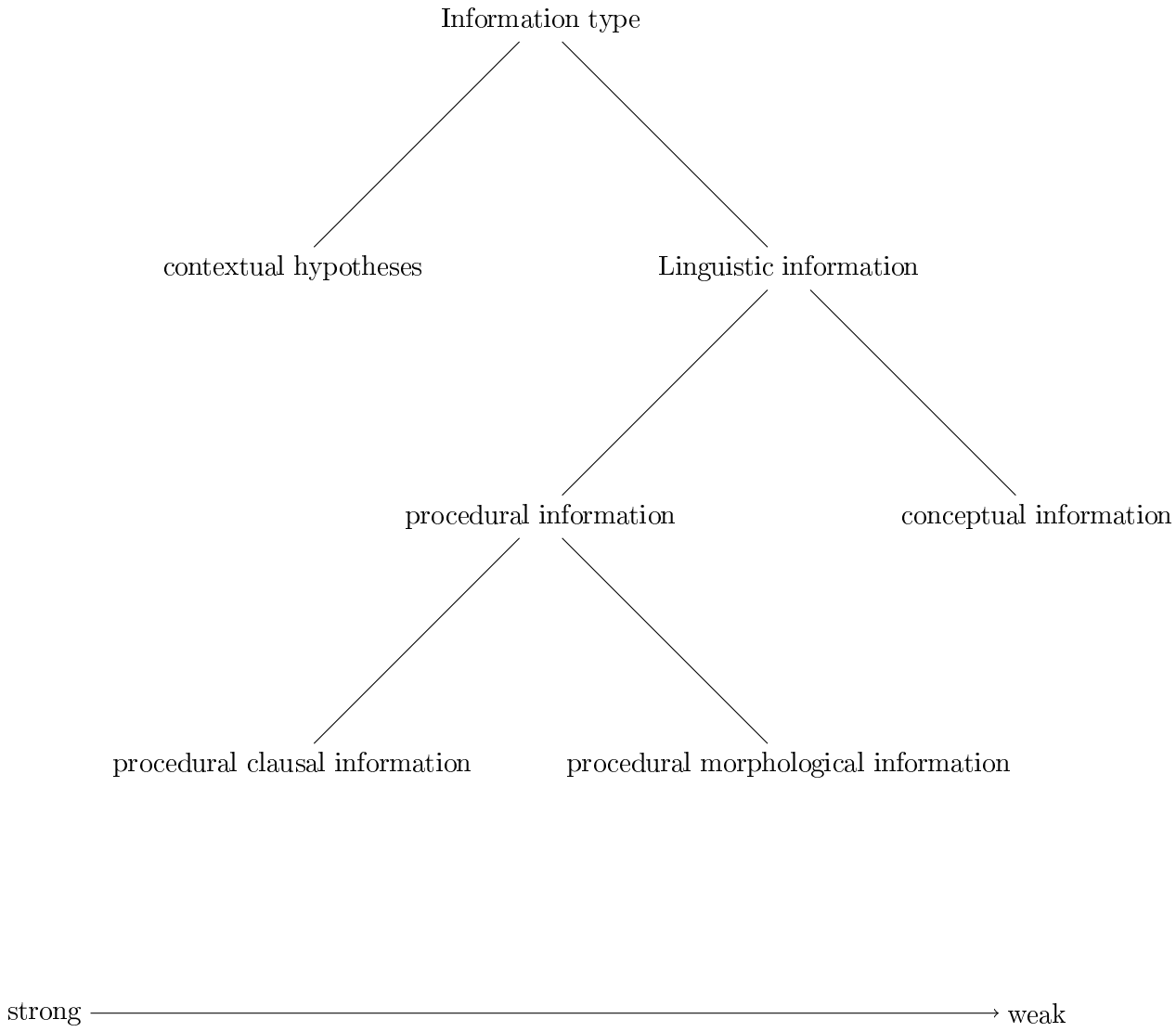


Figure 2.2: Directional inferences information type hierarchy.

clausal information in the form of the connective *so* prevents this reading and the falling is interpreted as happening before the pushing. Hence, procedural information is stronger than conceptual information.

(2.17) John fell, so Max pushed him.

(2.18) John fell, Max pushed him.

Similarly, procedural morphological information is information on the eventualities order coded at the word level, it is typically encoded as verb tenses such as in (2.22) and (2.23). Procedural morphological information is a weak feature for determining eventuality order because morphological information is weaker than clausal information, as can be seen in example 2.19 from Moeschler's original paper in French. Here, the utterance bears conflicting information as the use of the *passé simple* French tense gives a procedural morphological information of time going forward, while the presence of *parce que/because* bears a procedural clausal information of time going backward. The procedural clausal information is stronger and the utterance reads as the pushing happening before the falling. However, such instances, giving contradiction procedural information can be unnatural and clumsy.

(2.19) Jean tomba parce que Max le poussa. / *John fell (passé simple) because Max pushed (passé simple) him.*

(2.20) John fell because Max pushed him.

(2.21) John fell and then Max pushed him.

(2.22) John fell, Max had pushed him.

(2.23) John had fallen, Max pushed him.

Finally, conceptual information corresponds to SDRT world knowledge that is encoded in concepts or relevance theory's encyclopedic entries. It is a general knowledge about the way eventualities tend to interact with each other. The causal rule linking pushing and falling is an example of such predicate information. Another example would be script knowledge that

encodes the fact that, for example, when one goes to a restaurant, one first eats and then pays.

Moeschler (2000a) gives the following algorithm for using the information that we reviewed in the previous paragraphs to determine the order of eventualities E_1 and E_2.

1. Determine the direction of E_1 based on linguistics information, where strong features overweight weak features.
2. If possible, produce a contextual hypothesis.
3. Determine the direction of E_2 based on linguistics information, where strong features overweight weak features.
4. Determine the ordering of E_1 and E_2 based on the computed features.
5. Validate the order with a contextual hypothesis.

Let us use this algorithm for the SDRT explanation example *John fell, Max pushed him*. Because French makes use of procedural morphological information for this example, we will analyse the French version given in 2.24. In this case, *Jean tomba / John fell* only bears a procedural morphological information in the form of the tense of the verb, which leads to a weak forward ordering. *Max le poussa / Max pushed him* also bears a weak forward procedural morphological information in the verb tense. When taken together, the clauses lead to a weak backward predicate information due to the conceptual information given by the *push-fall* rule. In the absence of any strong information, such as a connective encoding of the temporal order of the clauses and of context, one will consider the default conceptual information– the push-fall rule– as plausible and make the contextual hypothesis that the falling was caused by the pushing. Being a strong feature, the contextual hypothesis outweighs the other types of information and the resulting analysis is in anti-chronological order.

(2.24) Jean tomba, Max le poussa. / *John fell (passé simple), Max pushed him (passé simple).*

2.2.3 Other theories of causation

In the next sections, we will review some other important theories causation that give less focus to the way an interpreter can distinguish between causal and non-causal implicit relations.

The paradox of causal complexity and Sanders' causality-by-default hypothesis

Sanders (2005) gives a coherence-based approach to causality, with a strong emphasis on cognition. The idea of coherence in linguistics is that interpreting a discourse requires the interpreter to understand a set of *coherence relations* that are associated with the discourse. These relations include the cause-consequence relation. Sanders and his colleagues (1992, 1993) give a complete taxonomy of these relations. The cognitive accuracy of his taxonomy is empirically validated by experiments with human subjects.

Sanders (2005) introduces the notion of the paradox of causal complexity. The paradox arises from the hypothesis that causal relations are more complex than simple additive relations. This hypothesis is justified by the fact that a causal relation such as `A cause B` also entails the corresponding additive relation `A and B` and thus contains more information than the additive relation alone. We could then expect causal relations to take more cognitive resources and thus more time to be processed. However, this not the case in empirical experiments with human subjects. Causal relations are actually faster to process than are additive relations. The contrast between an expected longer reading time and the actual faster reading time observed is called the paradox of causal complexity.

Sanders proposes to solve this paradox by introducing the hypothesis of *causality by default*. The idea is that interpreters will tend to interpret an utterance as causal because humans have a natural tendency to see patterns and structures rather than simple concatenations. He gives examples 2.25 and 2.26, which are intuitively interpreted as causal, although no causal rule seems to link their predicates. A more informative causal reading is thus preferred to the simpler additive reading, in this case. This tendency to look for the more informative relation first would explain the faster reading times. The additive relation is more cognitively costly, as the causal relation must

be ruled out first.

(2.25) Bill entered the room. Bob left.

(2.26) Bob left. Bill entered the room.

Sanders also proposes another solution to the paradox of causal complexity: the *schematic expectation hypothesis*. The idea here is that discourses tend to follow known schemas and that an interpreter will look for the normal schema continuation. For example, there is a problem-solution schema that will trigger an expectation for a causal relation. This approach is similar to the relevance theory approach, where a causal relation can achieve relevance with one of the utterances answering a question raised by the previous utterance.

Causation and inverse temporal order

Moeschler published several papers about the causation and temporal order and text order of the events. As we have seen in section 2.2.2, this order is central to the directional inferences model, as, the conceptual causation knowledge is used in this model to determine the temporal order of the events; that is, to determine whether the presentation of two events in the text follows their order in the world or is anti-chronological.

Moeschler (2000b) contrasts the two possible discourse orders and calls *narration* the temporal discourse order and *explanation* the anti-chronological order. Answering the question whether chronological order is the natural way to present events in a discourse, he argues that a reverse temporal order must be indicated, which is the case of *causal inversion*, in which a causal discourse is presented in a reverse temporal order. In this case, the inversion must be indicated either by temporally explicit verb tenses, or by the use of events that are normally causally connected by world knowledge.

Moeschler (2007) also notes that narration and explanation discourse are not symmetrical; that is, although their truth conditions are the same, they do not fulfil the same pragmatic and cognitive role. Indeed, not every temporally ordered discourse entails a causal relation. Some narration discourses only present temporally ordered events without any causal links between them. Causation is then of a different nature than temporal ordering. Moreover, causation is associated with *causal chains*. A causal chain

is a chain of direct causes and consequences. In these chains, the connected eventualities must alternate between the state and event aspectual classes. A causal discourse, however, may connect events that are not adjacent in the causal chain, and is then a case of indirect causation (see also Moeschler, 2003b, for more details on direct and indirect causation).

Finally, Moeschler (2010) argues that the inverse temporal order for causation is, at least in French, reflected by the fact that the only truly causal connective is *parce que/because*, which presents the eventualities in the anti-chronological order. In this paper, Moeschler shows that, contrary to *parce que*, other connectives in French lack the ability to represent causation between certain pairs of eventualities, depending on their aspectual class. This fact, together with a pilot study demonstrating that temporal order plays an important role in reading times of pairs of causally linked utterances, is clue that causation temporal ordering is cognitively motivated.

2.3 Computational linguistics framework: machine learning for categorisation tasks

Most of the computational linguistics work that we will study in the next sections pertains to the general framework of *supervised machine learning*. This framework provides a set of algorithms that allow a computer to learn rules based on examples. In the case of causation recognition or of discourse relation classification, which, as we will see later in section 2.4.2, is a super task of causal relation recognition, the task at hand is a classification task. That is, the programme should learn to classify items –in these cases, pairs of clauses– into several predefined classes. These classes are *causal* or *not causal* for the causation recognition task, and there is one class per discourse relation in the case of discourse relation classification.

Supervised classifiers learn how to classify items by looking at examples that are already correctly classified. A *learning corpus* is a set of correctly classified examples from which the classifier can learn. This corpus usually consists of a set of manually classified items. In computational linguistics, annotated corpora are usually used. In order to generalise from these examples, the classifier will take into account a set of predefined features – clues that can be used to assign the correct class to the example. Imagine, for

example, a classifier that classifies pictures of fruits into two classes: apples and oranges. The training corpus would be a set of pictures, each associated with the correct tag *apple* or *orange*. Classification features could be the shape of the fruit and its colour. When trained, the classifier will be able to classify as an orange or an apple a picture that it has never seen before.

In this work, we will make use of the *Boostexter* classifier, which is especially easy to use for text classification tasks (Schapire and Singer, 2000), as it can natively compute textual features, such as n-grams, on a segment of text. We will describe experiments with Boostexter in section 4.5.

Boostexter uses a boosting algorithm over one-level decision trees associated with single features, so it repeatedly does the following: compute a one level decision tree from a feature, and then integrate all decision trees into a classifier. One level decision trees simply state in which category an example should be classified, given a specific feature. In the case of the apples and oranges classifier, such a decision tree could state that *if the colour is green, then the label is apple*. These single feature decisions are not very accurate by themselves, but they are then combined into an accurate classifier. Basically, Boostexter uses the votes of each feature as a basis for recognising the item's class. A class is attributed to an item if most of its important features point to this class.

2.4 Previous work on the automatic recognition of causal relations

In this thesis, we will study the automatic recognition of causation. In this section, we will summarise previous work that aims at solving this task. We will show that although previous work on implicit causation recognition achieved results that are better than a random baseline, these results are still far from usable in other systems, in practice. We will also show that most of previous work relies in one way or another on a feature that broadly represents world knowledge.

Previous work on the automatic recognition of causation can be classified into two types. First, work that focuses on causation itself, and that we will present in section 2.4.1, and second, work that tackles the more general task of recognising discourse relations. These relations contain the explanation

relation, which is similar to causation, the contingency relation, which is a super-type of causation, or causation itself. We will present previous work on discourse relation classification in section 2.4.2.

Despite their formalism, linguistics theories cannot be straightforwardly implemented. Most of the clues that they rely on are computational linguistics open problems. Mainly, as we will develop in this whole thesis and specifically in chapter 5, we do not have access to a world knowledge database, and this knowledge is not easily represented nor acquired. As we will see later in this section, world knowledge has often been represented by previous authors as pairs of words. However, as we will elaborate in chapters 4 and 5, word pairs are both sparse and not sufficiently representative of the eventualities represented by the clauses.

2.4.1 Causal relation recognition

In this section, we will describe previous computational linguistics work on the recognition of causation by itself.

Authors of this previous work have to rely on a definition of causation. We will study definitions of causation in more detail and propose our own, in the form of an annotation manual, in chapter 3. Few of the previous papers in computer science mentioned a strong linguistics theoretical background. The authors mostly used expert annotations as a standard to which they would compare their extracted relations. One notable exception is the work of Garcia (1998). She extracted and classified causal relations using Talmy's force dynamic theory (1988a) as the background theory. Specifically, her sub-classification of the causal relations relied on this framework.

The eventualities that can be linked by a causal relation can be represented in the text mainly in two different syntactic ways: at the sentence level, as noun phrases and at the discourse level, as clauses. Previous work focused on the extraction of causal eventualities represented as noun phrases, such as in (2.27) (Garcia, 1998; Khoo et al., 2000; Girju and Moldovan, 2002; Girju, 2003). These relations can be retrieved using syntactic patterns such as (2.28).

(2.27) The findings caused a renew of interest in the domain.

(2.28) NP_1 causal-verb NP_2

Khoo, Chan and Niu (2000) also extracted causal relations at the discourse level, using the presence of connectives such as in (2.29). They also used patterns such as (2.30) for this task.

(2.29) The trains do not travel because there was a landslide on the tracks.

(2.30) *clause*1 connective *clause*2

More relevant to the present thesis is work that seeks to recognise implicit causation: causation that is not signalled by any marker such as connectives. This has been attempted in the general framework of causation recognition by (Pechsiri et al., 2006). Pechsiri's programme rely on the verb pairs feature, using a Bayesian classifier on an extremely domain restricted corpus in Thai (the Department of Agricultural Extension corpus). Their classifier uses the *WordNet* (Miller, 1995) hypernym of the head of both discourse units that they want to classify, as well as a causal marker, if there is any. They report 86% precision and 70% recall on their corpus.

More recently, Beamer and Girju (2009) sought to extract causation relations between verb pairs. Their work lies in a completely unsupervised framework. They introduce a statistical measure of causation: the *causal potential*. They compute this potential for verb pairs using a screenplay corpus in which action indications are temporally ordered in the text. This temporal ordering allows them to compute accurate temporal dependencies between verb pairs, and to compute the causal potential using these dependencies. They show that this potential, which resembles real-world measures such as the ones used in biology and relies on conditional probabilities of temporally ordered eventualities, is indeed correlated with the number of times the verb pair does appear in a causal relation in the text. Using only pairs of verbs in an unsupervised fashion, Beamer and Girju's work does not aim at maximal accuracy, but they propose to use their results as a baseline for supervised extraction of causal relations.

Finally, in a somewhat similar way, Riaz and Girju used topic models to extract the broader *contingency* relation (Riaz and Girju, 2010). Their system works on extremely domain-specific corpora (they tested it on news concerning hurricane Katarina and on news concerning the Irak war). Using topic models, they identify several topics in their corpus and then assign a topic to each clause in the corpus, using a bag of words approach. Clauses are

then clustered into eventualities, where clauses are considered to represent the same eventuality if they share their central verb, and if their subject and object arguments are similar. They then use eventuality co-occurrence statistics and conditional probabilities between the eventualities to compute a contingency score.

2.4.2 Discourse relation classification

Causation relations are part of or close to some discourse relations. Their automatic recognition is thus a super-task of causation recognition, which has spawned much recent computational linguistics work. This work was encouraged and facilitated by the release of two discourse annotated corpora, which can be used as an evaluation tool or, more generally, as a training and testing tool. These corpora are frequently used as a gold standard[1], and thus allow researchers to make more objective comparisons of discourse relation recognition systems.

These two large discourse annotation projects are the rhetorical structure theory corpus RSTC (Carlson et al., 2002) and the Penn discourse treebank PDTB (Prasad et al., 2008). The RSTC is annotated with rhetorical structure theory relations (Mann and Thompson, 1988). It covers 385 Wall Street Journal articles and is about 176 000 words long. The PDTB claims to be theory neutral and annotates three levels of relations from the broader *contingency* relation to its finer grained grandchild: the *reason* relation. A *cause* relation exists and is a mid level relation. PDTB spans the whole 1 million word Wall Street Journal corpus.

Since it is relatively easy to predict explicit discourse relations based on the presence of a connective or cue phrase, many authors have focused on the recognition of implicit discourse relations.

Several supervised approaches to the problem of discourse relation classification exist (Pitler et al., 2009; Zhou et al., 2010; Wang et al., 2010; Louis et al., 2010). All approaches use machine learning methods based on

[1] As we will study in more detail in section 2.6.1, a gold standard is a set of answers to a computational linguistics problem that are considered perfect and against which computer programmes are tested. In this case, utterances are manually annotated with their corresponding discourse relations. To test a computer programme, the raw corpus is given as input and the programme's found relations are compared to the correct hand annotated relations in the gold standard.

several features. A constant is the use of lexical features to represent world knowledge. Pitler and her colleagues notably introduced polarity tags, which state whether a word has a positive, negative or neutral connotation. They give the example of *popular* as a positive word and *oblivion* as a negative word. These tags can be used for sentiment analysis and should allow determination of whether two utterances carry parallel or contrastive sentiments. Pitler and her colleagues also introduced the use of Levin verb classes (Levin, 1993), and the use of a context feature, which states which relation immediately precedes or follows in the text the relation that must be classified. They used this feature only if the preceding or following relation was explicit. They used the PDTB and classified occurrences into the highest level relations.

Similar to many other authors, Pitler and her colleagues reported their results with two measures: f-score and accuracy. Accuracy is the simplest measure. It reports the total number of correctly classified items divided by the total number of items, as formally defined in the equation 2.31, where the result is multiplied by 100 to get a percentage. This equation does not take into account the individual classes and the score can be computed over all classes for a task requiring any number of classes.

$$\text{accuracy} \frac{\text{number of correctly classified items}}{\text{total number of items}} \times 100 \tag{2.31}$$

For some tasks, however, there is a need to not treat all incorrect results in the same way. Take information retrieval, for example. Information retrieval is the task that search engines, such as *Google*, perform. Given a search term, the task is to retrieve all documents that are about this term. This can be viewed as a classification task with two classes: relevant documents and irrelevant documents. In a big document collection, such as the web, one can assume that the number of relevant documents is much smaller than the total number of documents. Moreover, presenting an irrelevant document, called in this context a *false positive*, does not have the same implications as omitting a relevant document. A wrongly omitted document; that is, a relevant document presented as irrelevant, is called a *false negative*. In this context, it is important to report different measures taking into account the qualitative difference between mistakes arising from false positive and mistakes arising from false negative. Moreover, in this case,

accuracy is not a good indicator of quality because of the many more items pertaining to one class compared to the other. Imagine, for example, that the document collection contains 90% irrelevant documents and 10% relevant documents. A classifier that classifies all documents as irrelevant will get a $\frac{90}{100} \times 100 = 90\%$ accuracy –a good score– while being completely useless. Because accuracy is not a good measure in such cases, the computational linguistics community often reports other measures –precision and recall– instead of, or as well as, accuracy.

Precision gives an idea of how precise the system is. That is, in the case of information retrieval, given a retrieved document, it presents the likelihood that the document is indeed relevant to the search term. Precision is the number of relevant items that are correctly retrieved compared to the total number of retrieved documents. This measure takes false positives into account but ignores false negatives. Precision is formally defined in equation 2.32.

$$\text{precision} = \frac{\text{correctly retrieved documents}}{\text{total number of retrieved documents}} \times 100 \tag{2.32}$$

Conversely, recall takes into account false negatives but ignores false positives. Recall is a measure of how thorough the system is, of how many documents it misses. Given a relevant document, recall indicates how likely it is to not be omitted by the system. Recall is the number of correctly retrieved documents on the total number of documents that should have been retrieved. Recall is formally defined in 2.33.

$$\text{recall} = \frac{\text{correctly retrieved documents}}{\text{total number of relevant documents}} \times 100 \tag{2.33}$$

Finally, f-score is a measure that encompasses both precision and recall. It is a summary of both measures, and is often given as a single measure of the overall quality of a system. Mathematically, f-score is the harmonic mean of precision of recall. The harmonic mean is a mean that has the property of penalising large distances between the numbers used to calculate the mean. For example, the arithmetic mean (the mean usually used in everyday situations) of 5 and 5 is 5. The harmonic mean of 5 and 5 is also 5. However, while the arithmetic mean of 1 and 9 is still 5, their harmonic mean

is 1.8, reflecting the penalization for numbers that are far away from each other. In the case of the f score, intuitively, a system that is both average on precision and recall is better than a system that is very good in one measure and very poor in the other. F-score is formally defined in equation 2.34.

$$\text{fscore} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{2.34}$$

Pitler and her colleague evaluated their results separately for each class and reported both accuracy and F-score. They reported an F-score of 47.13% (67.3% accuracy) for implicit contingency relations (the relation that contains the *cause* sub-relation).

Zhou and colleagues (2010) decomposed the task of implicit discourse relation recognition into two subtasks: determining the best possible marker for the implicit relation and then using this marker, together with other features, to classify the occurrence into the correct relation. They used a language model based on trigrams for the first task and a supervised classifier for the second task. They classified occurrences into the highest level PDTB relations and reported an average 3% improvement when added to a one feature system, where the feature had been chosen from the features in (Pitler et al., 2009).

Wang and colleagues (2010) used syntactic features more precise than the ones from previous work. Their system classified occurrences into the four PDTB top level relations and achieved an overall accuracy of 40% for implicit relations.

Finally, Louis and her colleagues showed that entity features perform better than random at classifying top level PDTB relations. Their features captured several aspects of the entities (or arguments) participating in the eventualities to classify; for example, whether both eventualities have a common participant, what its syntactical role is and whether the entity is new in the text or has already been seen. They used a gold standard to determine some of the features. Their system did not perform as well as lexically rich systems but was better than a random baseline. A baseline is the result of a (often fictive) system to which the results of a new system are compared. When the field is mature enough and a gold standard and evaluation protocol have been defined and are used by all authors, one usually takes a previous system as baseline, with the goal for a new system being to be better than

older systems. For a younger unexplored field, one would often be satisfied, for a first try, with doing better than random. In this case, one compares one's results to the result of a fictive system that takes all decision randomly; that is, a random baseline.

Therefore, lexical features seem to be necessary for implicit relation recognition, but they are intrinsically sparse, making the task of a classifier especially difficult on relatively short annotated corpora. To reduce this sparseness problem, the trend has been to use unsupervised systems, where the classifier would be trained on examples generated automatically by removing the marker of explicit occurrences and would then be used to recognise implicit occurrences.

Marcu and Echihabi (2001) were the first to publish on the idea of automatically generating the training corpus from explicit occurrences. Their work relied on the hypothesis that their features would be similar enough in their occurrences in explicit and implicit examples to allow a classifier trained on explicit examples to recognise implicit ones. They used only one feature: pairs of words (one word per clause). Their system recognised the relation that the removed marker signalled with a pairwise accuracy between 64% and 89%. They calculated accuracies in distinguishing between two relations from their four relation set (contrast, cause-explanation-evidence, condition and elaboration), calculating accuracy for every possible pairing.

Sporleder and Lascarides (2007) did a similar work, using a much more complete feature set. They used lexical, syntactic, coherence and other features. They reported a 57.55 accuracy on their five relations, outperforming previous work by about 20%.

However, later work (Sporleder and Lascarides, 2008) showed that results were not as good as hoped when the systems were evaluated on native implicit examples as opposed to synthetic examples made from the removal of a marker. These unsatisfactory results were independent of the classifier or the feature set used. They showed that systems trained on marked examples did not generalise well to unmarked ones.

Sporleder (2007) then investigated features that are predictive of both implicit and explicit examples and showed that these features are few, occurrence dependent and tend to be lexical. She also showed that automatically selecting the explicit examples that most resemble the implicit ones does not

allow for a significant improvement of the system.

Pitler and her colleagues (2009) then showed that the good results obtained by word pair features in this framework might be due to an artefact caused by the evaluation of these systems on synthetic examples, as the absence of a marker (removed to make the example a synthetic implicit) was then a good predictor of the relation it originally marked.

Finally, Hernault and his colleagues (2010) used a semi supervised approach to decrease the feature sparseness on small data sets. They used feature co-occurrences in a general corpus to improve the supervised classification of infrequent discourse relations. Their approach led to an improvement of the training quality on small datasets.

2.5 Corpora annotated with implicit causal relations

In chapter 3, we will present experiments in annotating implicit causal discourse relations in French, and we will seek to find a working definition of causation in the form of an annotation manual. We will present this annotation manual that we tested with annotators and that allows for a high inter-annotator agreement between our educated predictions and the majority of the human annotators.

We discuss here a number of previous studies that influenced ours. To the best of our knowledge, no French corpus exists that is annotated for causal relations. However, an ambitious project of discourse relation annotation, the ANNODIS project, does exist, and should lead to useful and interesting results in the near future. In this section, we will discuss the annotation of the RST corpus, of the PDTB and the work of Inui (2005) who performed a work close to our own in Japanese by annotating a corpus with causal relations. We discuss this work in section 2.5.3. Finally, in section 2.5.4, we discuss the relations between our methodology and that of (Hovy et al., 2006), which inspired the work we will present in chapter 3.

2.5.1 The rhetorical structure theory corpus and the Penn discourse treebank

In English, an important corpus exists that is annotated with syntactic trees (Marcus et al., 1994). This is the corpus annotated by Carlson and colleagues (2002) with relations from the Rhetoric Structure Theory (RST). One of these relations is the *cause* relation, which is subdivided into three sub-categories: *cause*, *result* and *consequence*. The difference between *cause* and *result* is the importance given by the text to the consquence or to the cause. The *consequence* relation is distinguished from other relations by the fact that it represents a more or less direct causal link. We consider the relations annotated by Carlson and colleagues as *reason* to be a type of causal relation. It is distinguished from the others as it features a consequence realised by an animated agent. The annotation instructions of Carlson and his colleagues do not provide a definition of the concept of cause. Nevertheless, annotators are requested to always choose the less general relation possible, so that, for instance, a causal and temporal relation must be annotated as causal (more precise) and not as temporal (more general).

Similarly, the Penn discourse treebank annotates an exhaustive set of discourse relations. It claims to be theory neutral and it implements a *cause* relation that exactly matches ours. We will use this corpus in various experiments in chapter 4.

Compared with the work of these previous authors, the work that we will present in chapter 3 is more specific. Indeed, we only aim at precisely identifying causal relations. We do not try for a finer analysis of causal relations by subdividing them into sub-categories, as we do not distinguish, for instance, between causes and reasons. However, we attempt to obtain a much more detailed manual that allows for a very precise discrimination between causal and non-causal statements.

2.5.2 The ANNODIS project

The ANNODIS project (Péry-Woodley et al., 2009) aims at constructing a resource similar to the Penn discourse treebank or the rhetorical structure theory corpus in French. Similarly to the Penn discourse treebank, it aims at annotating theory-neutral discourse structures. Its relation set is derived

form a large number of linguistics theories such as SDRT.

ANNODIS presents two important differences from the equivalent English corpora. First, the corpus to be annotated is a handmade by the ANNODIS researchers. They do not, contrary to the English corpora, make use of an existing syntactically annotated corpus. This has the important advantage of allowing them to choose a very eclectic set of text genres, allowing their annotations to much more genre neutral, contrary to the English initiatives that make use of a very specific journalistic corpus.

Second, the ANNODIS annotation strategy makes use of the innovative concept of *ascendant and descendant* annotation strategy. Traditionally, these types of annotations are done ascendantly. That is, elementary discourse units are identified and then pairs of them are linked by specific relations, forming a new unit unit that can, in turn, be linked to existing units, creating a discourse tree. ANNODIS proposes to use this strategy simultaneously with a *descending* strategy, where discontinuities are first sought in the text, allowing for a high level text segmentation and then descending from these maximal discourse segments.

As of the writing of this thesis, the ANNODIS project has not yet produced an annotated corpus, but such a resource should be released in the near future.

2.5.3 Inui

Inui (2005) annotated a Japanese corpus with causal relations. His work is closer to our own than to that of Carslon and colleagues, as he only annotated causal relations. Inui's corpus is constituted of 750 newspaper articles dealing with social matters. The causal relations that he annotates can be signalled by a causal marker or they can be implicit. Inui not only annotated causal relations between clauses, but also between noun phrases, as in (2.35). To precisely identity these relations, Inui used a set of linguistic tests. Annotators had to try these tests on the extracts that they analysed. If one of the tests produced a sentence that was both semantically and syntactically correct, the extract was annotated as causal. Otherwise, the annotators had to try the other tests and consider those extracts that passed none of the linguistic tests as non-causal. In addition to annotating extracts as causal or non-causal, Inui added a label to quantify the degree of necessity of the an-

notated causation. This label indicated whether the causal relation between the two statements is usually present.

(2.35) The accident caused a traffic jam.

Our work differs from Inui's in three respects. Firstly, our interest lay in the causal relations between eventualities represented by statements, and not in noun phrases. Secondly, we only tried and annotated causation, and did not concern ourselves with its degree of necessity. Eventually, we concerned ourselves more precisely with the annotation instructions for which we used not only linguistic tests, which can be ambiguous, but also other characteristics of causation such as temporal order, counterfactuality or asymmetry.

2.5.4 Hovy and colleagues

Hovy and his colleagues (2006) completed a large-scale annotation on word sense disambiguation on the Penn TreeBank (Marcus et al., 1994) corpus. Their aim was to link every word to its meaning in an ontology. This task could provide a low inter-annotator agreement. Hovy and colleagues had set their aim on obtaining an inter-annotator agreement over 90%. Therefore, they used an iterative system in which instructions were tested in 50 sentence samples and modified until a high agreement was obtained. Modifications in instructions were not only clarifications, as in the work we will present in chapter 3. It also aimed to define groups of optimal meanings to which words could be attached. Hence, for each iteration, instructions were clarified, and groups of meanings were modified until agreement was sufficiently high. Then, Hovy and colleagues matched their groups of meanings to meanings of an ontology and these meanings were then annotated. An important difference between this work and ours is that we did not adapt the task, we only attempted to clarify it.

2.6 Evaluation

In chapter 3, we will argue that, for subjective tasks such as causation recognition, it is important to take the subjectivity of the task, as measured by the inter-annotator agreement, into account in order to evaluate automatic

systems that aim at solving the task. We will propose an evaluation metric that does so.

Evaluating automatic systems that perform natural language processing part is an important task of the computational linguistics community. The goal of evaluating such systems is not only to measure their general accuracy, but also to compare them to other systems. A common evaluation framework for a specific task allows authors to easily compare their systems to the current state of the art, by simply looking at previous authors' results on the evaluation test set, without the need for reimplementing the state-of-the-art systems.

The need for a common evaluation has given rise to a number of hand-annotated evaluation corpora –gold standards– for specific natural language processing tasks. Each system can then be evaluated by comparing its output to the gold standard. The Linguistic Data Consortium [2], for example, regularly releases gold standards for computational linguistics tasks; one famous example is the Penn TreeBank (Marcus et al., 1994), which is widely used for parsing evaluation. Several important tasks have also given rise to evaluation campaigns where a task and an evaluation framework are defined, and evaluated systems are presented in an specific conference or workshop. This is the case, for example, for the TREC [3] conference for information retrieval or for SENSEVAL [4], for word sense disambiguation.

As we will study in more detail in chapter 3, recognising causation is a somewhat subjective task. Humans do not always agree on what is causal or not in the absence of an explicit marker. In this section, we will study how subjective tasks are evaluated using an annotated corpus, we will review how the subjectivity of such corpora can be evaluated and we will summarise some previous work on taking this subjectivity into account when evaluating automatic systems. Later in this thesis, in section 3.5, we will propose our own measure that takes the inter-annotator agreement of the gold standard into account when evaluating a system that attempts to perform the task automatically.

[2] `http://www.ldc.upenn.edu/`
[3] `http://trec.nist.gov/`
[4] `http://www.senseval.org/`

2.6.1 Gold standard and the problem of subjectivity and low inter-annotator agreements

The first step of developing any natural language processing tool is to define precisely the task at hand. This helps in the development of the system, as one needs to know one's exact goal, but it also aids in evaluating the final system and comparing it to its competitors. Providing exact rules on what the system should output for each possible input is typically very difficult, so the tasks are often defined by a set of example inputs together with their correct associated outputs: a *gold standard*. In the development stage of a system, one takes into account only a subset of these examples. When the system is ready, it is tested on examples that were never used in the development phase. The number of common answers in the system's output and the gold standard gives a measure of the quality of the system.

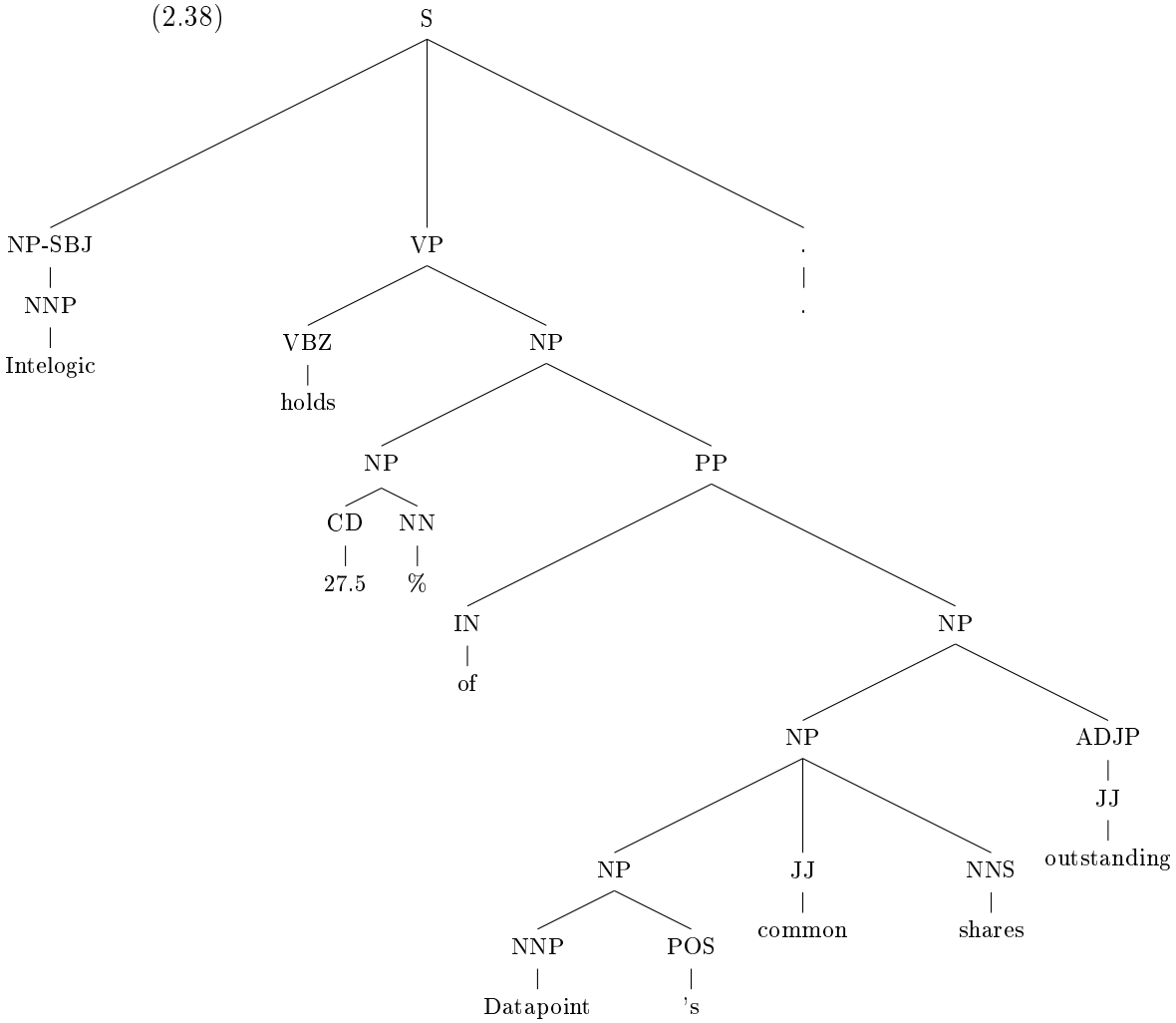
The gold standard is defined by human judges and, for natural language processing tasks, typically takes the form of an annotated corpus. In parsing, for example, the input is a sentence in natural language, and the output is a syntactic tree representing the sentence. A famous gold standard for parsing is the Penn Treebank corpus (Marcus et al., 1994), which is based on an English corpus, mainly from newspaper texts, that has been manually annotated with the corresponding parse trees. For example, the sentence (2.36) is annotated with the output (2.37) corresponding to the parse tree shown in (2.38).

(2.36) Intelogic holds 27.5% of Datapoint's common shares outstanding.

(2.37)
```
( (S
    (NP-SBJ (NNP Intelogic) )
    (VP (VBZ holds)
      (NP
        (NP (CD 27.5) (NN %) )
        (PP (IN of)
          (NP
            (NP
              (NP (NNP Datapoint) (POS 's) )
              (JJ common) (NNS shares) )
            (ADJP (JJ outstanding) )))))
```

(. .)))

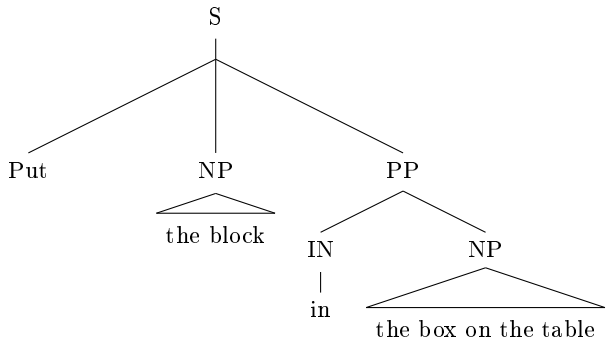
(2.38)



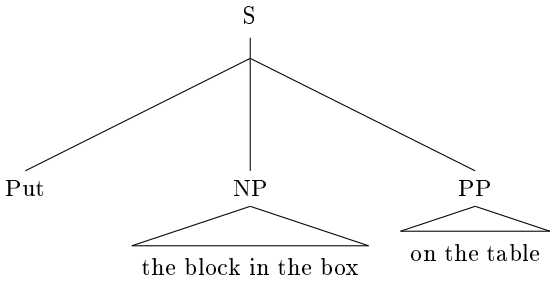
The gold standard is considered to be perfect, and an automatic system should reproduce the human annotations as closely as possible. However, the tasks are often intrinsically complex and subjective. Church and Patil (1982) showed, for example, that parsing is very ambiguous. They gave example (2.39) as an ambiguous sentence. The sentence (2.39) can be parsed with the two different groupings: (2.40), with the meaning that the box is on the table and (2.41), with the meaning that the block is in the box. In the absence of a disambiguating context, both parse trees are possible and the answer is subjective.

(2.39) Put the block in the box on the table.

(2.40)



(2.41)



Similarly, in the task of word sense disambiguation, human judges often differ. Word sense disambiguation is the task of assigning a specific meaning to a word that can have several meanings. For example, the word *bass* in the examples (2.42) and (2.43) from (Navigli, 2009) do not have the same meaning. *Bass* in (2.42) refers to a sound, while in (2.43), it refers to a fish.

(2.42) I can hear bass sounds.

(2.43) They like grilled bass.

However, while (2.42) and (2.43) are fairly clear, annotation efforts, using the senses described in *WordNet* (Miller, 1995), have led to low inter-annotator agreements (Snyder and Palmer, 2004). Snyder and Palmer cite the case of the word *national*, which has been very problematic, with one annotator consistently giving preference to sense two: *limited to or in the interest of a particular nation* and the other one choosing sense three: *concerned with or applicable to or belonging to an entire nation or country.* Snyder and Palmer note that the difference relies in sense two contrasting with *international* and sense three contrasting with *local.*

In such cases, human judgements differ, and annotators do not always agree. How much humans disagree on a specific task is a measure of the difficulty or subjectivity of the task. A low inter-annotator agreement indicates

either a difficult task or a subjective one. In the case of natural language processing, the goal for the system is to mimic human results on a task. As we do not aim for computers to solve the task better than humans, and we do not even know if this might be possible for any natural language processing task, we consider humans as perfect systems, and their judgements as correct answers. In this case, a low inter-annotator agreement is a mark of subjectivity, a measure of the divergences that are acceptable between different systems that do correctly solve the problem.

2.6.2 Creating a gold standard

Although natural language processing tasks almost always contain some amount of subjectivity, gold standards typically mention only one possible and supposedly correct output for each input. Releasers produce only one set of annotations, although they often make use of several annotators. In such cases, it is necessary to merge different and possibly divergent annotations into one gold standard.

Large annotation projects often make use of several annotators per example, and then of other annotators to adjudicate the disagreement cases. This is the case, for example, of the *Penn Discourse Treebank* (Miltsakaki et al., 2004) and the OntoNotes corpus (Hovy et al., 2006).

The Penn Discourse Treebank is a project of discourse annotation of parts of the Penn Treebank. The annotations mark the discourse relations that hold between segments of the corpus. Examples of discourse relations are causation such as in (2.44) (which is also an explicit relation, marked by *because*) or contrast such as in (2.45) (which is implicit).

(2.44) Typically, money-fund yields beat comparable short-term investments because portfolio managers can vary maturities and go after the highest rates.

(2.45) A figure above 50 indicates the economy is likely to expand; one below 50 indicates a contraction may be ahead.

The Penn Discourse Treebank releasers used the following annotation scheme: each example is first annotated by two persons, and diverging ex-

amples are then adjudicated by a four others, working as a team (Prasad et al., 2008).

OntoNotes is a corpus of word sense disambiguation. Each word of a subset of the Penn Treebank is linked to its sense in an ontology. As Snyder and Palmer (2004) have shown, the task leads to a very low inter-annotator agreement, indicating a need to redefine the task to get a higher agreement. The exact necessary granularity of senses is not a given and it is possible to merge close senses together and to redefine them prior to annotation. Hovy and his colleagues (2006) used a cycle methodology, calculating inter-annotator agreement on a sample of the corpus and then defining and clarifying the senses until they reach a sufficient agreement. The release of ontoNotes consists of annotations done in parallel and independently by two annotators and then adjudicated by one.

Other projects have released examples that were annotated by only one person, not necessarily the same person for each example, but also released agreement measures calculated on a sample of the corpus. This is the case for the Rhetorical Structure Theory corpus (Carlson et al., 2002), another discourse relation corpus, and for the Timebank corpus (Pustejovsky et al., 2003; Pustejovsky et al., 2006), which annotates the temporal order of eventualities.

The Penn Treebank follows a mixed scheme (Marcus et al., 1994). The annotation is first done automatically and then corrected twice by different annotators. Here again, Marcus and his colleagues conducted an experiment using multiple annotators' annotations of a sample of the corpus in order to give a measure of inter-annotator agreement.

Releasers of this types of corpora provide a measure of the inter-annotator agreement, either in the form of raw agreement or in the form of a more precise agreement coefficient such as Cohen's κ score (Cohen, 1960), which takes into account the fact that random annotations still lead to some raw agreement. The corpus builders strive for a good agreement and corpora with very poor agreement might be discarded. Nevertheless, a degree of subjectivity always exists for natural language processing tasks, which is reflected in the inter-annotator agreement.

2.6.3 Inter-annotator agreement measures

In this section, we will present several measures that have been considered suitable for reporting inter-annotator agreement on gold standard corpora, as well as measures that were actually used by several large annotated corpora initiatives. As we will see, the measures used in practice are sometimes not the ones recommended by the authors of the theoretical work.

Raw agreement

Raw agreement is the simplest measure of inter-annotator agreement. It consists of giving the percentage of items that annotators have agreed on, as in equation 2.46, where a is the number of examples on which annotators agree and d is the number of examples on which annotators disagree.

$$\text{ra} = \frac{a}{a+d} \times 100 \tag{2.46}$$

This measure is simple and is often the one given by corpora releasers. However, it has a major drawback: it is dependent on the distribution of classes in the specific corpus studied and thus it does not allow for a good evaluation and comparison of the subjectivity of the annotation tasks. If the corpus is highly skewed in favour of a class –say for example, a binary classification task where one class appears about 90% of the time– then even if annotation is completely random, if the annotators follow the proportion of each class, they will reach a much better raw agreement than if both classes had a 50% chance of appearing. This measure also reports better agreements when the classification task contains fewer classes. If each annotator randomly selects a class, they will find agreement more often if there are few classes than many. For some critics of the raw agreement score, see for example (Artstein and Poesio, 2008; Carletta, 1996).

Authors of the Penn treebank give a very similar measure for the tagging task: the raw disagreement, which is the number of disagreed examples out of the total number of examples, which is $100 - \text{ra}$ (Marcus et al., 1994). Similarly, the Penn discourse treebank authors give raw agreement measures. They argue that the κ, which has been more frequently recommended and which we will study later in this section, is appropriate only for discrete classes task, which is not the case of their span selection task (Miltsakaki et

al., 2004). The SALSA project –a German frame project– also reports raw agreement measures for similar reasons (Burchardt et al., 2006).

Coefficients

As we have seen in the previous section, the raw agreement measure, while straightforward, has important drawbacks. In particular it does not take into account the chance agreement, making the measure task dependent. Therefore, some tasks lead to better agreement measures because of their structure in terms of number of classes, for example, and not due to an intrinsic level of objectivity associated with them.

Several coefficients aim to normalise the the inter-annotator agreement by taking the chance agreement into account. Artstein and Poesio (2008) gave a thorough survey of such measures and their use in computational linguistics on which this section is based. Basically, measures differ by the number of classes and of annotators with which they can natively work (that is, without having to average the measures over all annotator pairs or over each class), and in the way they model chance agreement. Two famous coefficients are Cohen's κ[5] (Cohen, 1960) and Siegel and Castellan's K[6] (Siegel and Castellan Jr, 1988). Although both measures share a similar name, they are actually not the same and can give different results with the same data. Siegel and Castellan's K is a multi annotator score based on Scott's π (Scott, 1955). Both κ and π (pi) are normalisations of the raw agreement by an estimation of the chance agreement, using the following formula.

$$\frac{\text{observed\_agreement} - \text{chance\_agreement}}{1 - \text{chance\_agreement}} \tag{2.47}$$

The estimated chance agreement is subtracted from the actual agreement and the value is then renormalised on the accessible range; that is, the range going from chance agreement to one. The two measures only differ on the way the chance agreement is modelled. Both measure make use of the proportion of examples assigned to each class, but Scott's π assumes that each annotator

[5] lower case kappa
[6] upper case kappa

annotates samples using the same distribution of probability, while κ uses a different distribution for each annotator.

Cohen's κ has been used in several previous computational linguistics papers, see (Véronis, 1998; Wiebe et al., 1999), for example. Siegel and Castellan's K has also been widely used in the computational linguistics community in (Marcu et al., 1999; Barzilay and McKeown, 2001; Palmer et al., 2005), for example. Closer to the focus of this thesis, releasers of the Rhetorical Structure Theory corpus (Carlson et al., 2002) give both pairwise and general Siegel and Castellan's K scores.

2.6.4 Evaluation that takes into account the inter-annotator agreement of the gold standard.

In section 3.5, we will propose a way to evaluate automatic system that takes into account the subjectivity of the task reflected in the inter-annotator agreement.

The problem of subjective annotations is not new. (Reidsma and op den Akker, 2008), for example, analyse how machine learning can exploit such subjective annotations. However, to the best of our knowledge, only (Vieira, 2002) proposes to take this subjectivity into account when evaluating automatic systems. Vieira proposes to evaluate automatic systems using the same coefficient measure used to evaluate the inter-annotator agreement of the corresponding gold standard. The gold standard inter-annotator agreement then constitutes a maximum above which an automatic system is not expected to perform. In section 3.5, we will go one step further and propose a statistical test to compare the automatic system's results to the inter-annotator agreement scores of the gold standard.

2.7 Conclusion

In this chapter, we described linguistics theories that model the recognition of implicit causation by humans as well as previous computational linguistics work that aimss at doing so automatically.

We showed that previous linguistics theories rely mostly on world knowledge to recognise causation, and that this world knowledge is modelled in

computational linguistics work as lexical features, mainly as pairs of verbs or of words.

However, previous computational linguistics work on the recognition of implicit causation, be it their main task, or a subtask of discourse relation classification, does not give satisfactory results, although many features have been proven to predict causation better than a random baseline does. The task is far from being solved. Numbers such as a 47% f-score or 40% accuracy on four classes are far from being sufficient to integrate these systems into broader programmes.

Since manually annotated corpora are extremely costly to produce, and are thus short, and since world knowledge is the primary feature of causation and is straightforwardly represented as a lexical feature that is intrinsically sparse eventualities in a large corpus, the field suffers from an important sparseness problem. Attempts at overcoming this problem by using automatically labelled examples from big unannotated corpora have been unsatisfactory when evaluated on the real problem; that is, native implicit examples.

Previous computational linguistics work takes little account of previous linguistics theories, and although authors have presented thorough analysis on the which features help the task or not (Sporleder and Lascarides, 2008; Pitler et al., 2009), there is no explanation of exactly what the field is lacking to actually achieve satisfactory and usable results, perhaps because previous authors have focused on achieving quantitatively better results than the ones reported in the papers preceding them in the field.

Chapter 3

Human recognition of causation: defining objective criteria for a subjective task[1]

3.1 Introduction

In the present chapter, we will present work aimed at achieving a definition of what we consider to be causal within the framework of this thesis. The definition should allow us to calibrate and evaluate the quality of the programmes to be developed for automated extraction of these relations.

Causation has been abundantly studied in previous work stemming from several fields of research, such as cognitive sciences, philosophy, linguistics, physics and Law. Over two centuries have passed since the seminal work of Hume (1739), and nevertheless many questions remain unsolved as to the very nature of causation. Amongst many examples, we could mention a recently held virtual conference organised by `Interdisciplines.org` (2005) that tackled causation from the perspective of the diverse approaches of philosophy, cognitive sciences and social sciences. In linguistics, causation has played an important role in the work of such diverse authors as Talmy (1988b) for semantics or Sanders (2005) for pragmatics.

In spite of these many studies, finding consensual criteria that allow

[1] A previous version of this chapter was written in French for my *mémoire de pré-doctorat*.

easy identification of text segments representing causal relations remains a difficult task and, as we shall see below, human judgements conflict in many instances.

For example, without context, it is impossible to determine whether utterance (3.1) represents a causal relation. Simply reformulating the sentence with an explicit causal marker, such as in (3.2), does not suffice to disambiguate, as it remains difficult to determine whether sentence (3.2) is synonymous to (3.1). Even with a less extreme example, such as (3.3), judgement remains difficult, as one could consider that breaking one's leg does not stem from skiing, but from an eventuality such as colliding or falling, for which skiing is a mere contextual background. Finally, naive annotators have conflicting judgement for (3.4), which is considered non-causal by experts.

(3.1) John's ear fell off, he took a shower.

(3.2) The falling of John's ear was caused by his taking a shower.

(3.3) He broke his leg while skiing.

(3.4) This is a triangle, it has three edges.

As far as we know, no French annotated corpus exists for causal relations. Such a corpus would be useful for evaluation of programmes that attempt to automatically extract these relations. In this chapter, we will describe a method that provides a better definition for causation from its characteristics. We detail a set of annotation rules that translates our vision of causation and allows annotators to precisely identify causation in a corpus. In particular, our instructions manage to solve ambiguities in several problematic cases. This work was necessary, necessary because, to the best of our knowledge, no necessary and sufficient conditions exist that would easily allow an annotator to determine whether a text extract is causal or not. This manual may serve as the basis for the design and evaluation of an automated annotation programme.

Causation is a relation that exists between two eventualities. These eventualities may be realised in the text as either noun phrases, such as in (3.5), or by clauses, such as in (3.6). Within the scope of this work, we will concern ourselves only with causal relations between clauses. We further focus on

implicit relations that contain no causal markers, such as *because*, such as in (3.6).

(3.5) The earthquake has caused a landslide.

(3.6) The glass is broken, it has fallen from the table.

Our annotation instructions must fulfil two requirements: first, they have to be sufficiently clear for annotators to coherently identify causal relations. Secondly, they must match expert judgement on causation, in order to guarantee that causation is indeed being identified, rather than some different relation. In this chapter, we will test three hypotheses:

Hyp1 Intuitive characteristics of causation exist, and they can be consciously accessed by an individual for reasoning about causation.

Hyp2 A number of characteristics are correlated with human causation judgements.

Hyp3 Based on these characteristics, our manual allows for a coherent annotation of causation.

Hypothesis Hyp1 is partially invalidated by an experiment that elicits the characteristics in question in section 3.2. We test hypothesis Hyp2 using an annotation experiment for causation and its characteristics in section 3.3 and we check coherence and clarity of instructions using inter-annotator agreement in 3.4.7.

We provide the following results:

- We describe an experiment that allows elicitation of the characteristics of causation that are consciously used for reasoning (3.2).
- We test correlations between causation and its most notable characteristics in human annotation (3.3).
- We describe a methodology to refine and test our manual and detail the rules that result (3.4).
- We show that our annotating instructions are coherent since they allow us to reach a very good agreement between annotations resulting from

a voting scheme on several annotators and our predictions ($\kappa = 0.84$), and they lead to a good identification of causation that matches that of experts (3.4.7).

- Finally, we propose an evaluation metric that allows us to take into account the subjectivity of a task when evaluating an automatic system that aims at solving it.

3.2 Intuitive characteristics

In this section, we will describe an experiment that allows us to obtain a list of intuitive characteristics for causation. The aim of this experiment is to determine whether subjects use some characteristics or some tests to determine whether a text extract expresses a causal relation. This experiment allows us to elicit these characteristics and to check whether they match those described in theoretical work. Characteristics thus obtained can be integrated to annotation instructions.

This experiment allows testing of hypothesis Hyp1, which proposes that a set of characteristics are conscientiously used to achieve causal reasoning. We have partially invalidated this hypothesis, as we will see later in this section.

3.2.1 Methodology

To elicit characteristics or test for causation, we have asked subjects to determine whether a causal relation was expressed or not in a number of text samples in French. We have asked the subjects to systematically justify their answer, whether positive or negative. From these justifications, we have inferred characteristics used for causal reasoning.

Extracts for analysis were taken from a number of texts from the BAF (RALI laboratory, 1997) corpus, a heterogeneous corpus of legal, institutional, scientific and literary texts. They were made of one or two sentences. We chose these texts because they contain causal connectives *donc* (thus) or *parce que* (because) or the connective *mais* (but), which is typically non-causal. We thus obtained three types of extracts, each containing a different connective. These extracts are *explicit*: they contain a connective that iden-

tifies the relation that they feature. These extracts are therefore easier to analyse, and can limit useful justifications, since the presence of a connective may systematically be used to justify a judgement. We also created implicit extracts by removing the connective from some of the explicit extracts. We replaced the connectives by a comma when necessary. Each subject had to analyse extracts from each of the six types thus generated: for each of the three connectives, these would be extracts containing the connective and extracts from which it had been removed. Each of the extracts came with a contextual background that allowed understanding of its meaning. The whole batch can be found in annex B.1.

The subjects of the experiments were 9 linguistic students who had not yet studied causation and were fluent French speakers. Each student had to analyse 10 extracts in a different, random order. For each extract, they had to determine whether the extract represented a causal relation and they had to systematically justify their answers, whether these were positive or negative. Each student involved had to analyse 3 implicit sentences generated from extracts featuring a *parce que*, 3 from one featuring a *donc*, 1 from one featuring a *mais*, and one containing each of the explicit connectives. This distribution reflects the degree to which implicit causal sentence analysis matters to us. Analysis took about 45 minutes for each student. An example of a the text printed and given to a student can be found in annex B.2

3.2.2 Results

Most (72.2%) of the sentences were deemed to be causal, which is not surprising since 80.0% of them contained a causal connective or had contained one that had been removed.

We classified justifications obtained in the experiment into 5 classes: reformulation, linguistic test, presence of an explicit marker, presence of a non-causal relation and other. We further detail these categories in the following paragraphs.

Reformulations. We call *reformulations* the type of justification that does not provide any further information on causal reasoning and amounts to paraphrasing the instructions. (3.7) and (3.8) are examples of such justifi-

cations. (3.7) justifies a positive answer for causation and (3.8), a negative one. This type of justification is very common for positive answers (37.5%) and the most common for negative (41.4%).

(3.7) Une explication est donnée.
An explanation is given

(3.8) Je ne vois pas de relation causale.
I see no causal relation.

Linguistic tests. We call *linguistic tests* the justifications that amount to the subject introducing a causal marker into the extract being analysed and observing that the operation produces a valid sentence. These markers may be connectives such as *parce que / because* or *donc / so* (in this case, the subjects often replaced the very connectives that we had removed from the text). Predicates such as *est le fruit de / is the result of*, *entraîne / entails*, *permet / allows* and *est la cause de / is the cause of*. In this case, eventualities that are expressed as clauses are nominalised. Subjects then used the phrase *le fait que / the fact that*. The difference between this type of justification and reformulation is that the text to be analysed is repeated, at least partially, in the justification. For instance (3.10) is used to justify a positive answer to extract (3.9), and (3.12) for (3.11). This type of justification is the most common for positive answers (40.0%) and amounts to only 6.9% of negative ones.

(3.9) 'Les extensions de fichiers sont ajoutées automatiquement par le système, vous ne devriez jamais avoir à en ajouter manuellement."
"File extensions are automatically added by the system, you should never have to manually add them."

(3.10) C'est parce que les extensions sont ajoutées automatiquement que l'on ne doit pas les mettre manuellement.
It is because extensions are automatically added that one should not have to manually add them.

(3.11) 'Le Canada fait partie du Commonwealth britannique et à ce titre, nous acceptons la reine comme chef d'État. Les députés lui prêtent un serment d'allégeance."

"Canada is part of the British Commonwealth and as such, we accept the Queen as head of State. Deputies swear allegiance to her."

(3.12) Le fait que le Canada fasse partie du Commonwealth britannique rend possible le fait que la reine soit chef d'Etat.
The fact that Canada is part of the British Commonwealth makes it possible for the Queen to be head of State.

Presence of explicit markers. Another type of justification consists of reporting the presence of an *explicit marker* in the text to justify whether a causal relation is present. In negative cases, subjects report the presence of a non-causal marker to justify an absence of causal relation. (3.13) is an example attached to a positive answer, and (3.14) attached to a negative answer. This type of justification amounts to 16.7% of positive answers and 3.4% of negative answers. 36.9% of marked sentences that were identified as causal were justified in this manner, and only one negative out of the 9 was justified in this way. A few sentences that we considered to be unmarked were justified in this manner. In these instances, the subject had found another marker that they deemed to be causal, such as *par le fait / by the fact that*, or *induire / induce*.

(3.13) *Donc* apparaît. Donc introduit une conséquence.
Thus appears. Thus introduces a consequence.

(3.14) Le *mais* exprime une nuance, une restriction dans ce cas précis.
The *but* expresses a nuance, a restriction in the persent case.

Presence of a non-causal relation. Sometimes, subjects analysed their text and detected a non-causal relation. They then justified their negative answer by the presence of the other, different relation. An instance is (3.15). This category amounts to 31.0% of negative answers. It is featured with one positive answer, but in this case, the subject had changed their mind after giving their justification. The exact wording of this instance is given in (3.16).

(3.15) C'est une description, la seconde phrase apporte seulement une précision.
It is a description, the second sentence only provides a precision.

(3.16) La seconde proposition détaille la première mais n'introduit pas une cause liée par un connecteur. Mais si on met *parce que* entre les deux ça peut jouer.
The second clause details the first one but does not introduce a cause linked by a connective. But if *because* was inserted in between, it might work out.

Others. Finally, a few justifications did not fit either of the preceding categories. They were mainly interrogation marks drawn instead of a justification and one justification that we did not manage to understand.

3.2.3 Discussion of the results

We will discuss the results of this experiment here. We analyse the influence of connectives on causal judgements and provide arguments that lead us to invalidate hypothesis Hyp1.

Causation and connectives. The sentences that featured or had featured causal connectives overall are analysed as being causal, and those that featured or had featured a *but* as being non-causal, as figure 3.1 shows. All marked extracts were analysed as the marker indicates. The slight loss of positive answers in unmarked sentences can be explained by the suppression of the connective complicating analysis. These same extracts are often analysed as non-causal by a variety of subjects when they originally contained a causal marker. In particular, example (3.17) was analysed as being non-causal by 3 of the 6 subjects who analysed it. One subject detected the causal relation expressed between a lack of preliminary study and the validity of the study, but expressed their disagreement (3.18). It is possible that the other subjects might have made a similar reasoning.

(3.17) "Aucune étude semblable n'a été faite auparavant, ni au Canada ni à l'étranger et ce, malgré la place des TCI dans tous les aspects de la vie quotidienne. Cette recherche sera fort pertinente tant pour les aînés d'aujourd'hui et de demain, que pour les industries et les gouvernements."
"No similar study had ever been made before, either in Canada or abroad, and this in spite of the importance of TCI in everyday life.

This research will be very relevant both for seniors of today and tomorrow, and for industries and governments."

(3.18) Je ne vois pas comment le fait de ne pas avoir fait d'étude de ce genre entraîne une relation causale avec la pertinence de la recherche. Il y a certes une relation logique mais pas causale pour moi.
I can't see how not having performed a study of the kind would entail a causal relation with the relevance of the research. There is indeed a logical relation, but not causal in my opinion.

Extract (3.19), which originally used to contain a *but* between its two clauses, was analysed as causal by one of the 5 subjects who studied it with justification (3.20). This extract is difficult to analyse. The subject detected a relation, plausible but complex, with a cause that did not appear in the text (loss of terminology) but that is suggested by *without missing any of the terminological incoherences.*

(3.19) "Dans quelle mesure nous parviendrons à [assouplir certaines conditions] sans que nous échappe aucune des incohérences terminologiques réelles présentement détectées par le système, cela reste à voir. Il serait beaucoup plus difficile de concevoir et de rendre opérationnelles des stratégies permettant au système de distinguer entre les cas acceptables et inacceptables de termes cibles qui ont été omis, ou encore remplacés par un autre terme ou une périphrase."
"In what measure will we manage to [relax some conditions] without missing any of the actual terminological incoherences presently detected by the system, that remains to be seen. It would be much more difficult to design and make operational strategies that would allow the system to distinguish between acceptable and unacceptable cases of target terms being omitted, or replaced by another term or by a reformulation."

(3.20) La perte de terminologie rendrait difficile l'utilisation de la stratégie de distinction.
The loss of terminology would make it difficult to use distinction strategy.

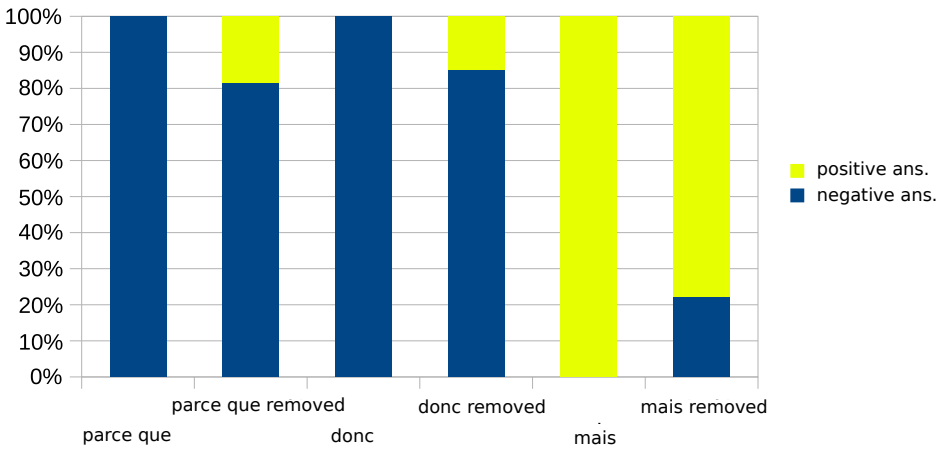


Figure 3.1: Types of sentences identified as causal. This graph represents the percentage of each type of sentence identified as causal or non-causal.

These results are not surprising. We noticed a global adequacy between answers and original connectives, with a slight noise caused by suppression of the connectives that simplified analysis.

Homogeneity of justifications. We were surprised to see no reasoning pertaining to the eventualities themselves, but only justifications of a linguistic sort. In particular, we would have expected to find counterfactuality as a justification: if the cause had not occurred, then the consequence would not have occurred either. This classical property of causation as been identified in (Hume, 1739), for example.

We attribute this lack of justification about the way the eventualities are linked in the real world, rather than as in the text, to the context of the experiment. Indeed, this was a linguistics experiment, on texts and with linguistics students. It would be interesting to know whether the phenomenon would be reproduced in an experiment advertised as psychological, with video and engineers. We remain convinced that such an experiment would produce many more justifications pertaining to world knowledge rather than to linguistic knowledge.

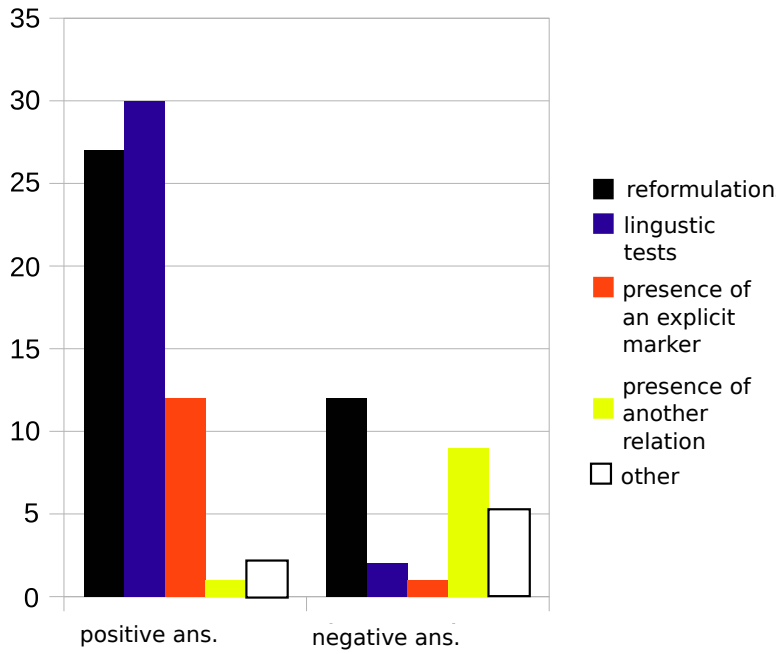


Figure 3.2: Justification types for causal intuitions. This graph shows quantities of justifications of each type for phrases considered to be causal (left) and non-causal (right)

Invalidation of hypothesis Hyp.1. Figure 3.2 shows the number of positive and negative answers for causation that contain a justification of either type. Answers that contain several justifications contribute to each of the corresponding types. The most frequent justification type for positive answers is linguistic test, closely followed by reformulation. For negative answers, reformulation is the most frequent. We think that this large representation of reformulation indicates that the subjects have had little access to their own causal reasoning. They have an intuition as to whether something is causal, but they do not know how they know it.

This conclusion is also motivated by the difficulties we encountered in obtaining justifications during preliminary experiments, in particular for negative cases. We have observed many failures to provide justifications in spite of instructions requesting systematic justification of answers. We obtained these only from instructions that introduce the question:

> Une relation causale est-elle exprimée entre certains éléments de ce passage ? Barrez s'il vous plaît la mention inutile et justifiez votre réponse.
>
> Je pense qu'une/aucune relation causale est exprimée dans ce passage parce que ...
>
> Is a causal relation expressed between some elements of this extract? Please cross out superfluous text and justify your answer.
>
> I think that a causal relation is/is not expressed in this extract, because...

This difficulty is a clue that this task is not spontaneous for subjects.

Homogeneity of justifications, large representation of reformulations and reluctance in justifying answers are strong clues that the first hypothesis is false. This hypothesis states that intuitive characteristics exist for causation that are accessible through reasoning. While there certainly exist characteristics of causation used for causal reasoning, our experiments indicate that they are not directly accessing through conscious reasoning.

Use for an annotation manual. All of these justifications will satisfy causation experts. They offer a good description of the characteristics of causation. We have integrated them to the annotation manual as follows: reformulation indicates a causal intuition, which is the most important test. As we will see below, the other tests are used mostly to clarify intuitions and to eliminate non-causal cases. We have also used linguistic tests in the manual. We did not pay attention to explicit markers as our aim was to disambiguate implicit cases. We also did not ask annotators to identify non-causal relations either. We think that a precise identification is pointless when identification as a non-causal relation is sufficient. Furthermore, some non-causal relations, such as temporal relations, are not incompatible with a causal relation, as in (3.21).

(3.21) Il gagne au loto et se croit tout permis.
He wins at the lottery and thinks he can do anthing.[2]

[2] Title of the *Matin Bleu*, 18 August 2009

3.3 Correlation between judgement on causation and on its characteristics

We describe an experiment here that allows us to test hypothesis Hyp2, which states that human judgements on causation are correlated with judgements on characteristics of causation. This experiment allows us to determine whether some factors do co-occur with causation, and to which extent they are linked to it. These characteristics may thereafter be used to create an annotation manual, as we will see in section 3.4.

3.3.1 Methodology

We used an ad hoc programme: CARMA (CAusal Relations Multi-user Annotations). CARMA displays sentences to annotators and allows them to identify characteristics of causation and to determine whether the sentence is causal. The programme displays sentences where the potential cause and the potential consequence are underlined. It records the answers of the annotators. Thus, an annotator has only to determine whether the sentence is causal or not, but does not have to delimit the segments that represent cause or consequence.

Subjects for these experiments were four linguistics students who had not yet studied causation and who spoke fluent French. We asked them to decide whether 24 sentences amounted to causal relations. The sentences were artificial sentences that were ambiguous as to causation (given in annexe B.3). Figure 3.3 represents an example of an instruction with a sentence. The different characteristics were detailed to the subjects before the experiment.

The characteristics to be analysed were chosen from our previous experiment that aimed at eliciting intuitive characteristics of causation, and from previous theoretical work. We shall detail them in the following paragraphs.

Detailed temporal order. Temporal asymmetry is a usual characteristic of causation. It was first identified by Hume (1739). In many causation cases, such as in the example *John fell, Max pushed him*, the cause happens before consequence. In cases of direct causation, such as (3.22), the cause immediately precedes the consequence. In some cases, cause and consequence are more or less simultaneous, such as in (3.23). In this experiment, we require

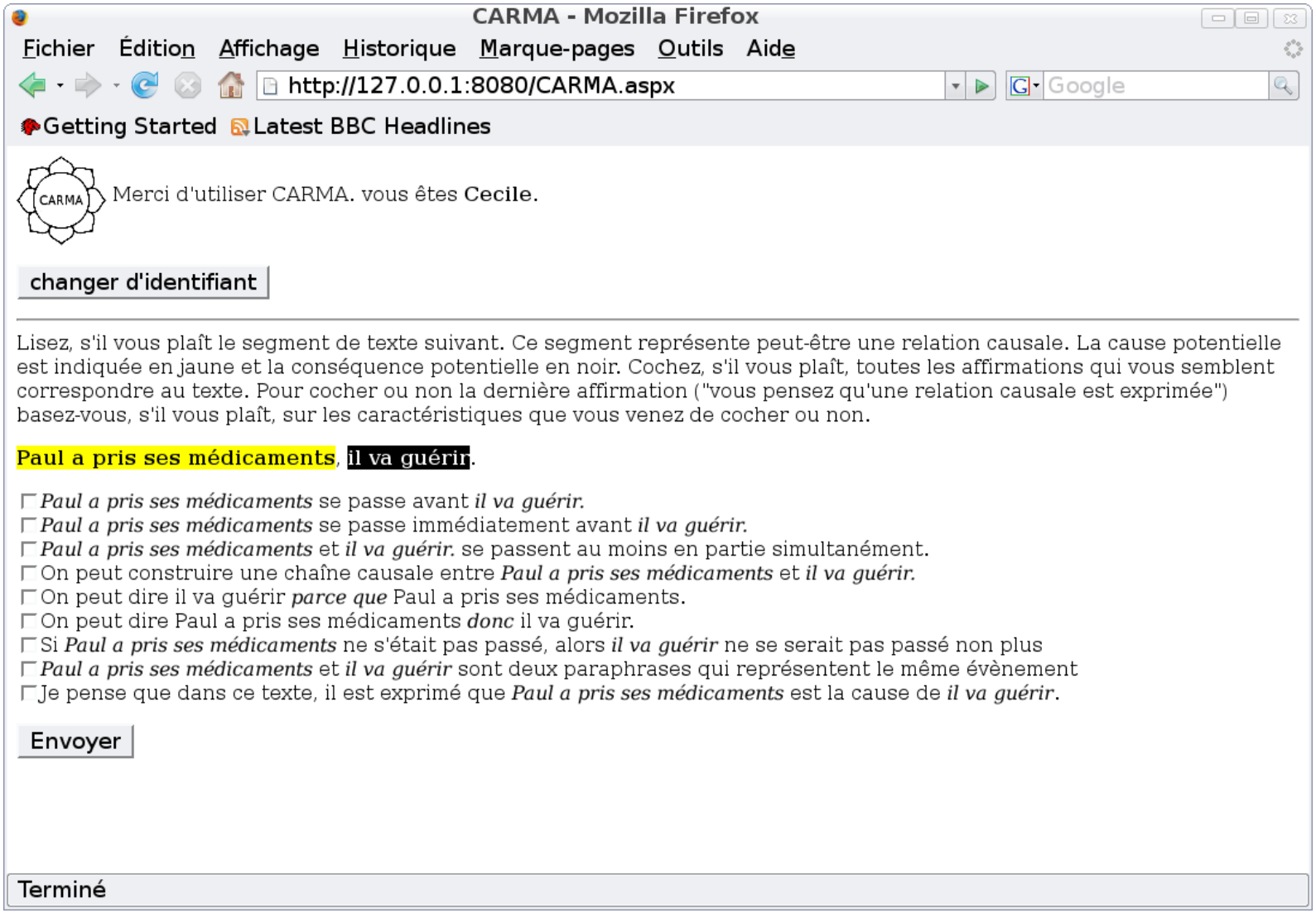


Figure 3.3: CARMA displaying instructions along with a sentence.

a detailed analysis of temporal order using identification of the following characteristics: cause preceding consequence, cause immediately preceding consequence, or partial simultaneity.

(3.22) The glass reached the floor and broke.

(3.23) He got tired driving.

Causal chains. A causal chain is a chain composed of causes and direct consequences that can be associated to a causal relation (as studied, for instance, by Moeschler, 2003a). For example, for the causal relation (3.24), causal chain (3.25) can be built. We asked annotators to determine whether a causal chain could be formed between eventualities that they had to analyse.

(3.24) John fell, Max pushed him.

(3.25) Max pushes John → John is unsteady → John falls.

Linguistic tests. Linguistic tests were very frequent in the previous experiment. Here, we asked whether is was possible to insert *parce que / because* and *donc / so* in sentences, and to determine whether the result is correct, syntax being adapted as needed.

Counterfactuality. The counterfactual property is an important feature of causation (see, for instance, Reboul, 2005). Counterfactuality refers to the fact that if the cause had not occurred, then the consequence would not have occurred either. This is one of the properties that distinguish causation from the logical *if... then*. This property allows easy elimination of some non-causal examples such as *My bus is about to leave, I just finished my breakfast*. Indeed, one can determine that these eventualities are not related by causation, since the bus would depart soon regardless of whether I had just finished my breakfast or not.

Paraphrases. We request annotators to identify potentially causal couples of eventualities that are paraphrases of each other. Indeed, an eventuality normally cannot cause itself to happen, and paraphrases of one same event, such as (3.26), should not normally be causal.

(3.26) This is a triangle, it has three edges.

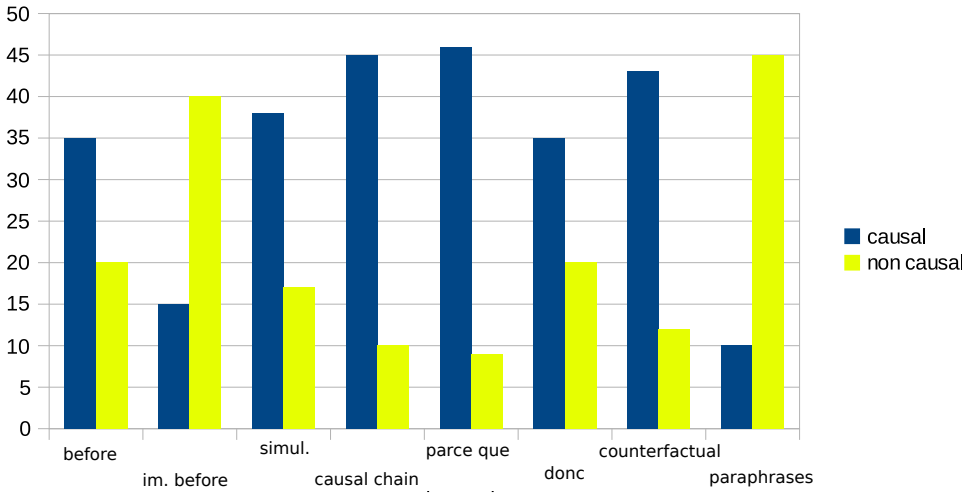


Figure 3.4: Occurrences of the characteristics of each type identified for sentences judged to be causal and non-causal.

3.3.2 Result analysis

We present results on correlation between characteristics and causation and on some first inter-annotator agreement tests on this experiment.

Verification of hypothesis Hyp.2 Figure (3.4) shows the quantity of positive answers for each characteristic for sentences analysed as causal and non-causal. We note that some characteristics are indeed often associated with causation; in particular, causal chains, linguistic tests with *because* and the counterfactual property happen much more often when a causal relation is occurring than when it is not. We now tested the actual statistical association of characteristics with causation using the exact Fisher test. To perform a test of statistical signification of data, we have to formulate an hypothesis called the null hypothesis, or H_0. This hypothesis expresses what we aim to disprove. In our case, the null hypothesis for each of the characteristics is that the characteristic in question is independent from causation.

The exact Fisher test is a correlation test for binomial variables especially well-suited to our case, as it allows us to take into consideration cells that contain few occurrences (we have only two sentences that have been analysed

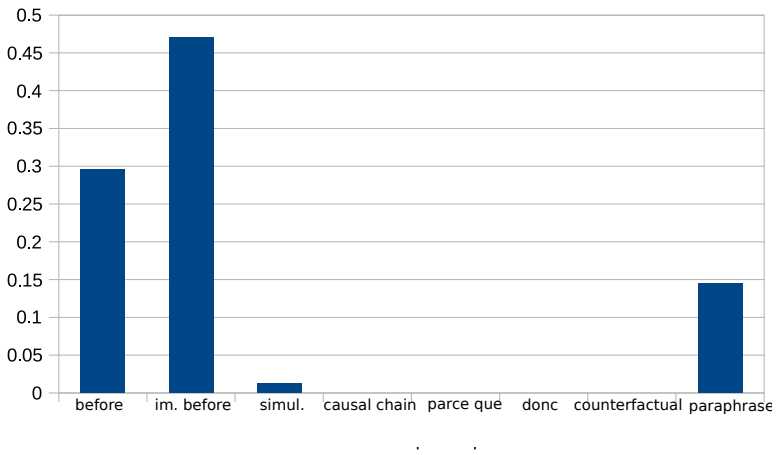


Figure 3.5: P-value of the exact Fisher test for the association of each of the characteristics with causation. The lesser the value, the more associated the variables.

as causal for which the annotator was unable to construct a causal chain, for instance). This test allows us to compute p-values. A p-value is the likelihood that a null hypothesis would generate data at least as extreme as observed. The smaller the p-value, the more the null hypothesis can be rejected. For a p-value of 0.05, for instance, we know that if the null hypothesis was true, there would be a 5% likelihood of generating the data as observed, or more extreme data. We would therefore have a 95% chance to be correct in rejecting the null hypothesis, given the observed data.

Figure 3.5 shows the p-value of the exact Fisher test. If the null hypothesis is to be rejected for values lesser than 0.05, it can be concluded that the following characteristics are indeed indicators of causation: simultaneity or partial simultaneity, possibility of forming causal chains, linguistic tests with *parce que / because* and *donc / thus*, and counterfactuality. Our hypothesis Hyp2, according to which a certain number of characteristics are correlated with causation in human judgements, is therefore verified.

Inter-annotator agreement We compute the agreement using Cohen's κ score (Cohen, 1960). As we have seen in section 2.6.3, this score is based on agreement percentages between annotators from which agreement caused by chance has been removed. If two annotators randomly consider a property

to be positive in 90% of cases, they will have a greater agreement percentage than if they annotated it as positive in 50% of the cases. The κ score allows this phenomenon to be taken into consideration and it will be equivalent in both cases. This score varies between -1 and 1. A score of 0 indicates a complete independence between annotations, and 1, a perfect match.

In this experiment, the mean inter-annotator agreement for each pair of annotators was a mediocre $\kappa = 0.3$ for causation. We also computed an agreement between our annotations and those of the majority of the students, in order to determine whether we have the same mean understanding of causation. The κ score between our annotations and those of a majority of the students was $\kappa = 0.43$. We computed this last value by discarding examples made ambiguous by a lack of context and that can therefore be interpreted as causal or as non-causal only if more information was provided, such as:

(3.27) After a long trip to Canada, he got married.

We also discarded sentences that did not yield a majority amongst the answers of the annotators (sentences that two annotators had considered to be causal and two others had considered to be non-causal). This score of 0.43 is mediocre but allows the determination of where divergences lie between between our vision of causation and that of the students, and generation of an improved annotation manual. Sources of disagreement between our visions and a majority of annotations notably contained cases of sentences that had been subject to linguistic tests but that we did not consider to be causal. These are sentences that can be linked with *parce que* or *donc* but that fall into categories of epistemic usages or speech acts. We will detail these issues in more detail in section 3.4.1.

We also had the same experiment performed by two causation experts. We obtained a mediocre inter-annotator agreement of $\kappa = 0.32$, but this experiment was the basis of discussions on the points of divergence that allowed us to clarify some points in the annotation manual. In the following section, we describe this process and the resulting annotation manual.

3.4 Annotation instructions

Annotation instructions are a set of characteristics that an annotator has to identify. Some characteristics are necessary for causation, some allow the annotator to eliminate non-causal cases and some provide help in disambiguating difficult cases. Since we do not have necessary and sufficient conditions to identify causal expressions, annotators have to rely on their intuition to decide, ultimately, whether the case is causal or not. Instructions allow them to decide on some ambiguous cases and should support and guide the intuition of the annotator. We will describe these annotations in more detail in the following paragraphs.

We relied on two criteria to create an annotation manual from the set of initial characteristics described in section 3.3: divergence between causation experts and divergence between our predictions and the annotations of a majority of annotators.

We used results from the experiment described in (3.3) to refine and clarify our instructions. We tested our instructions in a new experiment using the CARMA software with a modified interface. In particular, we added context to the sentences to be annotated and we requested systematic identification of a smaller set of important characteristics. The new CARMA allows annotators to choose a *does not know* option for each characteristic. Segments to be analysed are signalled, but it is up to the annotator to decide whether the text features clauses in the cause-consequence order, or in the consequence-cause order. We think that this choice allows us to bring attention to symmetric statements, and thus non-causal, as we shall see in section 3.4.4. Figure 3.6 shows the new interface.

We shall now describe the new instructions that allow us to obtain a good κ between a majority of the annotators and our predictions. In section 3.4.7, we shall discuss the results yielded by testing these instructions.

3.4.1 Linguistic tests

In our previous experiment, results of linguistic tests with *parce que* and *donc* were correlated with intuition of the annotators on causation of the statements. These tests were also very much present in justifications of the experiment that allowed us to find intuitive characteristics of causation.

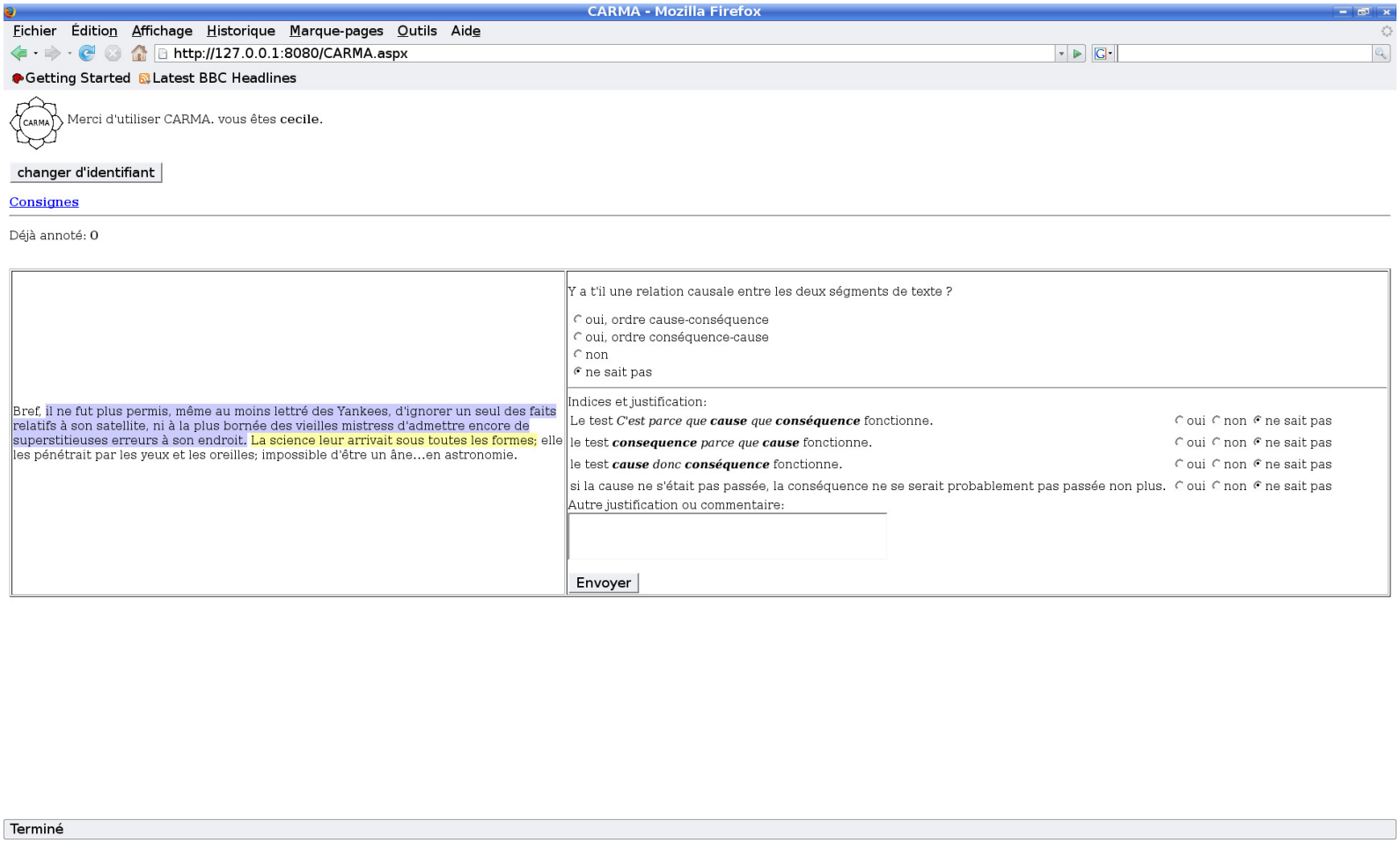


Figure 3.6: CARMA features a statement along with its context and records answers.

The issue of *parce que* and *donc* is that they are causally ambiguous. Indeed, *because* and *thus* can be used as speech acts, such as in *hurry up because we are going to be late*, where the second part of the sentence is not the cause of the first one, but of its enunciation. We did not wish annotators to solve these types of ellipses, but wanted sentences like this to be annotated as non-causal. Similarly, in *George has left because his jacket is not on the chair*, the second phrase is not the cause of the first one, but of the belief in it. We did not wish for *George's jacket is not on the chair* to be annotated as the cause of *George has left*. Nevertheless, in this case, the cause of the belief is a consequence of what is being believed. Indeed, *George has left* is the cause of *his jacket is not on the chair*. We therefore wished annotators to annotate this sentence as causal, but on the inverse order of that (consequence-cause), which the *because* connective usually requires. Moeschler (2011) has made a thorough analysis of the different usages of the French *parce que* as well as the relations between its causal and argumentative (speech act and epistemic) usages.

In the previous experiment, one cause of disagreement between our predictions and the majority of the annotations was sentences that pass linguistic tests but amount to epistemic uses or speech acts. In instructions, we detailed epistemic use and speech act of *parce que* and *donc* in the annotation manual. We asked annotators to systematically indicate whether the extract being analysed was compatible with linguistic tests. In particular, we asked them to try to insert *parce que*, *donc*, and, while accomodating possible syntactic issues, to decide whether the resulting sentence was semantically correct. We also added a requirement to systematically annotate the test with *c'est parce que / that is because*. Indeed, *c'est parce que* does not, up to our knowledge, translate epistemic use or speech acts. While 3.28 is a perfectly acceptable French utterance, 3.29 is at least very strange and does ask for a lot of accomodation, as its only interpretation would be that *George's jacket is not on the chair* is the reason why *George has left* (maybe, upon finding that somebody has stolen his jacket, which was supposed to be on the chair, George leaves to show his anger).

(3.28) George est sorti parce que sa veste n'est pas sur la chaise./*George has left because his jacket is not on the chair.*

(3.29) C'est parce que sa veste n'est pas sur la chaise que George est sorti./*It is because his jacket is not on the chair that George has left.*

We required systematic annotation of linguistics tests on this criterion because we think that they will allow us to very quickly perform discriminative and intuitive testing of causation.

3.4.2 Counterfactuality

Even though this property was discriminant in our previous experiment, it has not been associated with all positive answers. Indeed, for instance, in the sentence *John broke his leg while skiing*, some annotators had considered that the counterfactual property did not apply, even though they considered the sentence to be causal. They argued that if he had not been skiing, John could nevertheless have broken his leg, for instance, by slipping on a banana peel. Therefore, we mentioned and commented on this example in the new annotation manual. We also added *likely* to the formulation of the counterfactual property, in the following experiment : *if the cause had not happened, then the consequence would likely not have happened either.*

We requested that annotators systematically identify the counterfactual property in extracts that they were annotating, as we thought it would allow them to quickly discriminate between some non-causal cases and that it could be quickly identified.

3.4.3 Temporal asymetry

In the previous experiment, we requested from annotators that they precisely identify the temporal order of the eventualities that could potentially constitute a causal relation. By doing so, we wished to clarify their intuition through rigorous analysis of the temporal order. However, the analysis was time-consuming and proved to be of little use, as only a partial simultaneity could be significantly correlated with causation. This is the reason why, in the previous experiment, we only requested them to annotate as non-causal the cases in which the cause would happen after the consequence. Indeed, it was less costly to discriminate cases where the potential cause would happen after the potential consequence than to provide a refined analysis of the other cases.

3.4.4 Asymetry

In the previous experiment, we asked annotators to identify whether propositions were paraphrases one from the other. This property was very scarcely annotated (14% of all examples) and very little discriminative. We generalised it in the manual as the *asymmetry property*.

Causal relations are asymmetric (Hume, 1739), not merely temporally but also causally. If an eventuality is the cause of the other, the second one can at best, in anecdotal cases, be the cause of the first one. This property is present even in those cases where cause and consequence are simultaneous. For instance, in *John is getting tired while driving*, the fact that John would drive is the cause of his fatigue, but his fatigue does not cause him to drive. Cases of circular causation could be imagined, but we think that these cases are rare and that both causation directions would then be mentioned separately in the text, because of the unusual nature of the phenomenon. Should such a case arise, then both occurrences of causal relation should be separately annotated and each only in the direction that is expressed.

In our second experiment, for each potentially causal pair of event, we asked annotators to indicate whether the text features them in the cause-consequence order or in the consequence-cause order. We asked them to annotate examples for which it is difficult to determine the order or for which both orders could be possible as non-causal. In particular, we wished to avoid the condition where annotators would consider a sentence such as *this is a triangle, it has three edges* as causal, as it often was annotated as causal, which does not match with our conception of causation.

3.4.5 Causal chains

Although this property was very discriminative in the previous experiment, we do no know whether it allows the annotators to clarify their intuition or whether they needed to already have an intuition as to whether an extract was causal before they could construct a causal chain. If an intuition as to whether the extract is causal is necessary to construct the causal chain and if this intuition is not questioned by this exercise, then identification of a causal chain is redundant. It also constitutes a waste of time for annotators, all the

more since the process is relatively costly, as a number of possible causes and consequences must be considered between the clauses being analysed.

Therefore, we did not request that the annotators systematically identify causal chains in this manner in this annotation experiment. We did, nevertheless, suggest it as away to clarify intuition on ambiguous cases, in particular if eventualities being analysed were temporally remote.

3.4.6 Further annotation rules

We clarified two further cases in the annotation manual. First, sentences such as *he broke a leg while skiing* generated disagreements between causation experts. This issue stems from the fact that *skiing* is actually not the cause of *breaking one's leg*. The cause of the leg fracture is an unspecified eventuality within the context of the skiing, such as *falling* or *colliding*. We therefore had to take an arbitrary choice: decide whether the context of the eventuality can be considered to be causal or not. We decided to ask the annotators to annotate such contexts as causal, because we wanted annotations rules to define a precise, though wide, acceptance of causation.

We also discovered, through discussion about expert disagreement, that causal judgement becomes more difficult if the potentially causal eventuality is a negative event, such as in *as no similar study was performed before, this report will be very relevant*, or when its aspectual class is a state, such as in *Marjorie is of age, she can vote*. These cases are often of an inferential nature. Therefore, we advised the annotators to try to clarify their intuition in such cases by constructing causal chains or by trying to refer to a general law, such as *when of age, one can vote*. We also requested, in these cases, that they check asymmetry, for instance with the *c'est parce que* test. Indeed, *it is because Marjorie is of age that she can vote* sounds natural, while *it is because she can vote that Marjorie is of age* does not sound quite so natural.

.

3.4.7 Testing the manual

We shall now present the results of the experiment that allows us to test our annotation manual. The 15 extracts to analyse were taken from Jules Verne's novel *From the Earth to the Moon* (given in annex B.4). Each extract

consisted of one paragraph and contained two emphasised statements that were to be analysed as to determine whether they had a causal relation. We chose these extracts by eliminating the pairs of eventualities that were the most clearly non-causal, in order to save time for the annotators and to obtain more significant results. Each of these 15 passages was analysed by 4 linguistics students who had not studied causation and who spoke fluent French.

For this experiment, we obtained a mean κ score of 0.38 between annotators for causation. The improvement in κ score is therefore very small with respect to the previous experiment, for which it was 0.30. This slight improvement could be explained by the fact that the new experiment used extracts from an existing text, rather than statements deliberately chosen to be causally ambiguous.

We did, however, obtain a very significant improvement of the κ score between our own annotations and those of a majority of the annotators. Indeed, the new score amounts to 0.84, compared to 0.43 previously. As before, the score ignores extracts for which there was no majority in the answers of the annotators (which was two of the fifteen statements). Amongst the thirteen examples with a clear majority, the majority differed from our annotations only once, when three of the annotators considered as causal an extract that we deemed to be non-causal (the fourth considered it to be non-causal).

We think that this very high agreement between our annotations and the annotation resulting from a voting scheme on all annotator's decisions indicates that we managed to transmit our vision of causation in a clear and coherent manner. We succeeded to translate into our manual our understanding of causation with enough clarity for annotators to be able to obtain results close to our own. We could therefore develop a clear and coherent methodology to identify a precise type of relation in a text. Expert agreement and the fact that the manual is based on theoretical work allow us to affirm that these relations are indeed causal relations. We therefore validate our hypothesis Hyp3, according to which our manual allows for a coherent annotating of causation.

3.5 Taking the subjectivity of a task into account for its evaluation

As we have seen in this chapter, recognising implicit causation is a difficult task even for human annotators. Even with carefully constructed annotation guidelines, annotators diverge on many instances, as some level of subjectivity is involved in the task. These disagreements are often not due to human errors but to an intrinsic degree of subjectivity in the task. Sometimes, several different answers are all correct. As we have seen in section 2.6.1, this amount of subjectivity can be evaluated with the inter-annotator agreement. A task leading to a low inter-annotator agreement is more subjective than is one where each annotator agrees for each instance.

As we have seen in section 2.6, automatic systems are evaluated using a gold standard, which is one set of annotated examples considered to be the absolute truth. However, this gold standard is artificial in that it was either created by some sort of voting scheme that put together diverging annotations form several persons or it was annotated by only one person, whose opinion differs from that of others. For an automatic system to reproduce such an artificial standard is both extremely difficult and not very useful. One cannot expect a good programme to reproduce such a subjective standard. Ideally, an automatic system would be as good as any trained annotator, not the exact replica of an arbitrary standard. This means that a point of agreement exists between an automatic system and a gold standard above which a rise in the agreement is meaningless for the task. Above this point, the system is over-fitting the actual evaluation data and not getting any better at the abstract task it was created to do.

The actual best automatic system does not mimic a specific gold standard of a task. It mimics a trained human behaviour. The evaluation should then not measure the exact distance separating the system from the gold standard but should give an idea of how far the system is from the results generated by actual human annotators. We therefore need a statistical test indicating how far away the system is from actual humans; that is, a test giving the likelihood that the results from the automatic system were generated in the same way as the results of the human annotations.

For this type of evaluation to make sense, we need the system's eval-

uation measure to have the same units as the inter-annotator agreement. Both measure the same thing: the distance between two sets of annotations (human versus human in the case of the inter-annotator agreement and machine versus gold standard in the case of an automatic system's evaluation). Some measures used for both tasks actually only differ in their name. Raw agreement and accuracy are the exact same measure: the number of common annotations divided by the total number of annotations.

To compute this measure, we need an idea of the typical inter-annotator agreement, as a global measure or as a pairwise mean, and we also need a measure of the natural dispersions of these agreements. Let us take for example a mean pairwise raw agreement score of 80%. An automatic system scoring a 70% accuracy should be evaluated as much closer to humans, if the actual pairs are, say, 65%, 70%, 90% and 95%, rather than 79%, 79%, 80% and 82%, since in the first case, the machine is actually within the normal human range, while in the second case, it is outside of it. This measure is traditionally given by the *standard deviation* of the data.

Let us take a look at a possible modelling of the inter-annotator agreement. Assume that we have access to pairwise agreement measures. We then have several measures that should be the basis for us to understand what is the normal way the measures are distributed. We can then compare a new measure–the system-human agreement– to the set of human-human measures that we have, and see whether it looks similar. In order to have a mathematical measure of how similar the new measure is to the old ones, we need to deduce, from the human-human measures, what is actually normal. We know that each human human measure is normal, but, intuitively, we also know that similar measures should also be considered normal. What is exactly normal, and how normal it is, can be mathematically computed, given the *distribution* of our measures. Let us take a simpler example: somebody tries to hit a target with a dart. The distance the dart falls from the target will vary, with most being close to the centre and few being very far away. In this case, we can model the distance distribution with a gaussian –also called normal– function, which is a very widely used distribution that depends on the mean of the measures and on the deviation of measures from this mean. Given a few measures, a mean and a standard deviation can be computed. Given these two numbers, the whole distribution is known, and it is possible to compute accurately how normal a new dart throw is, given

previously measured distances.

In the case of pairwise inter annotator agreement, these will be few and intuitively, a normal distribution is reasonable, with most measures falling near the mean and measures away from the mean getting more and more improbable, with a general curve depending on the standard deviation. In this precise case, since inter-annotator agreement measures typically fall in a 0-1 interval (this is the case for accuracy, while κ lies between -1 and 1), the normal distribution is not an exactly accurate model of the data because it supposes that measures can become arbitrary far from the mean, which, because our measure is bounded, is not our case. A general model would be the beta distribution, which is a very general distribution for bounded variable, but the beta distribution is very unconstrained and it is difficult to fit an intuitively satisfying beta distribution with the typical extremely small sample of pairwise agreements. The theoretical best function in our case is probably the double truncated gaussian function, a version of the gaussian for bounded variables, but it gets quite difficult to compute measures for it. It is also rarely used in practice, leading to a lack of implementation of its related measures in software typically used for simple statistical calculations. The normal distribution, although not the theoretical best for our case, has the advantage of being widely used and of having a wide variety of available tools associated with it [3]. Furthermore, we claim that it suits our purpose as using it would preserve the ranking of evaluated system while still giving an idea of the distance separating two systems.

For the rest of this section, we will rely on one examples: evaluating the system of (Pitler et al., 2009). As all necessary data are not always available from the relevant papers, we will try to give a fair guess of any missing data. Because of this necessary guesswork, the results are not meant to be accurate; these are only examples of the way such an evaluation scheme could be used. Pitler and her colleague's paper aims at recognising implicit discourse relations. Two large corpora exist with annotations for this task:: the Penn discourse treebank PDTB and the rhetorical structure theory corpus RSTC (Carlson et al., 2002). Pitler and her colleagues use the PDTB for evaluation, but as inter-annotator agreements were measured for two annotators only, we do not have any idea of the dispersion of such agreements for the PDTB.

[3] even online. See `http://www.danielsoper.com/statcalc/calc53.aspx`, for example.

Releasers of the RSTC, however, give 6 pairwise Siegel and Castellan's (Siegel and Castellan Jr, 1988) K for the global task of relation classification, that is for implicit as well as explicit relations classification. We will use these numbers as an approximation, even though the system was evaluated on another corpus and on implicit relations only. From the PDTB inter-annotator agreement measures, we can compute a mean $\mu = 0.72$ and a standard deviation of $\sigma = 0.16$.

We now need to express the system performance on the same metric as the inter-annotator agreement on the corpus; that is, Siegel and Castellan's K. In Pitler and colleagues' paper, the evaluation is given as several measures, one of them being accuracy, which is the same measure as raw agreement. Siegel and Castellan's measure is raw agreement normalised, taking chance agreement into account. In this case, we can estimate chance agreement in the Siegel and Castellan sense, which is derived from the total number of occurrences annotated within each class by an annotator. We have an idea of this proportion, because Pitler and her colleagues give the proportion of each class in their test set. If we assume (which is far from precise) that the automatic system got the proportions right, so that it does not modify the overall proportions, taking both annotators into account, we can rely on these proportions for computing the K score corresponding to the accuracy score. As the classes are fairly skewed, in this case, chance agreement is about 0.47 (or 47%). We can then compute the corresponding K from the 63% accuracy : $K = \frac{0.63-0.47}{1-0.47} = 0.30$. This measure corresponds to a far from perfect, yet still significant, agreement between the automatic system and the human annotators.

Let us now compute the likelihood that this measure, or a worse one, was generated in the same way as the human-human inter-annotator agreements. The measure should be low, as it is far from the case, as humans consistently agree with each other much more than the system agrees with them. To compute this likelihood, we use the cumulative distribution function of the gaussian associated with the human-human inter-annotator agreements. The score is 0.004 or 0.4%. This means that the likelihood that the score was achieved by an actual human is a very unlikely 0.4%. The score is far from the human-human scores, but still above 0.

This method will give low results for most automatic systems, but this is

not surprising as it computes the likelihood that the system is as good as a human annotator. It has the advantage of taking the subjectivity of the task into account when evaluating the system, giving better scores for tasks that are more subjective and leading to the same accuracy as their more objective counterparts. The method is independent of the measure used to compute the inter-annotator agreement, only requiring the system to be evaluated with a similar measure to that used by the corpus releasers to assert inter-annotator agreement. As it relies on a measure of dispersion, it cannot be used for corpora that are associated with only one inter-annotator measure.

As work regularly appears arguing for the use of different evaluation measures for inter-annotator agreements as well as for evaluation metrics for automatic systems, we claim that the best way for a paper to report these results would be to give raw tables (or only a reference to a web site where these numbers can be found) as well as the results of an evaluation measure. The use of tables indicating the exact number of commonly and differently tagged classes for each classes and each annotator would allow for a more transparent evaluation of corpora and automatic systems, as they would allow the reader to compute different evaluation scores.

3.6 Conclusion and perspectives

In this work, we described a first experiment aimed at discovering intuitive characteristics of causation. This experiment allowed us to partially invalidate our hypothesis Hyp1 that there are characteristics of causation consciously accessible for causal reasoning. We are led to invalidate this hypothesis by the difficulty of obtaining justifications of causal judgements, by the homogeneity of justifications and by the high amount of reformulation in justifications.

We validated our hypothesis Hyp2 that some characteristics are associated to causation in human judgement with an annotation experiment on these characteristics and on causation of ambiguous extracts.

We took inspiration from these characteristics as well as from theoretical work and expert opinions to write annotation instructions for causation, which we then tested. We presented, in detail, the instructions that allow a majority of annotators to mimic our educated judgement. We then argued

that these instructions are coherent since they permit a high agreement between a majority of annotators and our predictions, and that they do indeed describe causation rather than some other relation, since they have been validated by experts. We thus validated our hypothesis Hyp3 that our manual allows for a coherent annotation of causation.

We claim that our methodology, somewhat inspired from that of Hovy and colleagues (2006), can be applied to numerous tasks that do not immediately permit a high inter-annotator agreement. We also state that first seeking intuitive characteristics in such tasks can be useful for understanding related cognitive processing, and most of all for designing annotation instructions that take this intuitive process into account. These instructions will have to integrate intuitive reasoning whenever it satisfies experts, and, if it does not provide a good understanding of the task, they will have to explicitly caution annotators who might be tempted to rely on it. Finally, we hold a firm belief that expert discussion on conflicting annotations can provide a better understanding and, most of all, a better formalisation of the task being defined.

Nevertheless, our annotation manual does not allow us to obtain a sufficiently high agreement between annotators, and it would need further refining to allow us to annotate an actually useful corpus. As we have seen in section 2.5.4, an important difference between our work and the more successful work of (Hovy et al., 2006) is that we do not change the task defined in the annotation manual, we only attempt to clarify it. This might explain why we do not obtain a good inter-annotator agreement in spite of our efforts. We are convinced that by performing more iterations and clarifying our manual further, our methodology might allow us to obtain a somewhat better agreement. Nevertheless, we think that the number of necessary iterations is high because of the intrinsic subjectivity of the task and because we do not want to adapt it, as we aim at satisfying linguistics experts of causation. This is the reason why we satisfy ourselves with a manual that allows transmission of our vision of causation to annotation resulting from the annotators votes.

Several point still remain to be clarified. In particular, we would like to explore the notion of subjectivity. It seems clear that causation is a subjective notion and that it is the point of view of the text itself, rather

than that of the annotator, that is to be analysed. We will have to also try to precisely annotate which parts of the text represent cause and which part represents consequence. The limits of these segments will probably not be immediately consensual and will require clarification in the annotation manual.

Designing this annotation manual has allowed us to clarify the task that we study in the present thesis and to highlight the high level of subjectivity that is associated with it, as even a very carefully designed annotation manual does not lead to high pairwise inter-annotator agreement.

We have shown that the pragmatic task of identifying occurrences of implicit causation is very subjective. We believe this result to stem from the fact that our task entails two other subjective tasks. First, an annotator has to decide whether the two eventualities are likely to stand in a causal relation in the world. We believe that finding causation in the world is very subjective. An eventuality never happens for a single reason. A set of converging eventualities or lack of eventualities all form causes. Each of them is itself part of a causal chain encompassing numerous causes. Finding a single cause in all of these is highly subjective. Deciding how long a causal chain can be formed between two eventualities before one cannot consider the first as the cause of the second is highly subjective.

Second, on top of the intrinsic subjectivity of deciding whether a real-world eventuality is the cause of another one, our task adds a layer of subjectivity due to the degree of focus cast on causation in the text. An annotator has indeed to decide whether the text presents the eventualities as causal. Two eventualities, indeed, never occur together in the text linked only by a causal relation. A causal relation is always at least associated with a temporal relation. Deciding whether the author wanted to give focus to the causal relation or to the temporal relation is also a subjective task. Therefore, it is perhaps less surprising to discover a low inter-annotator agreement for this task. In this chapter, we could precisely evaluate it and we measured in in several situations.

Finally, we propose an evaluation metric that allows us to take into account the subjectivity of a task, as measured by the gold standard inter-annotator agreement, when evaluating a system that aims at automatically reproducing the task that the annotations exemplify. Our measure relies on

a mathematically accurate likelihood that the system is as good as a human annotator. Given the state of the field, where systems are very far from humans, this measure tends to be low. However, it has the advantage of highly evaluating systems that are close to humans rather than highly evaluating systems that are close to an arbitrary standard that takes no account of variations induced by normal human subjectivity.

Chapter 4

Basic world knowledge: verb pairs

4.1 Introduction

As described in more detail in chapter 2, previous work in theoretical linguistics has postulated that the main feature used by humans to understand implicit causal relations is *world knowledge*. In their seminal paper about the segmented discourse relation theory (SDRT), Lascarides and Asher (1993) describe this world knowledge as a set of rules connecting eventualities that are normally causally linked. Although these eventualities can *a priori* be represented in many different syntactical ways, in the cases that interest both the authors of SDRT and ourselves, they are represented by clauses, which contain verbs. The archetypal causal example, used in SDRT and in *relevance* papers, such as (Wilson and Sperber, 2004), is example 4.1, which, following Lascarides and Asher, is associated with the *push-fall law* that allows humans to understand it as causal. In this case, the semantically rich verbs represent an obvious feature of causation.

(4.1) Max fell. John pushed him.

Modelling world knowledge laws that represent causally connected eventuality types – the archetypal feature of causation – has been the focus of most previous work in computational linguistics that seek to automatically

recognise causation, whether this is the main purpose, or is a subtask of recognising discourse relations.

The late 2010s has seen a trend in automatic discourse relation classification, to learn, on explicit examples, features that could then be used to recognise implicit examples (Marcu and Echihabi, 2001; Blair-Goldensohn et al., 2007; Sporleder and Lascarides, 2007; Sporleder, 2007). An important feature used in these methods was *word pairs*. These word pairs were intended to represent world knowledge. Although early results seemed promising, doubts were cast on the effectiveness of the method, as it did not generalise well to naturally implicit examples (early results were evaluated on artificial implicit examples: explicit examples with their connective removed) (Sporleder and Lascarides, 2008). Recently, Pitler and her colleagues (2009) have shown that successful early results might have been due to a method artefact, and that we lacked proof that world knowledge could indeed be represented and extracted in this way (for more details, see 2.4.2).

Although all content words could bear the necessary world knowledge and be useful to recognise the expression of eventualities that are normally linked by a causal relation, in this chapter, we will focus on the archetypal event-representing words: verbs. We wanted to study how useful verbs are for recognising causal relations, and how they could be used for eventuality classification. In this chapter, we will show that although world knowledge is necessary to recognise causal relations, word pairs in general and verb pairs in particular do not carry this type of information, either because causal knowledge cannot be acquired from them in texts or because they are not adequate for recognising specific causally linked eventualities (we will develop these reasons more in chapter 5). In this chapter, we carry out an analysis of verb pairs in relation to causation. We study the frequency curve of verb pairs that appear in causal relations, showing that causal relations do not show any sign of restricting verb pairs compared to general text. We then compute a measure of the predictability of verb pairs for causation, showing that the predicting verbs are highly corpus-dependent and do not correspond to an intuitive understanding of world knowledge. We then show that, as our predictability measure becomes more reliable, it tends to show statistical independence between verb pairs and causation.

While we do not reject the idea that world knowledge is the main tool in causation recognition, we show, in this chapter, that this knowledge either is not encoded in verb pairs or cannot be statistically extracted from them in a big corpus. We will argue later in chapter 5 that causation recognition relies much more on general understanding of the world than on linguistic skills, and that natural language processing tools are not mature enough to capture this knowledge in texts. In this chapter, we focus on showing that verb pairs – a very intuitive and linguistically motivated causation feature – are not a reliable indicator of causation.

4.2 Methodology

In this section, we describe our methodology for analysing the predictability of verb pairs for causation. We will first describe the corpus that we used for our experiments, and then the way we extracted frequencies from the corpora. We will give results later: in section 4.3, we will give results on the distribution of frequency of verb pairs in causal and non-causal contexts, and in section 4.4, we will provide results on the predictability of verb pairs for causation. Finally, in section 4.4.3, we will give a more thorough analysis of verb pair predictability on the Penn Discourse Treebank and show that as its reliability improves, our predictability measure tends to show statistical independence of the verb pair and causation.

4.2.1 Description of our corpora

For the set of experiments described in this chapter, we use 3 different corpora, that we will now describe: the Penn Discourse Treebank, the Associated Press Worldstream part of the North American News Text corpus, and the juvenile literature part of the Gutenberg project.

The Penn Discourse Treebank (PDTB) (Prasad et al., 2008) is a corpus annotated with discourse relations. The base corpus is the 1 million word Wall Street Journal corpus, the same as was used as a base for the Penn Treebank (Marcus et al., 1994). The corpus is hand-annotated with several discourse relation that are classified on a hierarchic level in types and subtypes. Figure 4.1 gives the relation hierarchy for the contingency relation.

As can be seen in figure 4.1, the sense hierarchy has three levels. Causation is one of the middle level relations and is called *cause*. It has two subtype relations: *reason* and *result*, and its supertype is *contingency*. *Contingency* relations are defined as situations where a causal link exists between two eventualities (Prasad et al., 2007). It has three child relations: *cause*, *pragmatic cause* and *condition*. *Pragmatic cause* is used when an argument is given for an utterance. For example in (4.2), *his jacket is not on the chair* is an argument for the utterance *John is out*, not an eventuality causing the fact that John is out. *Condition* is similar to *cause*, but contrary to *cause*, it does not assert the truth value of the argument. Example (4.3) gives no indication whether John will leave or stay.

(4.2) John is out, his jacket is not on the chair.

(4.3) If John leaves, I am going to miss him.

We are thus interested in the *cause* relation. The PDTB corpus further classifies *cause* relations into *reason* and *result*. In result the cause appears first in the text, and the consequence afterwards, whereas *reason* exhibits an anti-chronological order. The annotations of the PDTB corpus contain explicit and implicit relations. In explicit relations, the connective is signaled. It is typically *because* for *reason* relations and *so* or *as a result* for *result* relations. For implicit relations, the explicit connective that the annotator has judged the best for the implicit relation is also indicated.

The raw Wall Street Journal on which the Penn Discourse Treebank is annotated weights about 13 MB, uncompressed, and contains about 1 million and 100 thousand words. The Wall Street Journal publishes financial, economics and business news in American English.

The Associated Press Worldstream part of the North American News Text corpus (APW), is a corpus of general domain news distributed by the Linguistic Data Consortium (LDC). It weights about 1.8 GB uncompressed and contains about 267 millions words.

The ChiC corpus (CHIldren Corpus) is a corpus of literature intended for a younger audience that we created from the Gutenberg project. We hypothesise that children's texts are simpler and present fewer phenomena

Figure 4.1: The contingency relation in the Penn discourse treebank

unrelated to causation and that makes its identification more complex than in their adults counterpart texts, while still presenting implicit causal relations. We thus wanted to compile a relatively large corpus of texts intended for children. In the case of verb pairs analysis, we also hypothesised that causation relations intended for children would be ontologically less complex, resulting in a more constrained selection of verb pairs.

Project Gutenberg (`http://www.gutenberg.org`) is a web site that publishes a large quantity of public domain books and makes them available in several formats, one of which is simple *.txt* files. Project Gutenberg provides lists of books by Library of Congress classes. This classification system provides a category for books intended for a younger public: *fiction and juvenile belle lettres* which represent the PZ class in this system. We automatically retrieved all books from this category from Project Gutenberg.

Our ChiC corpus contains about 5000 books. It contains, for example, *Peter Pan* by J.M. Barrie, *Treasure Island* by R.L. Stevenson or *Alice's Adventures in Wonderland* by L. Carroll. Although intended for a younger audience, the books might not be only for children, but also for teenagers. Since most books have fallen into public domain, some data are from a few centuries ago, and all English is not perfectly contemporary.

The ChiC corpus consists of about 500 MB uncompressed and contains about 85 million words.

These three corpora have several differences, particularly in the text styles (news and fiction) and in the way they allow us to represent causation, either using explicit causation for the unannotated corpora or several precise forms of causation –implicit, explicit, reason, result,...– for the Penn Discourse TreeBank. This variety allows us to test our hypothesis on the predictability of verb pairs on a large spectrum of corpora and representations, making the results more accurate as they are not corpus dependent. Table 4.2 gives an overview of the different corpora used in this chapter.

	APW	ChiC	PDTB
Source	LDC	Children texts selectively retrieved from the Gutenberg project	LDC
Annotations	none	none	discourse relations
Size	1.8 GB	500 MB	13 MB
Number of words	267 millions	85 millions	1 million
Type of texts	General news articles	children fiction	economic and financial news articles

Figure 4.2: Corpora used for predictability experiments with verb pairs

4.2.2 Modelling of causation

The goal of the experiments that we describe in this chapter is to compare verb pairs in causal and non-causal contexts, to see if causal contexts select specific verb pairs, or have a statistically significant impact on pair selection. We also wanted to quantify the predictability of specific verb pairs for causal contexts. This required us to classify the studied extracts into causal and non-causal occurrences. For the raw corpora that do not carry any annotation – the AP corpus and the ChiC corpus – the only indicators of causation that are mostly reliable are the explicit markers. We selected the connective *because*, because it is the archetypal causation connective and because it is relatively unambiguous regarding causation, although it can also mark pragmatic causes, such as in *Max is sick because I didn't see him at work today*, where the second clause is an argument in favour of the first and not the representation of a causing eventuality (for more details on these usages, see, for example, Moeschler, 2011). Predictable noise in such an automatically annotated corpus are false positives made up of pragmatic causation and false negatives made up of implicit causation and differently marked causation, as summed up in table 4.3. We claim, however, that this partition is the most

reliable possible in an unannotated corpus, and that such a partition should still exhibit statistical differences in the selected verb pairs if causation does indeed restrict its possible verb pairs.

	false positives	false negatives
nature	pragmatic causation	implicit and not because-marked causation
example	George is out because his jacket is not on the chair	Max fell, John pushed him

Figure 4.3: Predictable false positives and negatives due when using the presence of *because* as a model of causation.

The partition is a relatively easier task for the PDTB corpus, as it is annotated with discourse relations and with causation. The most obvious classification would be between *cause* relations and non-*cause* relations. We carried out several analyses of the PDTB verb pairs, analysing the following relations: *cause* relations, explicit examples containing *because*, *explicit cause* relations, *implicit cause* relations and *reason* relations.

We wanted to compare explicit and implicit *cause* relations, to see if verb pairs were more predictive in implicit *cause* relations (as implicit relations might rely on more obvious world knowledge). Examples with *because* were studied in order to allow us to compare the PDTB corpus with the other corpora, using the same modelling, and we studied the *reason* relation from the same perspective, as verb pairs would be in the same textual order for all reason relations.

4.3 Frequency curves results: a very long tail

In this section, we will present the results of our analysis of the frequency curves of verb pairs in causal and non-causal contexts.

4.3.1 Distribution of verb pairs in general corpora

To compute the general distribution of verb pairs in the different corpora, we counted pairs of verbs that appeared in the same sentence, the first one appearing before the second one, in the text. In order to make these calcu-

lations, we needed to sentence-split the corpora. To this end, we used the Lingua::EN::Sentence perl module by Shlomo Yona. This module uses regular expressions to split texts into sentences. It avoids mistakes by containing a set of built-in acronyms that would otherwise be likely to give false positive sentence splits.

We then lemmatised and tagged the text in order to recognise verbs and to count all conjugate forms of a verb as the same verb. To this end, we used *Tree Tagger* (Schmid, 1994). As a probabilistic tagger, *Tree Tagger* gives an analysis for each word in the text it analyses. It gives a part of speech tag and a lemma. We used all verb tags.

Zipf (1935) has shown that most words in a corpus are rare and that natural language corpora tend to follow a Zipf law, a family of laws where the frequency of a word is inversely proportional to its rank. In the present case, verb pairs follow quite closely a $\frac{1}{x}$ Zipf law; that is, if $maxF$ is the frequency –the number of occurrences– of the most frequent verb pair, then $\frac{\text{maxF}}{2}$ is the frequency of the second most frequent verb pair, $\frac{\text{maxF}}{3}$ of the third most frequent verb pair and, generally, the x^{th} most frequent verb pair has a frequency of about $\frac{\text{maxF}}{x}$.

Plotting the frequencies against the ranks makes a very unreadable stiff curve that follows the axis closely, as can be seen in figure 4.4 [1]. The resulting graphs are much easier to read when plotted in log-log scales, so we used this method to display our results. The axes are then not linear. In a linear graph, the space between 1 and 10 on an axis is ten times smaller that the space between 10 and 100. On a log scale, on the contrary, the space between 1 and 10 can be the same as the space between 10 and 100. In a log scale graph, the same distance on the graph represents a larger and larger distance between the plotted points, as one gets away form the origin of the axes.

Since most words in texts are rare, word pairs are even rarer. In the corpora that we studied, verb pairs show similar frequency curves for each corpus, as can be seen in figures 4.5, 4.6 and 4.7, which show the frequency curves in log scales. Each curve in log scale looks similar to the typical Zipf law line as the $\frac{\text{Max Occ}}{x}$ line, plotted together with our empirical results, show.

[1] In the following figures, ye+x means $y \times 10^x$. For example, 2e+06 means $2 \times 10^6 = 2000000$, that is, two million.

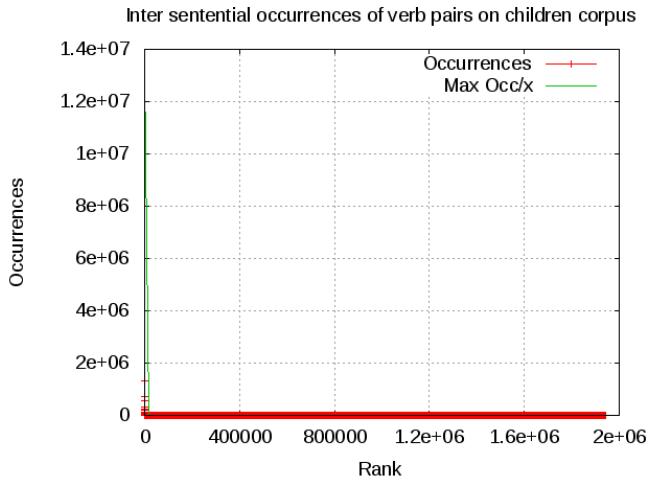


Figure 4.4: Total inter-sentential occurrences of verb pairs in linear scale on ChiC

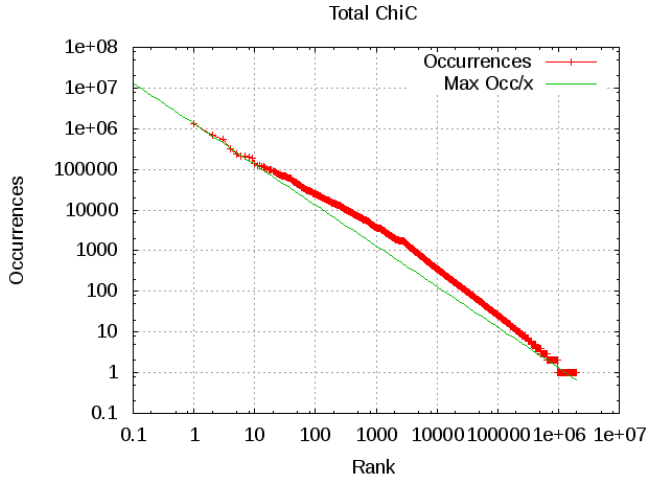


Figure 4.5: Total inter-sentential occurrences of verb pairs in log scale on ChiC

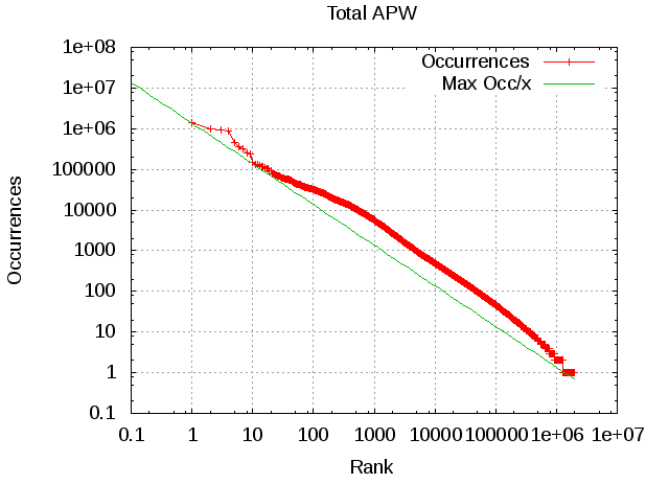


Figure 4.6: Total inter-sentential occurrences of verb pairs in log scale in APW

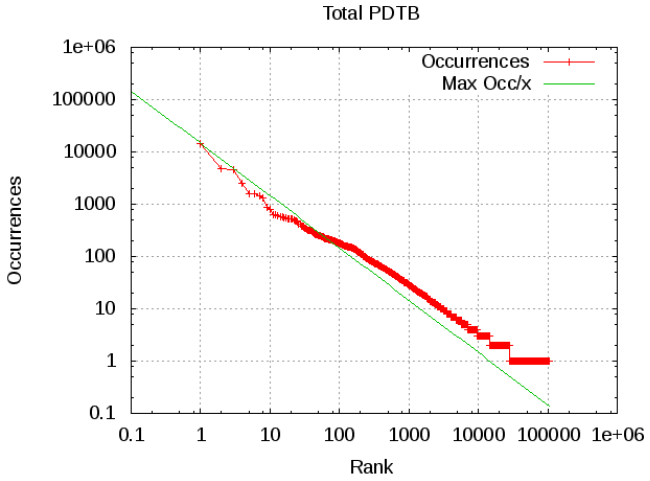


Figure 4.7: Total inter-sentential occurrences of verb pairs in log scale in PDTB

Figure 4.8 shows the top ten verbs in each corpus, without the pairs containing *be*, *have* or *do*, which may be auxiliaries. Not surprisingly, they are simple pairs of frequent verbs.

Rank	APW	ChiC	PDTB
1	say say	say go	say say
2	go say	go say	sell sell
3	say make	say say	make make
4	say take	come say	say make
5	make say	say come	rise rise
6	take say	go go	say expect
7	reserve distribute	know say	yield yield
8	get say	say see	price price
9	quote say	see say	expect expect
10	want say	go see	buy buy

Figure 4.8: Frequent general verb pairs.

4.3.2 Distribution of verb pairs in causal relations

We wanted to study the specific distribution curves of verb pairs in causal relations. Figures 4.9, 4.10 and 4.11 shows the number of causal occurrences of each verb pair for each corpus, using occurrences that appear with a *because* for the unannotated corpora and occurrences marked as *cause* for the PDTB. Apart from showing many fewer occurrences, the curves are similar to the ones for the general corpus. These verb pairs are sparse. This means that almost all of them appear very rarely in corpora. Most of them appear only once in the quite large corpora that we used for this study, and if we took all possible verb pairs in English, even more of them would not appear at all in the corpora. Because they are so rare, it is very difficult, if not impossible, to compute useful statistics for them. This sparseness of verb pairs in causation would make it more difficult to use them as supervised classification features, as a classifier would have to learn most pairs with a very limited number of occurrences.

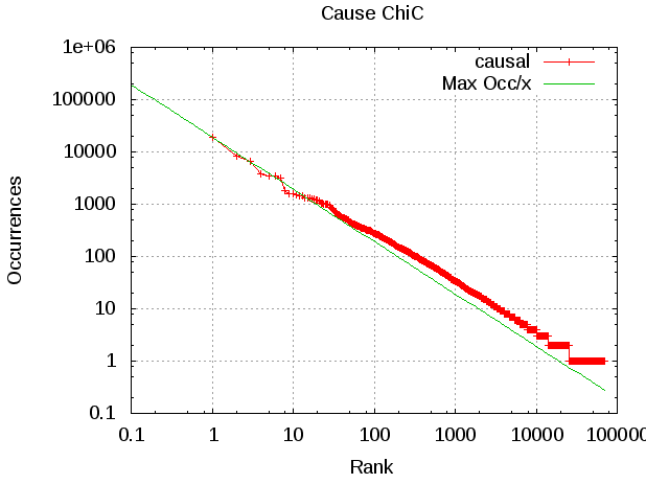


Figure 4.9: Causal occurrences of verb pairs in log scale on ChiC

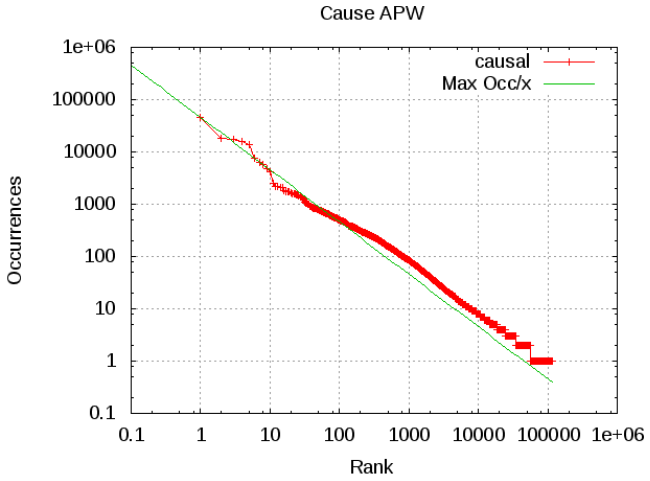


Figure 4.10: Causal occurrences of verb pairs in log scale in APW

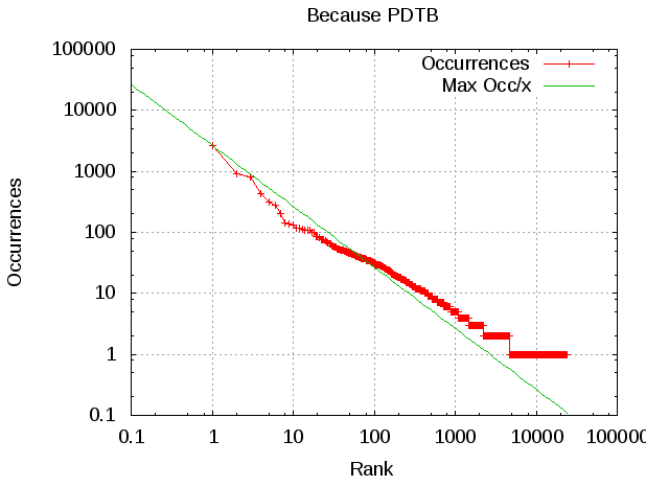


Figure 4.11: Causal occurrences of verb pairs in log scale in PDTB

Figure 4.12 shows the top ten verb pairs appearing in causal occurrences; that is, appearing with a *because* for the unannotated corpora and annotated as *cause* for the PDTB. They are generally frequent verbs. We do not observe any obvious change in distribution here either.

Rank	APW	ChiC	PDTB
1	say say	say know	get get
2	say want	say say	make make
3	go say	add change	go get
4	think say	say go	get say
5	say make	go go	want go
6	want say	change change	say say
7	say take	say get	say make
8	make say	say see	sell sell
9	come say	say come	buy buy
10	say go	say make	sell get

Figure 4.12: Top ten verb pairs in causal relations

4.4 Predictability

In this section, we describe how we computed the exact predictability of verb pairs for causation.

Although the distribution curves do not obviously differ between general inter-sentential and causal occurrences of verb pairs, we wanted to compute the predictability of each verb pair that appeared in causal occurrences in our corpora. We wanted to discover whether the verb pair appears more frequently with causation than without and whether it is in any way a clue for causation.

As we have seen before in chapter 2, linguistic theories that attempt to explain how humans can recognise implicit causation rely mostly on world knowledge. This means that some eventualities are known *a priori* to be typically causal. If verb pairs are good indicators of the eventualities that the clauses containing them represent, then they must be highly predictive of causation. If we take the order of the pairs into account, then some will be predictive of reason relations (anti-chronological order) and the same pair in the reverse order will be predictive of a result relation (chronological order). In our experiments, since we mostly used *because* as a causal marker, verb pairs should be predictive of reason relations.

4.4.1 Choice of a predictability measure

A predictability measure is essentially an association measure. It tests whether the clue appears significantly more often with the outcome than without. In our case, the clue is the verb pair and the outcome is the presence of causation.

Several statistical association measures exist. Tan and his colleagues give (2002) an overview of measures of interestingness of association patterns for data mining. They give a set of characteristics associated with each measure. Based on these characteristics, we chose to use the *Yule's Q* score, which is a normalisation off the *odds ratio* score because we wanted a score that is invariant to row or column scaling. That is, the total frequency of a verb pair should not have any impact on its interestingness classification as both a frequent or a rare pair might be a good predictor of causation. Similarly, the total number of causal relations, which varies from corpus to corpus, should

not have any impact on the ranking of predictability of the verb pairs that are found in this corpus. This property is rare, among the measures that Tan and his colleagues studied. It is only associated with variations of the odds ratio measure.

The odds ratio is computed in the following way:

$$\frac{a/b}{c/d} \tag{4.4}$$

In our case, for a specific verb pair v, a is the number of causal relations that contain the v, b is the number of causal relations that do not contain v, so that a/b is an indicator of how often causation appears together with v. Similarly, c is the number of times v appears outside of a causal relation and d is the number of sentences in which neither v nor a causal relation appears, so that c/d is an indicator of the predictability of the verb pairs for the absence of causation. Overall, the odds ratio compares the predictability of the verb pair for causation and for the absence of causation. If the verb pair predicts as much causation as its absence; that is, if the verb pair is not predictive at all, then the odds ratio is 1. Values greater than 1 indicate a positive association: the verb pair is predictive for causation. Values less than 1 indicate a negative association: the verb pair is predictive of the absence of causation.

Yule's Q is a normalisation of the odds ratio, and is computed as:

$$\frac{O-1}{O+1} \tag{4.5}$$

where O is the odds ratio. It varies between -1 and 1 where 1 represents the perfect association, -1 perfect negative association and 0 statistical independence.

Choosing this property means that the global frequency of a verb pair does not have any impact on its predictability score, allowing very rare verbs to top the predictability ranking. Because statistics on a few occurrences of verb pairs are not reliable, we computed the predictability of verb pairs that appear at least 10 times in a causal relation, for the big corpora, and 2 times for the PDTB. This is because the PDTB is relatively small, and preliminary results showed that a higher threshold gave a ranking of generally very frequent verb pairs only, resembling the ones shown in the table in 4.8.

This means that the PDTB results are much less reliable when compared to the other results because of the size of the corpus.

4.4.2 Verb pair predictability for the presence of *because*

After filtering out very rare verb pairs, we ranked them by Yule's Q score for each corpus. Verb pairs that appeared only in causal relations obtained a maximal Yule's Q of 1. We further ranked them, by their number of causal occurrences (c.o,), so that verb pairs that were assigned the lowest ranks were perfectly correlated with causation as well as frequent in the analysed corpus. The table in figure 4.13 shows the lowest ranked 20 verb pairs for each corpus, together with their Yule Q score and their number of causal occurrences in the corpus.

APW	Q	c.o.	ChiC	Q	c.o.	PDTB	Q	c.o.
plead mock	1	20	praise flow	1	16	set reinforce	1	2
initial link	1	18	praise toss	1	16	mean package	1	2
track smooth	1	18	add suprise	1	14	go perceive	1	2
prefer smooth	1	18	change suprise	1	12	cause escape	1	2
achieve intrude	1	17	flow toss	1	12	overstate include	1	2
question complicate	1	17	praise grow	0.99	12	have deposit	1	2
erode define	1	16	close change	0.99	45	want denounce	1	2
honor wage	1	15	grow toss	0.99	16	go reinforce	1	2
puncture shut	1	14	blow flow	0.98	16	denounce denounce	1	2
disconnect shut	1	14	remove change	0.98	135	qualify do	1	2
balance guarantee	1	13	sing toss	0.98	12	have man	1	2
test cheat	1	12	close add	0.98	24	package package	1	2
punish submit	1	12	add change	0.98	315	figure lose	1	2
disillusion test	1	12	blow toss	0.97	16	lead invest	1	2
quit dismember	1	12	remove add	0.97	72	mention die	1	2
surface guard	1	12	sing flow	0.97	12	be sense	0.99	2
defect feed	1	12	happen change	0.96	49	accrue have	0.99	2
bar overload	1	12	add add	0.96	169	go tie	0.99	2
interfere submit	1	12	add replace	0.95	14	trouble have	0.99	2
fall congest	1	11	add remove	.95	35	mention want	0.99	2

Figure 4.13: Top 20 predictive verb pairs for the presence of *because*

The general idea underlying these experiments is to verify whether verb pairs can represent eventualities that are highly causally linked in the world. Previous linguistic theories state that the main causation feature is world knowledge; that is, the knowledge that some specific eventualities in the world tend to cause some other specific eventualities, the *push-fall* causal law is often given as example, as we have seen in section 2.2. The question is then whether the predictive verb pairs presented here are representations of world knowledge causal laws between eventuality pairs.

The pairs are presented in their textual order. In the original text, the causal pairs appeared with a *because* in between them. This means that when judging their intuitive causation power, one should either insert a *because* in between, or judge them in the reverse order that corresponds to the temporal order of the eventualities expressed by the verbs. It is probably possible to construct some very specific contexts were some pairs do seem to enter a causal relation. In our own judgement, however, the verbs pairs do not reflect highly causal eventuality pairs.

The large number of words that might, without context, be assigned several part of speech tags, such as *honor* or *man* might indicate difficulties in the part of speech tagging.

The predictive verb pairs are extremely corpus dependent. No pair of verbs appears twice in the top 20 predictable pairs of two corpora. In fact, none of the 20 best pairs of any corpus appear in the ranked pairs of another corpus at all, because they appear either less often than the threshold in the other corpora, or do not appear at all in causal relations outside of the tested corpus.

4.4.3 PDTB predictability

In the above section, we presented results for the predictability of verb pairs for causal relations. To represent causal relations, we used sentences containing the word *because*. This modelling has several drawbacks. First, all representations of causation in a natural language text do not involve the use of the word *because*. There are many occurrences of implicit causation that do not make use of any marker at all, such as in (4.6), and there are causation relations that are marked with other words such as *as a result* or by much more ambiguous words such as *and*, as in (4.7). These relations are not captured by searching for *because* occurrences, and they are part of false negatives of such a modelling.

(4.6) Max has wet hair, he forgot his umbrella.

(4.7) John was sad, and he cried.

Moreover, not all instances *because* mark pure causation. Some mark pragmatic causation, such as in (4.8), where the second eventuality is not

the cause of the first, but the cause of the utterance of the first. Because of a lack of pure causation between the expressed eventualities, these cases are parts of the false positive of such a modelling.

(4.8) Are you looking for me? Because I'm right behind you.

In the absence of any human annotation, we claim that *because* as a marker of causation is the best possible choice, because it give few false positive and is a quite reliable causal marker. However, in this study, the PDTB was available to us, and this corpus is annotated with causation, allowing us to make a much finer classification of causal versus non-causal relations. In this section, we present the results of our predictability experiments on different partitions of the PDTB. We wanted to determine how much the choice of the sectioning affects the predictability results.

For the PDTB, we experimented with the following partitions: *cause* relations, explicit examples containing *because*, *explicit cause* relations–occurrences of explicitly marked causal relations, the marker being either *because* or something else–, *implicit cause* relations and *reason* relations. *Cause* relations contain all causal relations regardless of the chronological or anti-chronological order of the two eventualities in the relation. These relations can be marked, unmarked, or marked with an ambiguous marker. Using the *cause* versus non *cause* partition has the advantage of being very accurate and the disadvantage of being very different from the unannotated corpora partitioning, and of being blind to the eventuality order, so that the same relation expressed as a reason or as a result will produce two different verb pairs, one for each eventuality order.

The *because* versus non-*because* partitioning is exactly the same as the other corpora partitioning and has the advantage of being comparable to them and the same disadvantages; that is, it contains a lot of false negatives (causal relations expressed in any other way) and some false positives (pragmatic causes). Example (4.9) gives an example of a *because* relation from the PDTB.

We also tested with *reason* relations only, whether marked or unmarked. We had two reasons for this choice. First, *reason* relations present verb pairs in the same order as examples containing *because* and thus would be more comparable to other corpora. Second, as we detail more in chapter 5, we argue that reason relations are more defined and clearer in regard

to causation when compared to *result* relations. We thus hoped that verb pairs used in *reason* relations might be more predictable than those of *result* relations. Example (4.10) shows an unmarked *reason* relation, which can be made explicit using *thus*.

We tested verb pair predictability on implicit examples only, because we hypothesised that implicit relations might present more causally linked eventualities. Our reasoning is that marked relations might be marked precisely because they would be difficult to interpret without the marker. Conversely, implicit relations do not need a marker to be interpreted, and thus might hold between eventualities that are more generally causal than are their marked counterparts (For a much more thorough analysis of this phenomenon, see Moeschler, 2007). We thus wanted to see if verb pairs in implicit relations are more predictive of the relations than their marked counterparts. Example (4.11) shows an implicit relation from the PDTB (the annotators state that the best insertable connective in this example would be *as a result*).

We also wanted to test the predictability of verb pairs on explicit examples only, to compare them to implicit examples. Example (4.12) shows an explicit example from the PDTB.

(4.9) Typically, money-fund yields beat comparable short-term investments **because** portfolio managers can vary maturities and go after the highest rates.

(4.10) In mid-October, Time magazine lowered its guaranteed circulation rate base for 1990 while not increasing ad page rates; with a lower circulation base, Time's ad rate will be effectively 7.5% higher per subscribe.

(4.11) In July, the Environmental Protection Agency imposed a gradual ban on virtually all uses of asbestos. By 1997, almost all remaining uses of cancer-causing asbestos will be outlawed.

(4.12) **As** individual investors have turned away from the stock market over the years, securities firms have scrambled to find new products that brokers find easy to sell.

The table in 4.15 gives a overview of the 10 most predictive verb pairs for each partitioning. The verb pairs were extracted from the same corpus,

and a high level of intersection is seen between the different partitioning. All but the *because* partitioning are a subset of the *cause* partitioning as shown in figure 4.14. The *because* partitioning still has a high intersection with *cause* that is made of all the causal usages of *because*. The not intersecting occurrences come from pragmatic usages of *because*. Not surprisingly, then, some highly predictive verb pairs are shared between the different partitioning. Despite this fact, a very high degree of variation still remains. Of the 44 distinct verb pairs shown in the table, 10 appear in more than one partitioning.

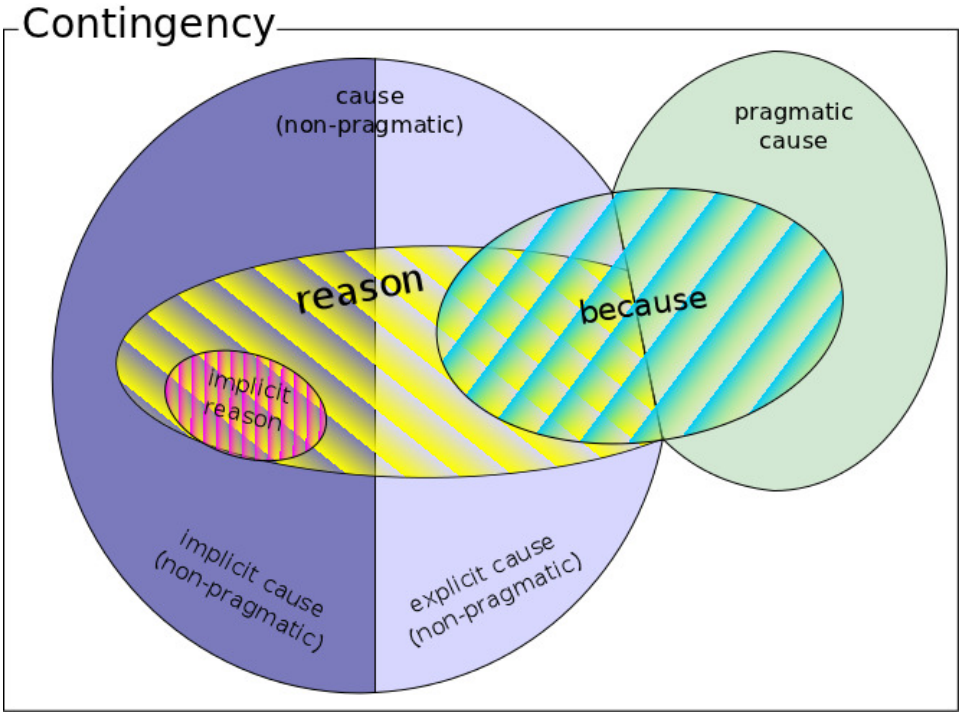


Figure 4.14: Intersections and unions of several PDTB partitioning

because	cause	reason
set reinforce	invoice be	accord concentrate
mean package	flash flash	begin spur
go perceive	accord concentrate	become contend
cause escape	operate need	plan believe
overstate include	contain pay	know disclose
have deposit	reject act	contain pay
want denounce	have chase	reject act
go reinforce	schedule buy	create rank
denounce denounce	create rank	face mean
qualify do	be annualize	meet knock
explicit cause	implicit cause	implicit reason
close react	involve seek	whack educate
cataloge be	like develop	think stop
begin spur	continue press	perform own
have exert	have pool	hedge fall
help double	invoice be	plan believe
know disclose	flash flash	contain pay
repay roll	get stomp	reject act
be reposition	plan believe	latch look
produce serve	contain pay	create rank
receive condemn	learn enter	face mean

Figure 4.15: Predicitve verb pairs in the Penn Discourse Treebank

The more examples used to compute a statistic, the more reliable is the statistic. In our case, a measure of the reliability of the predictability statistic is the number of causal occurrences of a verb pair. If verb pairs were generally good predictors of causation, then accurate measures should show high predictability. That is, frequent verb pairs would have a high Yule's Q We observe the reverse.

In figures 4.16, 4.17, 4.18, 4.19, 4.20 and 4.21, we plotted Yule's Q: the predictability measure of each verb pair, against the number of causal occurrences on which this measure was computed. Because most frequencies are very low but some are very high (because of the long tail effect), we plotted the frequency in log scale. Each figure shows a different partitioning of the PDTB.

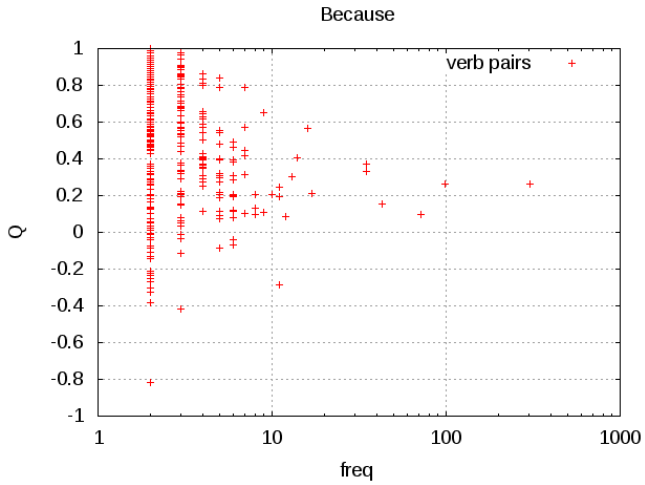


Figure 4.16: Frequency vs Q on the *because* partition

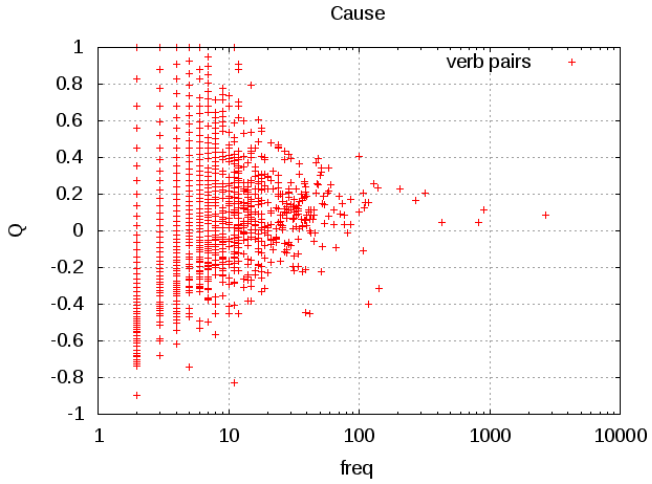


Figure 4.17: Frequency vs Q on the *cause* partition

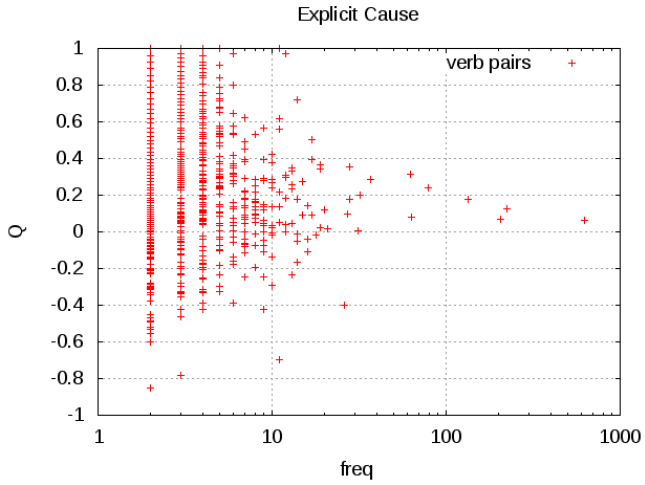


Figure 4.18: Frequency vs Q on the *explicit cause* partition

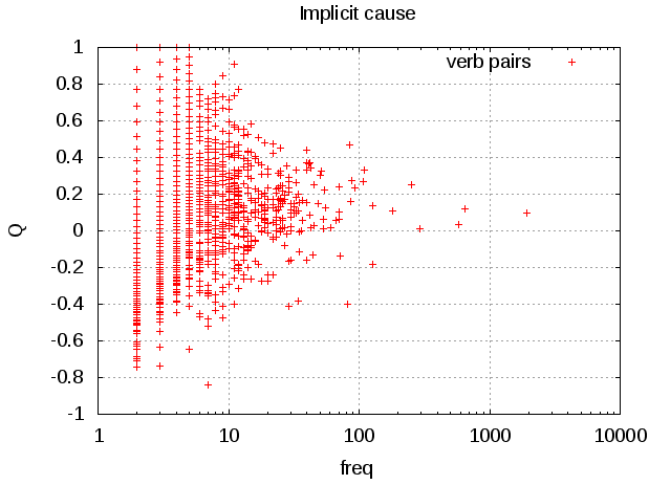


Figure 4.19: Frequency vs Q on the *implicit cause* partition

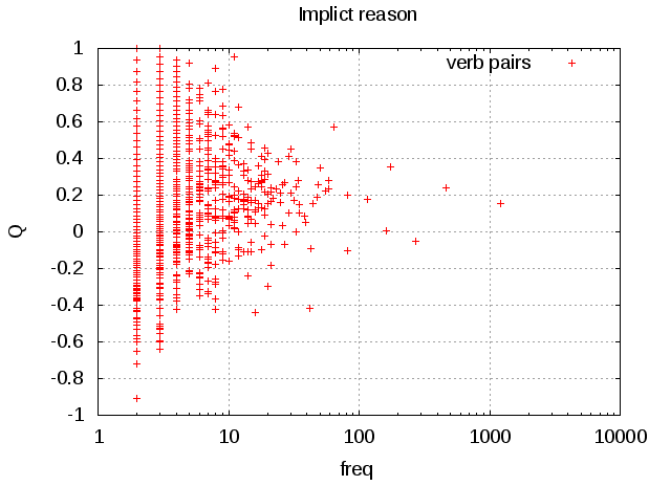


Figure 4.20: Frequency vs Q on the *implicit reason* partition

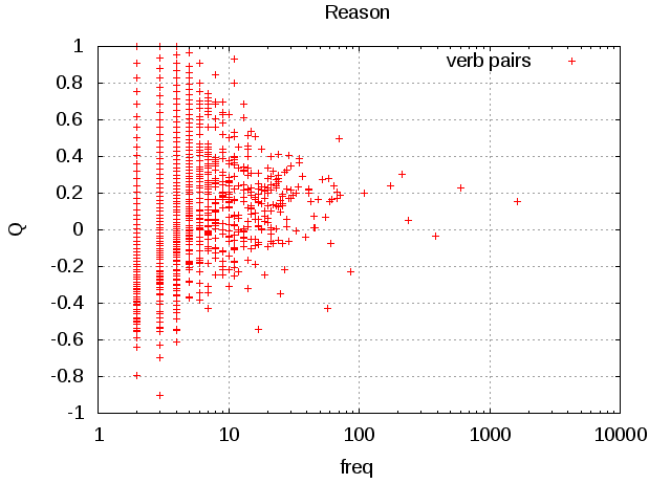


Figure 4.21: Frequency vs Q on the *reason* partition

Negative values of the Yule's Q score indicate negative association: verb pairs that present this score are more predictive of the absence of causation than of causation. A Yule's Q of 0 indicate statistical independence. The figures plot only Yule's Qs of verb pairs that appear at least twice in a causal relation. This means that verb pairs that do not appear at all in a causal relation and thus are very predictive of its absence are not represented in these figures. In our case, verb pairs that present negative Yules's Q scores, even though they might effectively be used by machine learning algorithms, are difficult to interpret in a linguistic way.

What interests us in these figures is the fact that as the frequency of a verb pair goes up, its Yule's Q score tends 0. That is, as the reliability of the statistical measure goes up, the predictability of the verb pairs tends to 0. Verb pairs that can reliably be measured tend to show statistical independence with causation.

4.5 Some classification experiments

In this section, we will present some experiments with an automatic classifier for implicit causation recognition. We used the Boostexter (Schapire and Singer, 2000) classifier to test several classification features. Boostexter uses a boosting algorithm (basically, a feature voting scheme) and is highly adapted to natural language processing tasks, offering built-in features such as the possibility to use n-grams. It is also the classifier used in some previous work (Sporleder and Lascarides, 2007; Sporleder and Lascarides, 2008). We used an existing classifier; however all classification features needed to be automatically extracted from the texts to classify. We programmed all features extraction in PERL.

We used Boostexter to classify items using lexical features only. We used content-word pairs as well as pairs of WordNet (Miller, 1995) hypernyms of content words. We evaluated all classifiers on 1400 implicit occurrences from section 20 to 24 of the PDTB –700 causal and 700 non-causal items. We trained the classifier either on 200,000 occurrences of the APW corpus – 100,000 occurrences generated from extracts originally containing a *because* and 100,000 from extracts originally containing a *but*– or on 6400 occurrences from sections 0 to 19 of the PDTB – 3200 implicit *cause* examples and 3200 non-cause.

The different one-feature classifiers are: chains of content words from each clause, where we used up to 3-grams, chains of WordNet hypernyms, also with up to 3-grams, and pairs of content words or of hypernyms. In the latter case, we hard coded all possible combination of pairs with one word from each clause, as a feature. Our hope in using WordNet was to reduce the sparseness of the lexical features. We took one hypernym per content word, the first one, that is the most frequent one. If WordNet did not provide any hypernym for the word, we input the word itself, instead. We used

Tree Tagger (Schmid, 1994) to determine which are the content words, and the WordNet perl module[2]. The subroutines that lemmatize clauses and get content words hypernyms from strings of text are given as annexe as example (D).

Table 4.22 shows the results of each classifier. Although always a bit better than the random 50% baseline, they are far from satisfactory. We obtained each result by running 1000 rounds of classification (that is 1000 refining of the classifier) and choosing the best round on the evaluation set. The evaluation set always gives much worse results than the training set, showing an over-fitting of the data. Given the slightly better results achieved when training on PDTB implicits, we give the pair results only for this kind of training.

	Best round number	training errors	evaluation errors
APW content words 1-grams	2	0.455	0.428
APW content words 2-grams	264	0.454	0.307
APW content words 3-grams	497	0.454	0.286
PDTB content words 1-grams	750	0.405	0.198
PDTB content words 2-grams	336	0.414	0.287
PDTB content words 3-grams	145	0.416	0.360
APW hypernyms 1-grams	2	0.455	0.428
APW hypernyms 2-grams	246	0.451	0.312
APW hypernyms 3-grams	279	0.453	0.307
PDTB hypernyms 1-grams	443	0.391	0.262
PDTB hypernyms 2-grams	357	0.399	0.287
PDTB hypernyms 3-grams	986	0.399	0.143
PDTB pairs of content words	881	0.462	0.188
PDTB pairs of hypernyms	882	0.457	0.165

Figure 4.22: Automatic classification results

[2] `http://people.csail.mit.edu/jrennie/WordNet/`

4.6 Conclusion

In this chapter, we described various experiments involving the predictability of verb pairs for causation. Our reasoning was that if, as linguistic theories predict, the recognition of implicit causation is mainly dependent on the recognition of two eventualities that are generally causal in the world, then it might be possible to capture those eventualities in the form of predictive verb pairs. Our idea was that, as the kernel of clauses representing eventualities, verbs should be representative of eventualities and thus, highly causally connected eventualities would lead to verb pairs being predictive of causation.

We used three different corpora: a general news corpus, a children stories corpus, and the much smaller Penn Discourse Treebank, which is manually annotated with discourse relations, even when they are implicit, one of this discourse relation being causation. In the larger unnannotated corpora, we modelled causation using the presence of the explicit marker *because*.

We showed that the distribution of verb pairs in causal relations is similar to the one of inter-sentential verb pairs in general. This is a clue that causal relations do not restrict verb pairs to a subset of the general language.

We computed a measure of verb pair predictability, using the Yule's Q statistic and we presented the most predictive verb pairs for each corpus and for several partitioning of the PDTB. We showed that predictive verb pairs do not correspond to intuitively causal eventualities and that they are extremely corpus dependent, and, to a lesser degree, partition-dependent, which is a clue that they might be over-fitting the task of recognising causal from non-causal relations on the specific corpora that we performed our experiments on. That is, they cannot be used for generalisation as they are artefacts of our setting.

We then showed that, as the reliability of our statistic grows, the measure tends to show statistical independence between the verb pairs and the presence of causation.

Finally, we present some classification experiments and show that the classifier at early stages over-fits the training data and does not give satisfactory results on the evaluation data.

The similarity of the distribution curves of verb pairs in causal usages

and in general language, the strong corpus and even partition dependency of the predictable pairs and the fact that, as the reliability of the sample grows, the predictability of verb pairs tends to 0, are strong indicators that verb pairs are not good predictors of causation in general. Even such a simple semantic clue does not seem to be learnable on the reasonably large corpora that we used for this experiment.

We claim that verb pairs are not good predictors of causation, and we are confident that similar experiments with any kind of open class word pairs will lead to similar results. We will give a fuller linguistic explanation to this belief in chapter 5. However, we argue that this lack of predictability of verb pairs can be blamed on the fact that verb pairs are not sufficient to determine the nature of the represented eventualities. Actually, it is extremely difficult, and perhaps even impossible, given the current state of the art, in the current state of the art, to reliably classify the expression of eventualities to the level of granularity necessary to represent general causation rules. We will argue that these rules rely on much coarser grained eventualities than the ones generally linguistically represented in texts.

Chapter 5

World knowledge necessary for implicit causation recognition: a manual analysis of causal occurrences in French children's tales

5.1 Introduction

In this chapter, we will present a manual analysis of the world knowledge (the set of *causal rules*) that is necessary to recognise implicit causation. As we have described in chapter 2, the two main linguistic theories that provide a model of how humans can recognise implicit relations rely mostly on world knowledge. These theories are the segmented discourse relation theory (SDRT) (Lascarides and Asher, 1993), and relevance theory (Wilson and Sperber, 2004). SDRT is a general theory that describes the way humans can recognise discourse relations. Causation relations are similar to the *result* and *explanation* relations that are part of these general discourse relations. Relevance theory describes several types of pragmatic enrichment of what is said. The temporal ordering of eventualities is one such enrichment and can take the form of temporal order in the text, where the eventualities follow, in the text, their order in the world or of causal inversion, where the

consequence – the temporally second eventuality – appears before the cause in the text.

Both SDRT and relevance theory rely on the idea that a common world knowledge exist that is necessary to recognise some implicit relations. This is also the case of Moeschler's directional inferences model (Moeschler, 2000a). However, these theories do not describe the specific nature of this world knowledge in much detail. The typical example – *Max fell, John pushed him* – is only interpretable using the knowledge that *pushing* normally causes *falling*. In their seminal paper on SDRT, Lascarides and Asher (1993) state that the following law is necessary in such a situation (their wording, page 7):

> **Push Causal Law:** If e_1 where x falls and e_2 where y pushes x are e-connected, then normally, e_2 causes e_1.

In this law, e_1 and e_2 are eventualities [1]. Two eventualities are *e-connected* if a discourse relation holds between them.

Following these linguistic theories, world knowledge is necessary for the human recognition of implicit causation. We claim that world knowledge would also be a necessary constituent of an automatic system. This knowledge might be implemented and represented in the system in many different ways. It might be a simple clue in a statistical machine learning system, it might be a part of a complex rule-based system, it might be hand coded or it might be automatically learnt. In any case, we argue, following linguistic theories, that it is an indispensable part of an automatic system, whether it is a statistical automatic learner or a rule based programme, and we hypothesise that it is the current bottleneck in automatic causation recognition systems.

World knowledge has been a part of modern systems, generally implemented as a word pair clue in machine learning programmes (Marcu and Echihabi, 2001; Sporleder and Lascarides, 2007; Pitler et al., 2009). Other authors simulated the way a human could learn this world knowledge from experience by calculating the joint probabilities of temporally ordered eventualities (Beamer and Girju, 2009). In the Beamer and Girju system, eventualities are represented by verb pairs. However, as Pitler and her colleagues

[1] For more details on eventualities, see (Reboul, 2000) who proposes a typology of eventualities as well a model for their cognitive representation.

(2009) have shown, the relative success of word pair based systems might be an artefact of the artificial construction of implicit discourse relation for their training and evaluation. In addition, as we have shown in chapter 4, verb pairs are not directly predictive of causation in natural language texts. If world knowledge is necessary, its exact nature is unknown.

In this chapter, we will study the exact nature of this necessary world knowledge. To this end, we manually annotated and studied a small corpus of fictional work intended for very young children. We chose this corpus because we hypothesise that causation might be linguistically and ontologically simpler in texts intended for a young public, which, by simplifying the analysis of interacting but non causation-related phenomena, would render this study easier. As we will describe later in this chapter, the study of causal relations and of the world knowledge that they necessitate for their understanding, even in this simple corpus, is already much more complex than in *ad hoc* constructed utterances.

This chapter is organised in the following way. We first describe our corpus in more detail, giving some statistics on the number of causal relations of different types in section 5.2. We then argue that *reason* relations, where the cause follows the consequence, are much more well defined and, by giving clearer intuitions, easier to annotate than *result* relations, where the cause precedes the consequence (section 5.3). Given this fact, we carried out manual analysis on reason relations only.

We then give a sketchy algorithm of the necessary steps to automatically use world knowledge for the interpretation of causation (section 5.4). We later focus our analysis on causal rules and analyse them in two different dimensions: their causal chain span (section 5.5.1), and the level of granularity of the eventualities that they connect (section 5.5.2). Finally, we give an analysis of different types of foreseeable difficulties that the acquisition and use of causal rules entails (section 5.5.2), arguing that, given the amount and type of necessary world knowledge, the current state of the art in computational linguistics does not allow us to reach accurate classification of causal versus non-causal implicit relations, based on world knowledge, the most salient clue in previous linguistics works.

5.2 Corpus description

In this chapter, we want to give an analysis of the exact nature of world knowledge necessary to recognise implicit causation. To this end, we studied a corpus of fiction intended for young children, which we will describe in this section.

Our reasoning for using such a corpus was that texts intended for young children would be simpler than their adult counterparts in many aspects, such as the length of the clauses, the required general world knowledge or the syntax, while still presenting fully developed implicit causal relations. These implicit causation relations might also be ontologically simpler. We reasoned that, as a young child has a more basic understanding of the world, the relations that are implicitly expressed in text that they can understand would require less and simpler world knowledge for the recognition of implicit causation [2].

This thesis aims at studying the automatic recognition of causation in a general way, and not in a specific language, and we believe that the nature of the necessary world knowledge is not a specificity of a given language. Since we focus our study on French, and since our own intuitions are better for French, which is our mother language, we used a French corpus. We do not doubt, however, that our findings would be similar for other languages – at least for a close language such as English. We even believe that they are completely language independent, because we hypothesise that causation is more of the domain of human psychology than a linguistic phenomenon, following previous authors (see, Moeschler, 2003a; Talmy, 1988a, for example). This hypothesis should be empirically tested but this testing is not within the scope of the present work.

Our corpus is a set of 33 independent tales intended for children 3 to 7 year old. The tales themes are *tales from all the world* for 16 of them. This includes traditional tales from different countries such as Japan, China or Russia or made-up tales that take place in countries that would be judged exotic by people coming from a French-speaking European country. The tra-

[2] One could hypothesize that children learn world knowledge by recognizing causation in the world as well as in linguistic forms, where causation can be recognized and learned, for example from stories, and added to the existing world knowledge. In this case, a younger child, having been exposed to fewer causal occurrences might have a more basic world knowledge than an older child.

ditional tales are rewritten and adapted for French-speaking young children. The other 17 tales have a *kings, queens, princes and princesses* theme. Each tale has a different author and is a few pages long, with plenty of illustrations. The following is an extract of it (our translation).

> Eisa soulève une oreille du dragon et elle crie de toutes ses forces :
>
> - HOU, HOU, DRAGON !
>
> Le dragon ne bouge pas. Eisa passe par-dessus le dragon et elle ouvre la porte de la caverne. À l'intérieur, il y a le prince Ronald. Le prince Ronald regarde Eisa et il lui dit :
>
> - Eh bien, Eisa, vous voilà dans un bel état ! Vous sentez le brûlé, vous êtes décoiffée et vous portez un affreux sac en papier. S'il vous plaît, Eisa, revenez quand vous aurez l'air d'une princesse !
>
> Eisa répond : - Mon cher Ronald, vous sentez la lavande, vos cheveux sont parfaitement coiffés et vos habits sont merveilleux. Vous avez tout à fait l'air d'un prince, mais vous n'êtes qu'un rien du tout et je ne vous épouserai jamais !
>
> Eisa lifts the dragon's ear and she screams as loud as she can:
>
> -HOU, HOU DRAGON!
>
> The dragon does not move. Eisa gets over the dragon and she opens the door to the cave. Inside, is Prince Ronald. Prince Ronald looks at Eisa and he tells her :
>
> -Well, Eisa, you are in quite a state! You stink of a burning smell, your hair is a mess and you are wearing an ugly paper bag. Please, Eisa, come back when you look like a princess!
>
> Eisa answers: - My dear Ronald, you smell of lavender, your have a perfect hairstyle and your clothes are wonderful. You look perfectly like a prince, but you are nothing and I won't ever marry you!

The complete corpus is about 22 000 word long and 137 KB, with tale length relatively even.

We annotated all causal relations between clauses, whether they were marked or implicit. We annotated the corpus with two different types of

	marked	unmarked
reason	21	20
result	51	180

Figure 5.1: Different types of causal relations in French tales

causal relations: reasons and results. The only difference between the two is the order of the eventualities in the text. Reasons present the consequence first and the cause second, whereas results present the eventualities in the chronological order. We annotated adjacent clauses only. We used the annotation manual described in chapter 3, relying mostly on linguistic tests for implicit relations. We annotated causation only if it stood between two eventualities represented as clauses in the text. Each clause had to contain at least one conjugated or gerund verb. The cause and the consequence had both to be true. For example, we did not annotate conditionals such as example (5.1) where it is not known whether the eventualities will actually happen or not.

(5.1) If you behave, you'll get double serving of chocolate cake tonight.

If there was a marker for the relation, even an ambiguous one, we identified it.

In this corpus, we identified 231 result relations and 41 reason relations. Table 5.1 shows the numbers of different types of causation occurrences in the corpus, *Unmarked* denotes causal occurrences that were either completely implicit or highly ambiguously marked. For *reason*, the ambiguous markers are *':'* and the gerund. For results, they are *:*, gerund, *et/and*, *quand/when*, *comme/as*, *dès que/as soon as*, *lorsque/when*, *puis/when*, *chaque fois/each time*, *ainsi/so*, subordinate and *ça y est/that's it.*

Table 5.2 shows the numbers of each marker for each relation. The sign ":" represents the use of the colon as a causal marker. Usages of *puisque/because* as a result marker are sentence initial uses of the connective. Unsurprisingly, the distribution shows a long tail of ambiguous markers.

marker	reason (ante chronological order)	result (chronological order)
implicit	17	100
car/because	12	0
parce que/because	7	0
:	2	2
gerund	1	10
puisque/because	1	3
c'est parce que/it is because	1	0
et/and	0	41
alors/so	0	20
si ... que/so ... that	0	15
quand/when	0	12
comme/as	0	9
tellement ... que /so ... that	0	4
aussi/so	0	2
dès que/as soon as	0	2
lorsque/when	0	1
c'est pourquoi/that's why	0	1
c'est ainsi que/this is how	0	1
causative verb	0	1
puis/then	0	1
chaque fois/each time	0	1
si bien que/so that	0	1
ainsi/so	0	1
tant ... que/so ... that	0	1
subordinate	0	1
ça y est/that's it	0	1

Figure 5.2: Causal markers in French tales

5.3 Textual order of the eventualities

We annotated two different types of causation depending on the order of the eventualities in the text: reason and result. This annotation style follows the one of the Penn Discourse Treebank (Prasad et al., 2008), where the *cause* relation is also subdivided into *reason* and *result*.

There is, however, an important difference between *reason* and *result*. The interpretation of reason relations as causal with a correct attribution of the cause and consequence eventualities is often the only way to correctly interpret the temporal relation between the segments, whereas, in result relations, the chronological order is that of narration; that is, temporal order. For example, in the famous *Max fell, John pushed him*, verb tenses do not give any clue of the temporal order of the eventualities, and the recognition of causation is necessary to interpret the correct eventuality ordering: reverse temporal order.

(Moeschler, in press, Moeschler et al., 2006) showed that a cognitive difference exists in the treatment of implicit causal relations when expressed in temporal or reverse temporal order. This difference is reflected in differences in reading times for pairs of causally related clauses in chronological and ante-chronological orders. The strength of association of the cause and the consequence segments also plays a role in reading time differences. Reason and result relations are thus cognitively different.

When annotating the corpus, we found that reason relations lead to clearer judgements. Although a quantitative experiment with several annotators would be necessary to assert this fact, it can be explained by the idea that a continuum exists between pure narration and pure result relations, whereas the need to make temporal order clear leads to a real dichotomy between reason and other relations.

Theories diverge when proposing a complete set of discourse relations. However, if we take into account the relations studied in (Lascarides and Asher, 1993), which are said to be important for temporal interpretation, we have to consider the following set of 5 relations: *explanation* which is similar to reason, *narration*, *elaboration*, *background* and *result*. Narration is the simple description of temporally ordered eventualities such as in (5.2). Elaboration relations are relations where a global eventuality is first stated,

then sub-eventualities that are part of it are described, as in example (5.3) where elaboration stands between the first two sentences and then narration between the sub-eventualities. Background relations give information about a surrounding state to an eventuality, such as in (5.4).

(5.2) Mary ran for 45 minutes, came back home and took a shower.

(5.3) John cleaned the whole house. He first washed the bathroom. Then the dishes, ...

(5.4) Max jumped into the pool. It was ice-cold.

In these relations, result and narration present the same temporal order: chronological order. *Explanation* is not temporally clear, but the similar reason relation present eventualities in the consequence-cause order, of which we know that cause can never follow consequence (partial or even total overlap is possible, though). It is similar to elaboration and background. However, in our experience, confusion arises much more between narration and result than between reason and other relations with similar temporal order.

While annotating our corpus, we found several instances where it was difficult to make a clear judgement between narration and result. Some instances are very clearly results. Example (5.5) shows a result relation that we found very clearly causal, and example (5.6) shows 3 narration related clauses.

(5.5) À midi, le soleil est très haut, et tout reluit dans le village..
At noon the sun is very high and everything shines in the village.

(5.6) Vite, Ali et Zora remplissent un gros sac de fils, ils versent l'eau des seaux dans une outre, et ils serrent une poignée de blé dans une bourse.
Quickly, Ali and Zora fill up a big bag with threads, they pour the water into a goatskin and they clench a handful of wheat into a purse.

However, some instances were more difficult to classify. Example (5.7) shows an instance leading to such a difficult decision. Here, the counterfactual test holds: the messenger would not have bowed if he didn't arrive, but intuitively, it is not a cause, but more of an enabling condition, and no causal

law seems to rely a messenger entering a room running and him bowing in front of the emperor. However, when activating a causal law such as (5.8), the eventualities look more causal with the causal chain in (5.9). This can be made more clear when removing particularities of the eventualities that are not part of the causal law and adding information that is part of the causal law as in (5.10). Therefore, the same pair of eventualities can be seen as narration or result depending on the focus of their textual representation. In (5.7), the focus is on the description of the eventualities, with the precision that the messenger arrives running and that the bowing is low, whereas in (5.10), the description of the same eventualities follows more closely the causal rule, and can be more easily classified as a result relation.

(5.7) Un messager arrive en courant. Il s'incline très bas devant l'empereur.
A messenger turns up running. He bows low in front of the emperor.

(5.8) If a messenger sees the emperor to whom he has a message to deliver, then he has to bow.

(5.9) the messenger turns up in the room where the emperor is → he sees the emperor → he bows

(5.10) Un messager arrive dans la pièce où se trouve l'empereur. Il s'incline devant lui.
A messenger comes in the room where the emperor is. He bows in front of him.

Similarly, in (5.11), the relation can be classified as narration or result, depending on whether a causal law such as (5.12) is activated or not.

(5.11) Mais voilà les oies de la sorcière Baba-Yaga qui passent par là. Elles emportent le petit gars.
But the witch Baba-Yaga's geese pass by. They take the little guy.

(5.12) When they see a lonely child, Baba-yaga's geese take him.

Because the temporal order of the eventualities are clear, the same problem does not exist for reason relations. And the eventualities are clearly temporally ordered in a text, because otherwise, it would hinder its understanding. This means that either the temporal relation is made clear in the

text (be it with the use of verb tenses or because the eventualities are part of a well known prototypical script), and then the reverse ordering yields a focus on their causal relation, or the eventualities are not clearly temporally ordered, but are easily identified by world knowledge as standing in a causal relation, and then this identification of the eventualities as causal makes their temporal ordering clear.

In the case of reason relations, because the eventualities are presented in reverse temporal order, they cannot be in a narration relation. An annotator then has to decide between several non temporal relation that present specific characteristics such as elaboration, background or reason. Example (5.13) and (5.14) show reverse order versions of the previous examples. A change of verb tenses is necessary to preserve the temporal ordering of the eventualities, and they still do not sound extremely natural. In the reverse version, however, the examples are much more clearly causal. In context, elaboration is relatively easy to rule out, as the elaborated eventuality must completely temporally surround the elaborating eventuality.

(5.13) Un messager s'incline très bas devant l'empereur. Il est arrivé en courrant dans la pièce.
A messenger bows low in front of the emperor. He came running into the room.

(5.14) Les oies de la sorcière Baba-Yaga emportent le petit gars. Elle passaient par là.
The witch Baba-Yaga's geese take the little guy. They were passing by.

Narration is negatively defined as a temporal relation with a *lack* of causation, or even a lack of emphasis on causation, whereas other relations all present particular characteristics. It is easier to differentiate between several defined characteristics, as it is the case with reason versus other relations with similar temporal order, than to differentiate between causation and the lack of causation, as it is the case with result versus narration classification.

Because they can be more reliably identified, we will focus the rest of this study on reason relations.

5.4 Using causal rules to recognise implicit causation

In this section, we present results for the manual annotation that we carried out in order to evaluate and categorise necessary world knowledge for causation recognition. We wanted to find the exact nature of the world knowledge that an automatic system would need to have access to for causation recognition. In order to study this knowledge, we manually analysed the 20 occurrences of causation from our corpus: the 17 completely implicit reason occurrences and 3 highly ambiguously marked occurrences appearing with a colon (":") or a gerund (given in annexe C).

In this chapter, we take a purely rule based approach for our manual analysis. It does not mean that this approach cannot be approximated by a simpler statistical model, but our intuitions are much clearer for manual analysis when thinking in a rule-based way.

As we have seen in chapter 2, other factors than world knowledge are involved in the recognition of causation. However, world knowledge is the single most important factor and is often enough to determine that a causation relation takes place. In SDRT, for example, if world knowledge is in favour of causation, causation normally holds. It has to be specifically overpassed by another rule for it not to hold. In this chapter, we focus on world knowledge, and on the way it can be used to recognise the presence of causation.

In this section, we give a sketchy intuitive algorithm for recognising causation based on world knowledge. This algorithm relies on a world knowledge database in the form of causal rules that normally hold between several eventualities. These rules would need to be acquired, either from a hand made database, or automatically.

Let s_1 and s_2 be the pair of clauses that we want to classify into *causal* or *not causal*. Then we can give the following pseudo code:

for all rules **do**
  if s_1 matches current-rule.cause **and** s_2 matches current-rule.consequence **then**
    causal = TRUE
  end if
end for

The idea is simple: compare the eventualities in the text to analyse to the eventualities linked by causal rules in the world knowledge base, and mark the pair of clauses as causal if a matching rule can be found. Hence, for each rule in the database, one needs to match one clause to the causing eventuality and the other to the consequence eventuality. If this can be done, then the clauses stand in a causal relation, if nothing else (a specific context, for example) prevents them from doing so.

This general algorithm entails several difficult steps. First, it is not clear whether it is at all possible to automatically acquire the causal rules in the first place, and as we will see later, the necessary granularity of the rules makes them extremely expensive to hand code. Second, the algorithm itself leads to a major difficulty: matching the clauses to the eventualities that the causal rules rely.

5.5 Dimensions of causal rules: causal chains and eventuality hierarchies

In this section, we describe the precise nature of the causal rules that we found necessary to interpret the causal occurrences from our corpus. This analysis can be done in two different dimensions: the horizontal span of causal rules on a causal chain, and the vertical position of the two eventualities that they rely on an eventuality *is-a* hierarchy. We will analyse the rules that we found necessary on these two dimensions: in section 5.5.1, we will investigate the necessary length of the rules on the causal chain, in section 5.5.2, we will give an analysis of causal rules in respect to the eventuality hierarchy.

5.5.1 Causal rules and causal chains: how far does a causal rule go?

Eventualities in a causal relation are linked by causal chains (see for example Moeschler, in press; Moeschler, 2003a). They are chains of elementary causes and consequences links. These elementary direct causations cannot be decomposed further into a causal chain. For example, the archetypal causation example *John fell, Max pushed him* can be associated with the causal chain in (5.15).

(5.15) Max pushes John → John is unsteady → John falls

We analysed each occurrences of our corpus using a causal chain. For example, we analysed (5.16) into (5.17).

(5.16) Il s'ennuyait très souvent. Dans ce pays-là, l'empereur n'avait pas le droit d'inviter des amis pour s'amuser.
He was very often bored. In that country, the emperor didn't have the right to invite friends over to have fun.

(5.17) The emperor doesn't have the right to invite friends over → he doesn't invite friend → there is no friend to play with → he is bored.

(5.18) He fell asleep very often. In that country, the emperor didn't have the right to invite friends over to have fun.

(5.19) The emperor doesn't have the right to invite friends over → he doesn't invite friend → there is no friend to play with → he is bored → he falls asleep.

In our corpus, the causal chains are made of between 1 and 3 causal links. Moeschler (in press) has shown that longer chains tend to give difficulties in interpreting them. We see the same here. While (5.16) is easily interpretable, adding one step (5.19) in the causal chain, as in (5.18), makes the resulting text difficult to understand.

Some of the direct causation occurrences with causal chain of length 1 are state-state causal occurrences. That is, occurrences where both the cause and the consequence have the aspectual type of a state. In these cases, the causal chain often has a length of one, as the occurrence is governed by a *when* causal law that directly relies the cause and the consequence which are in a complete temporal overlap, such as in example (5.20) that corresponds to the causal chain in (5.21) and to the causal rule in (5.22). We use the sign ⇒ to mean *normally causes*.

(5.20) Le tailleur de pierre était content, il aimait sa vie.
The stone-cutter was happy. He liked his life.

(5.21) The stone-cutter likes his life → he is happy.

(5.22) x likes his life $\Rightarrow$ x is happy.

Similarly, example (5.23) shows an occurrence from our corpus associated with the size 3 causal chain in (5.24). This example might be explained by the law in (5.25).

(5.23) [Les chats] ronronnent doucement en la voyant si travailleuse.
[The cats] purr gently, watching her working so hard.
Context: the maid works for the cats.

(5.24) The maid is working hard $\rightarrow$ the cats see her $\rightarrow$ they are happy $\rightarrow$ they purr

(5.25) x is working hard for y $\Rightarrow$ y shows a sign of happiness.

We wanted to determine the optimal length of causal rules. Let us define a *direct causal rule* as a rule that corresponds to a size one causal chain such as the rule in (5.22) that corresponds to the chain in (5.21). In a world knowledge database, it might be tempting to use direct causal rules only, as they could then be recombined into longer causal rules.

However, previous linguistic theories give world knowledge rules of arbitrary sizes on the causal chain. As we have seen earlier, the famous *John fell, Max pushed him* is size 2, not one.

Depending on the way one would acquire causal rules, there is no theoretical impossibility of using direct causal rules. However, we will argue that longer rules are more optimal, and this for a major reason: direct causal rules are often not natural, because they are often made of entailment relations that are not defeasible and are not consciously accessible (for more on entailment and causation, see Moeschler, in press). Thus, they do not appear in discourse. Direct causal rules such as in example (5.26) sounds awkward, and causal occurrences that can be explained using these rules only would be much rarer, if even present, in a corpus than the higher level rules such as the one in example (5.27). Thus, using direct causal rules would worsen the scarcity problem linked to acquiring causal rules.

(5.26) x pushes y $\Rightarrow$ y is unsteady.

(5.27) x pushes y $\Rightarrow$ y falls.

It is more straightforward to use one causal rule per causal occurrence in the text, which often seems optimal. However, for some occurrences in our analysis, we preferred to use two causal rules because one or both looked more intuitive and generic than a global longer chain spanning rule. For example, (5.28) could be associated with (5.29), but we preferred to associate it with (5.30) and (5.31). In this case, both rules are much more generic than the one step long causal chain rule, and one can imagine a number of situations where one or the other would be useful. Using the one step rule would lead to the multiplication of the creation of similar rules for different situations. Moreover, in this case, both rules are intuitive and might be useful for naturally occurring one step causation relations such as the ones in (5.32) and (5.33).

(5.28) Il s'ennuyait très souvent. Dans ce pays-là, l'empereur n'avait pas le droit d'inviter des amis pour s'amuser.
He was very often bored. In that country, the emperor didn't have the right to invite friends over to have fun.

(5.29) x doesn't have the right to do something entertaining $\Rightarrow$ x gets bored.

(5.30) x doesn't have the right to do y $\Rightarrow$ x doesn't do y.

(5.31) x doesn't do something entertaining $\Rightarrow$ x gets bored.

(5.32) She didn't hitch-hike because her parents didn't allow her to.

(5.33) Yesterday I couldn't play with my chemistry box, so I got bored.

Choosing to use one or several rules is arbitrary, and the perception of how much a rule sounds natural is subjective, so little can be said without a complete corpus analysis. However, a clear trade-off exists between using rules with a small causal chain span that are likely to be more generic, and using longer rules that might appear more natural and thus have a higher chance of directly linking two eventualities in a corpus. Moreover, choosing to store long rules will lead to several problems. First, this will produce a multiplication of the necessary rules and thus to a sparseness problem. Second, long rules are not combinable to form new rules. On the other hand, choosing to store short rules will lead to the issue that such rules, not being natural, are generally not found in natural language texts.

5.5.2 Causal rules and eventuality hierarchies: how specific is a causal rule?

The causal rules that we have seen in this chapter usually do not link two eventualities that are directly expressed in the analysed clauses, but two more generic types of eventualities. We proposed, for example,(5.34) instead of (5.35) that would stick to the expressed eventualities.

(5.34) x doesn't have the right to do something entertaining $\Rightarrow$ x gets bored.

(5.35) the emperor doesn't have the right to invite friends over $\Rightarrow$ he gets bored.

This categorisation of eventualities into increasingly more generic eventuality types can be represented as a conceptual hierarchy or ontology of eventualities. This hierarchy is a directed acyclic graph where each eventuality is linked to a more generic eventuality with an *is-a* relation. This hierarchy is not a tree, as an eventuality can have several different and more generic eventuality ancestors. For example (5.36) can be made more generic in (5.37) or in (5.38), which are not linked to each other with an *is-a* relation.

(5.36) The cat purrs.

(5.37) An animal purrs

(5.38) The cat makes a noise.

Even if we separately match the verb and its arguments with distinct hierarchies, some different characteristics of eventualities can be used as the basis to make several different super eventualities that might all have a causal power. For example, (5.39) can be of type (5.40) or (5.41).

(5.39) Men wash their beards in wine.

(5.40) x cleans y.

(5.41) x does something surprising.

Causal rules can link eventualities that are anywhere in the hierarchy, from the direct mapping of the clauses to eventualities and up in the hierarchy linking increasingly more generic eventualities until the rule reaches a point of optimal genericity that corresponds to an intuitive causal rule. At this point, the causal rule is both intuitive and generic, but far from the clauses and, as we will see in (5.6.4), the automation of classifying clauses into this type of eventualities is extremely challenging. Examples (5.42) and (5.46) can be associated with the close fitting rules (5.43) and (5.47), or with the more generic rules (5.44) and (5.48), and one reaches maximal genericity with rules such as (5.45) and (5.49).

(5.42) Mais, bientôt, on ne trouva plus de barbier. Ils étaient tous en prison !
But soon it became impossible to find any more barbers. They were all in jail!

(5.43) all barbers are in jail ⇒ it is impossible to find anymore of them.

(5.44) x is in jail ⇒ it is impossible to find x.

(5.45) x is hidden ⇒ one cannot find x.

(5.46) À ce moment, on entend un air de trompettes. La mère de l'empereur vient d'arriver dans le palais !
At this time, a trumpet tune is heard. The emperor's mother just arrived at the palace.

(5.47) the emperor's mother arrives at the palace ⇒ a trumpet tune is heard.

(5.48) a member of the royal family arrives at the palace ⇒ a tune is heard.

(5.49) an important person arrives at an official place ⇒ they are announced.

In these examples, the most intuitive and generic rule is obtained by going up in the hierarchy until an optimal point is reached.

Specific causal types of super-eventualities. The causal power in eventualities often arises from a specific characteristic of the cause eventuality that can be modelled as a super-eventuality in the hierarchy or as a specific eventuality type. Take the preceding (5.42) occurrence, for example. Here the causing eventuality *to be in jail* is a type of *hiding event*. This can be modelled as a specific characteristic of the eventuality *x is in jail* which is then associated with the characteristic *hiding* or as a super eventuality in the directed acyclic graph hierarchy, where *to be in jail* is linked with an *is-a* relation to a super eventuality *hiding event*. In this case, the super-eventuality is clearly defined by its causal power only and not by any other conceptual characteristic. *Hiding* eventualities, for example, are defined by their causal power: they cause the fact that their patient cannot be seen or found.

In our corpus, this was especially true for emotion causing eventualities. They are types of eventualities that cause a specific emotion. Example (5.50) and (5.51) show a *surprising event* as the cause of a surprise sign. Similarly, example (5.52) shows a sad eventuality as the cause of a sadness sign.

(5.50) Ils poussent un cri d'étonnement. Ce bébé a un oeil noir et un oeil tout gris.
They scream with surprise. This baby has a black eye and a completely grey eye.

(5.51) Arrivé au pays de l'Abracabizcorne, le prince n'en croit pas ses yeux : les hommes lavent leur barbe dans du vin, les escaliers montent vers rien et les chats chantent tralalalila.
When he arrived in the land of Abracabizcorne, the prince doesn't believe his eyes: the men wash their beards in wine, the stairs climb up to nothing and the cats sing tralalalila.

(5.52) La maman, assise près de la cheminée, reste là à pleurer. Elle est si fatiguée qu'elle ne peut plus travailler, et elle n'a plus du tout d'argent pour acheter à manger.
The mother, seated by the fireplace, stays here crying. She is so tired that she can't work any more, and she doesn't have any money left at all to buy something to eat.

In these examples, the causing eventualities are intrinsically associated with an emotion causing characteristic. This type of causation creates difficulties in automation, as we will see in section 5.6.

Simplicity-genericity trade-off. Overall, there is a trade-off between using generic rules and using rules relying eventualities that are easy to link to their representation in the actual text clauses. Take (5.53), for example. A generic rule that could explain its causality would be (5.54). However, it brings the problem of recognising *his life* as a *present situation*, if one uses a compositional model, or even to match *liking one's life* to *liking a present situation* otherwise.

(5.53) Le tailleur de pierre était content, il aimait sa vie.
The stone-cutter was happy. He liked his life.

(5.54) liking a present situation ⇒ being happy.

More generally, using generic rules calls for a more complex overall system. When using generic rules, one needs specific ways to treat coordination and negation, for example, that could be captured by specific rules on coordinated or negated eventualities. Moreover, treating coordination or negation on causal rules would call for a specific causation logic, dealing for example with multiple necessary causes.

5.6 Predictable difficulties for automation

This approach leads to several foreseeable difficulties in automation. The current state of the art in natural language processing is far from sufficient to achieve usable results on causal world knowledge acquisition and on its usage for causation recognition. In this section, we list some difficulties that are major obstacles for the automatic recognition of causal relations. These difficulties are linked to the different tasks involved in using world knowledge for causation recognition. We will analyse foreseeable difficulties in representing and acquiring world knowledge in the form of causal rules, and in matching new instances to existing rules.

5.6.1 Representation of causal rules

Representation of causal rules is fairly straightforward, only requiring an eventuality graph and some argument constraints. Causal rules link two eventualities or eventuality types that usually appear together in a causal relation. World knowledge can be modelled as a graph of causal rules where eventualities or eventuality types are linked by a directed causal rule. With simple, very general eventualities, such as *if it rains, the floor normally gets wet*, it is possible to stop here. However, as soon as the rule links eventualities that are a bit more generic, complex constraints are necessary to decide if the eventualities can be linked together in the specific instance. In particular, constraints on arguments are necessary. For example, in (5.55) a generic causal rule explaining it would be (5.56) which has some complex argument constraints.

(5.55) Le soir même, la mère envoie Ludivine chez la sorcière. Elle se dit qu'avec un peu de magie, on ne sait jamais, après tout !
The same night, the mother sends Ludivine to the witch. She thinks that maybe with some magic, one never knows after all !
Context: The mother wants one of her daughters to marry the king. The witch can make them prettier. She already sent two daughters to the prince but he didn't want to marry them. The mother thinks that Ludivine is less pretty than her sisters.

(5.56) x things that y might be useful $\Rightarrow$ x does y.

Although a somewhat complex constraint system needs to be implemented, rules representation does not seem to us to be a particular issue for automation.

5.6.2 Acquiring causal rules

Acquiring causal rules is one of the main difficulties for a world knowledge based approach. At the clauses level of granularity, eventualities are extremely sparse. As we have seen in chapter 4, verb pairs are already very sparse in causation occurrences. When adding arguments, the sparseness will increase, and it is inconceivable to calculate any statistics on such instances, even on extremely large corpora, and even if this was possible for

a few eventualities, the new eventualities in a test corpus will probably not ever have been seen.

Intuitive causal rules, such as the ones that we used to manually analyse our corpus, however, might have some genericity. Rules such as *when someone does not have the right to do something, they normally do not do it* or *a surprising eventuality normally causes surprise* are quite generic and would reduce the sparseness problem to a point where a big corpus would probably have several occurrences of instances of each of these rules. However, even in the unlikely case where someone would have access to an appropriate eventuality hierarchy, transforming eventualities in an instance into the their correct vertical (on the eventuality hierarchy) ancestors and then assigning to them the optimal horizontal (on the causal chain) rule would still be an open problem.

Because of the extreme sparseness of eventualities, a complete eventuality hierarchy is inconceivable. The top part of this hierarchy might be handmade, but would lead to the open problem of automatic eventuality categorisation.

For these reasons, automatic acquisition of causal knowledge is very far from the current natural language processing state of the art. Manual coding of these rules, for example by hand coding rules for causation occurrences on a corpus –as we did for this study, but on a much larger corpus– might be feasible in the current state of the art, but the costs of making such a resource would be prohibitive for many computational linguistic research groups.

5.6.3 Exploiting causal rules

Providing that one had access to causation rules, many difficulties still remain in using this knowledge for the automatic recognition of causation. We will study these issues in the next sections.

Globally, given a potential causation occurrence made of two clauses, two major steps must be taken to recognise causation. First, one needs to identify the eventualities that are textually represented as occurrences of abstract eventualities that are stored in the world knowledge database. We will study this step in section 5.6.4. Second, once the two clauses have each been matched to their corresponding abstract eventualities, one needs to verify in the world knowledge database whether these two abstract eventualities nor-

mally occur together in a causation relation. One needs to find out whether one of the clauses is the representation of an eventuality that normally causes the eventuality represented by the other clause. We will discuss this part in more detail in section 5.6.5.

5.6.4 Matching clauses to eventuality types

As we have seen in the previous section, a complete hierarchy of eventualities is inconceivable, given the number and sparseness of the finest granularity, clauses-fitting eventualities. Even the upper part of this hierarchy creates several problems. As we have seen, some eventuality types are linked to their specific causal properties. This is especially visible in emotion-causing eventualities. Example (5.50) linked a surprise causing eventuality to a sign of surprise. Using *surprising eventualities* as super-eventualities requires that one lists all the possible types of surprising eventualities, which is a non-sense. It is theoretically impossible to list all surprising eventualities, and listing all unsurprising eventualities is also completely unrealistic. Humans and even young children are able to recognise such eventualities, but, as they cannot make use of a such a list, they probably use another non linguistic system, such as theory of mind, to compute what emotion would be expected from the eventuality that they process. Such systems are beyond the scope of computational linguistics and, as far as we know, far from the current state of the art of artificial intelligence.

If those problems were overcome, and a hierarchy of eventualities existed, two issues would still remain: categorising new eventualities into the hierarchy, and choosing an appropriate level of granularity. Categorising new eventualities into their type might be partly feasible for simple eventuality types such as categorising *The emperor's mother just arrived at the palace* into *an important person arrives at an official place*, which could be done by *recognising textual entailment* systems[3] or systems based on distributional semantics. However, recognising a surprising event, for example, is still an open problem.

[3] the task of recognizing an utterance to entail another has been the goal of many recent computational linguistics systems and is the focus of a specific workshop (`http://www.nist.gov/tac/2010/RTE/index.html`)

Finally, the system that matches clauses to eventualities would need to categorise the clauses into the right granularity of eventualities, as different causal rules apply to an eventuality or its ancestor. The system would have to communicate with the *matching eventualities to causal rules* system and both should find an optimal match, taking into account the level of granularity of both eventualities and the number of causal rules necessary to link them, which is the horizontal causal chain dimension and the vertical eventuality granularity one.

5.6.5 Matching eventualities to causal rules

Once the clauses are matched to eventualities or to lists of ancestor eventualities, an implicit causation recognition system would need to discover whether a causal rule does apply to them. Given causal rules represented as a graph with constraints on the directed causation links, the system would have to find the optimal rule or rules that link the eventualities. This step and the previous step should work together to find the best causal relation; that is choosing the correct ancestors for each eventuality and linking them using one or several causal rules. Following the *penguin principle* [4] (Lascarides and Asher, 1993), specific eventualities should be preferred, and the path from an eventuality to the other should be short. These parameters will interact with each other, so they must be optimised together. One could then return a score to a statistical system that would take this score as well as other clues into account, or a threshold can be implemented in order to return a binary decision.

If all other steps were possible, this one should be far from insurmountable.

[4] This problem of contradicting rules is known as the penguin principle. It states that in the contradicting case of rule one and rule two, the most specific should apply.

Rule 1 all birds fly.

Rule 2 penguins are birds that do not fly

The segmented discourse representation theory (SDRT), proposes a logic that can deal with such contradictions, however its complexity makes it extremely difficult to implement in an automatic system

5.6.6 Other necessary information

We encountered several specific difficulties with this algorithm on causal occurrences in our corpus. As we have seen in section 5.5.2, the system needs a way to treat complex eventualities made of sub-eventualities with a coordination. It can either use a complex logic, which would also allow for the treatment of negation, or a complex eventuality can be coded as one very fine grained eventuality in the hierarchy, which would make the construction of the eventuality hierarchy even more difficult. One could also use a more brutal approach by taking into account only one of the coordinated eventualities, the one that achieves the best causal score, for example.

Other difficulties arise from instances where world knowledge alone is not enough to recognise causation. They can come from necessary contextual knowledge or from specific ways of expressing causation. For necessary contextual knowledge, look at (5.57) or (5.59). Without context, (5.57) is difficult to understand. First, one needs to resolve *he* as referring to *the king Leon*. Causation is then difficult to understand without the necessary information in (5.58), which can be deduced from context (the princess is the only possible object of love, here, and she has been trying to seduce the king). The text makes even more sense with the information that the king has never understood why kings marry princesses and has never seen a princess before.

(5.57) Petit à petit, il comprend pourquoi les rois épousent les princesses. Le roi Léon est amoureux !
Little by little he understands why kings marry princesses. The king Leon is in love!

(5.58) King Leon is in love with a princess.

(5.59) Justement, la servante s'en va, pleine de griffures et d'égratignures: elle a lancé de l'eau bouillante dans les pattes de trois chats tigrés, et ils se sont vengés.
At that time, the maid goes away, she's covered with scratches and grazes: she threw boiling water at three striped cats and they took their revenge.

(5.60) La servante est partie pleine de griffures et d'égratignures: elle a lancé de l'eau bouillante dans les pattes de trois chats tigrés, et ils se sont vengés.
The maid went away covered with scratches and grazes: she threw boiling water at three striped cats and they took their revenge.

(5.61) La servante est partie pleine de griffures et d'égratignures: elle a lancé de l'eau bouillante dans les pattes de trois chats tigrés.
The maid went away covered with scratches and grazes: she threw boiling water at three striped cats.

(5.62) x does something bad to $y \Rightarrow y$ takes revenge on x.

Example (5.59) is complex. The explicit inverse temporal ordering of the clauses is a strong causation clue, but the example could be transformed into (5.60) without substantially changing its meaning. In this case, the causation is much more difficult to understand without the last clause, such as in (5.61), where it is difficult to choose between a reason or a result relation. In the original text, the last and next to last clauses are linked by a result relation; that result relation is much more obvious than the reason relation in (5.61), as the eventuality *they took there revenge* is clearly linked to the general causal rule expressed in (5.62). The maid covered with scratches is an consequence of the revenge of the cats. Therefore, here we have an unclear causation made understandable by the later expression of the missing link in the causal chain. This analysis is extremely complex for a natural language processing system.

5.7 Conclusion

In this chapter, we described a manual analysis of causation occurrences in a children's corpus. We annotated with causation relations 33 tales intended for young children. We analysed 20 of them and we gave a sketchy algorithm for using world knowledge for causation recognition, and gave, based on the occurrences from our corpus, a list of tasks that would be needed but are not achievable given the current state of the art of computational linguistics.

We argued that intuition is clearer for relations that present the eventualities in the text in ante-chronological order, as it is not always possible

to clearly disambiguate *result* relations from the narration relations that do not carry any causal meaning. We further argue that a continuity exists between result relations and purely narration ones, depending on the relative focus given to the causal part of the relation in the text. For this reason, we analysed only the 20 implicit or highly ambiguously marked reason relations from our corpus in order to understand the nature of the necessary world knowledge for causation recognition.

We showed that extremely difficult problems arise for the automation of world knowledge extraction and of its usage for causation recognition. Difficulties lie in the massive sparseness of eventualities in a causal relation, rendering the automatic acquisition of causal laws not foreseeable in the near future with the current state of the art of natural language processing. Moreover, for similar reasons, the recognition of the correct type of eventualities in natural text clauses is far from being solved.

Finally, we showed that more difficulties lie ahead for occurrences where causation is not expressed in a simple consequence-cause way, but as different scattered parts of a causal chain, and for occurrences where contextual information is necessary for the understanding of causation.

A major drawback of this type of analysis on a small sample is that it does not allow us to estimate the number of false positives that such a method might include. Take, for example, the rule in (5.63) (given below): it is useful for explaining the causation in (5.64), but it is a very general rule and it might lead to false positives. It is, however, very difficult to estimate their number and nature with a manual analysis method.

(5.63) Strong desire → emotion sign.

(5.64) En pensant à sa mère et à Lola, elle a tout de même le coeur serré. Elle aimerait bien les revoir
Thinking of her mother and of Lola, she has nevertheless a broken heart. She would like to see them again.

Although it is conceivable that a statistical analysis could allow a less complex system to retrieve most of the causation without adding a lot of noise, any kind of system would still require the use of world knowledge. Every linguistic theory that explain the recognition of causation by humans centres its analysis on world knowledge. Following these theories, automatic

recognition systems make use of a simple world knowledge modelling, generally based on pairs of words. However these systems are much too simple. As we have shown in this chapter, causation is extremely sparse. Even if this sparseness can be reduced by aggregating eventualities into less finely grained super eventualities, some of the sparseness will remain and it is absolutely not clear whether we even have the technology to automatically cluster eventualities into these super eventualities. We do not have any annotated corpus of such eventualities classified into their super-eventualities. An unsupervised approach, maybe based on topic models, might be possible, but even this is very far removed from the actual state of the art.

We showed that the field suffers from an extreme sparseness problem, as the eventualities that are causally connected in world knowledge are represented in texts as complete clauses, which are even rarer in corpora than are the very rare verb pairs. We lack a way of abstracting these eventualities in order to overcome this sparseness that is present in the data and that prevents us from achieving automatic causation recognition.

Overall, the world knowledge based approach leads to difficulties that are extremely far from being solvable with the current state of the art of natural language processing, and if we believe previous linguistic work, world knowledge is an absolute necessity for any implicit causation recognition system. In this chapter, we have shown that the actual world knowledge needed is much more complicated than simply pairs of words. We hypothesise that humans get this knowledge and use it first in their day-to-day lives, and then use their experience and their imagination to understand causation in a text. That is, causation is first a psychological feature and not a linguistics problem.

Chapter 6

Conclusion

The aim of this thesis was to study the automatic extraction of implicit causal relations. A system that would carry out such a task would mimic human behaviour, as people can understand causal relations in natural language texts even without any causal marker. We showed, however, that this task is not well defined and is a very subjective one, leading to low inter-annotator agreements.

Theoretical work that aims at explaining how humans can understand implicit causation are complex and cannot be directly implemented as computer programmes. However, they all share the idea that the main way of recognising implicit causation is to access already existing world knowledge. This world knowledge states that some eventualities normally cause other eventualities in the world. Previous computational linguistics work does not aim at implementing specific linguistic theories, but the vast majority make use of some feature representing approximation of world knowledge. In this work, authors typically use pairs of verbs or, more generally, pairs of words, as clues as to which eventuality the text represents. These features are extremely sparse, and although they produce results that are clearly better than a random baseline, the results are not accurate enough to be useful in practice.

Verbs are a straightforward way of representing eventualities, but we showed that specific pairs of verbs are not predictive of causation. This shows, following previous linguistic work, that they are not good enough at representing eventualities that can pertain in causal relations.

Looking more thoroughly into necessary world knowledge for individual cases of implicit causation, we showed that the eventualities are indeed much more complex and finely grained than their central verbs. We showed that necessary world knowledge is complex. This complexity is such that it is not foreseeable, given the current state of the art, to cluster individual representations of eventualities. This clustering, however, would be the only way to sufficiently reduce the sparseness problem for world knowledge to be learnt. It would also be necessary to match individual textual representations to abstract eventualities. This, in turn, would be necessary to allow us to recognise implicit causation using a world knowledge database of causal links between abstract eventualities.

6.1 Main contributions

In this section, we will outline the main contributions of this thesis in regard to each of our research questions.

Question 1: What are the defining characteristics of causation ? Which characteristics of causation are associated to it in human judgement?

We showed that several characteristics of causation from previous theoretical work are indeed statistically associated with the recognition of causation in human judgement, and we listed them. We showed that causation is a highly subjective task and we produced a coherent characteristics-based annotation manual that also serves as a working definition of causation. This manual does not lead to high pairwise inter-annotator agreement but it rather leads to a high agreement between a voting scheme on all annotators answers and our own educated judgement.

Question 2: How can we take the subjectivity of a task into account when evaluating systems that carry it out automatically?

We proposed to compare automatic systems not to an arbitrary golden standard that gives only one set of answers to a subjective task, but to the actual human annotators. In doing so, we proposed to evaluate the system with the same inter-annotator agreement measure that served to evaluate

the human judges, and to use a statistical test to compare this measure to what human annotators usually achieve.

Question 3: Are there statistical correlations between individual verb pairs and causation?

We showed, on large corpora, that the verb pairs that are the most predictive of causal relations do not intuitively represent eventualities that actually tend to enter causal relations in the world. Moreover, we showed that, as the reliability of the predictability measure goes up, the measure goes toward showing statistical independence between individual verb pairs and the presence of a causal relation. This is a strong clue that verb pairs cannot be good predictors of causation.

Question 4: What is the exact nature of the world knowledge necessary to recognise implicit causation ? How can it be represented and acquired?

We manually analysed a corpus of implicit causation, identifying in each case the world knowledge necessary to recognise causation. We showed that the eventualities that are normally causally linked need to be represented in a much more complex and accurate way than just verb pairs. We showed that, to use a world knowledge resource, one would need to abstract the textual representation of eventualities into eventuality classes, which the current state of the art in natural language processing is not ready to provide. We also showed that the same problem arises for the automatic acquisition of such a database. Moreover, the intrinsic scarcity of individual eventualities renders the manual writing of such a database unforeseeable.

6.2 Perspectives

Overall, the field of implicit causation recognition or, more generally, of discourse relation classification, is not only suffering from a word or verb pairs sparseness problem. If it were the case, previous work that aims at using much larger automatically-acquired training corpora, such as the work of (Marcu and Echihabi, 2001) or of (Sporleder and Lascarides, 2007), would have been much more successful. The fact that individual verb pairs do not predict causation, even for large corpora, is a strong clue that the problem is

much deeper. We have argued, in this thesis, that the field currently suffers from another problem: the lack of world knowledge resources that would be usable in automatic systems. This lack is itself due to two fundamental problems: first, the representation of abstract eventualities that are normally causally linked. A fundamental problem, in this case, is the role of the verb arguments for its causal power. Drinking water, for example, does not cause someone to get drunk, whereas drinking whisky does. In this case, the argument is necessary to establish causation. As drinking always causes someone to feel less thirsty, regardless of the arguments. Second, the extreme difficulty in linking individual textual representations of eventualities to the abstract eventualities that they are specimen of. Unless these problems are solved, it will be impossible to automatically acquire such resources (from explicit occurrences, for example) or to use such a resource if it existed.

These problems apply to the natural language processing field in general. If we were able to solve them, we could use a system that abstracts textual representations into ideal eventualities for many tasks, as it is a core part of what we call understanding. This also asks the question of how humans do acquire world knowledge in general. Is it from experience only? Does language play a role in this acquisition? On a broader scale, we have a problem that is similar to the artificial intelligence and robotics problem of *symbol grounding*. *Symbol grounding* means to link an actual world phenomenon to its internal representation in an artificial intelligence system. We arrive at a point, in computational linguistics, where real phenomena in the world need to be linked to their representations. In our case, classes of eventualities exist in the world that tend to cause other classes of eventualities that tend to cause other classes of eventualities. We need to represent these classes, and then we need to link them to their representations in texts.

How might we recognise abstract eventualities or build eventuality hierarchies? Humans can clearly do it, but the space covered by all possible individual eventualities is much too big for the corresponding resource to be built manually. One approach, on extremely domain specific corpora, is that of Riaz and Girju (2010). They had some success clustering eventualities, when these eventualities were issued from a common topic, shared their central verb and had similar subject and object arguments. Although this approach undoubtedly leads to accuracy and scalability issues, a more complex system, using much more vectorial semantics or topic models in-

formation to cluster eventualities that do not share any lemma might prove interesting.

References

R. Artstein and M. Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

N. Asher and A. Lascarides. 2003. *Logics of conversation*. Cambridge University Press.

R. Barzilay and K. R McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, page 50–57.

B. Beamer and R. Girju. 2009. Using a bigram event model to predict causal potential. In *CICLing*, pages 430–441.

S. Blair-Goldensohn, K. R McKeown, and O. Rambow. 2007. Building and refining Rhetorical-Semantic relation models. In *Proceedings of NAACL-HLT*.

J. Bohnemeyer. 1998. *Time relations in discourse: Evidence from a comparative approach to Yukatek Maya*. Ponsen & Looijen.

A. Burchardt, K. Erk, A. Frank, A. Kowalski, S. Padó, and M. Pinkal. 2006. The SALSA corpus: a german corpus resource for lexical semantics. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006)*.

J. Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics*, 22(2):249–254.

L. Carlson, D. Marcu, and M. Okurowski. 2002. *Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory*.

R. Carston. 1993. Conjunction, explanation and relevance. *Lingua*, 90(1):2.

Robyn Carston. 2002. *Thoughts and utterances: the pragmatics of explicit communication*. Wiley-Blackwell.

K. Church and R. Patil. 1982. Coping with syntactic ambiguity or how to put the block in the box on the table. *Comput. Linguist.*, 8(3-4):139–149.

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Psychological Bulletin*, (20).

D. Garcia. 1998. «Exploitation pour l'élaboration de requêtes de filtrage de texte, des connaissances causales détecté par COATIS». *RIFRA'98 Rencontre internationale sur l'extraction, le filtrage et le résumé automatique*, page 44–54.

R. Girju and D. Moldovan. 2002. Mining answers for causation questions. In *In AAAI symposium on*.

R. Girju. 2003. Automatic detection of causal relations for question answering. In *Proc. of the 41st ACL, Workshop on Multilingual Summarization and Question Answering.*

H. P Grice. 1975. Logic and conversation. In P. Cole and J. L. Morgan, editors, *Syntax and Semantics: Vol. 3: Speech Acts*, pages 41–58. Academic Press, San Diego, CA.

P. Grice. 1989. *Studies in the Way of Words.* Harvard Univ Pr.

Hugo Hernault, Danushka Bollegala, and Mitsuru Ishizuka. 2010. A semi-supervised approach to improve classification of infrequent discourse relations using feature vector extension. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, page 399–409, Stroudsburg, PA, USA. Association for Computational Linguistics. ACM ID: 1870697.

E. H Hovy, M. P Marcus, M. Palmer, L. A Ramshaw, and R. M Weischedel. 2006. OntoNotes: the 90% solution. In Robert C. Moore, Jeff A. Bilmes, Jennifer Chu-Carroll, and Mark Sanderson, editors, *HLT-NAACL.* The Association for Computational Linguistics.

D. Hume. 1739. *A Treatise of Human Nature.* London, Millar.

2005. interdisciplines, `http://www.interdisciplines.org/medias/confs/archives/archive\_6.pdf`.

T. Inui. 2005. Creating an annotated corpus for the analysis of causal relations. *COE-LKR2005.*

C. S. G Khoo, S. Chan, and Y. Niu. 2000. Extracting causal knowledge from a medical database using graphical patterns. In *ACL '00: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, page 336–343, Morristown, NJ, USA. Association for Computational Linguistics.

A. Lascarides and N. Asher. 1993. Temporal interpretation, discourse relations and commonsense entailment. *Linguistics and Philosophy*, 16(5).

B. Levin. 1993. English verb classes and alternations: A preliminary investigation. *Chicago, Il.*

A. Louis, A. Joshi, R. Prasad, and A. Nenkova. 2010. Using entity features to classify implicit discourse relations. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 59–62.

W. C Mann and S. A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

D. Marcu and A. Echihabi. 2001. An unsupervised approach to recognizing discourse relations. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, page 368–375, Morristown, NJ, USA. Association for Computational Linguistics.

D. Marcu, E. Amorrortu, and M. Romera. 1999. Experiments in constructing a corpus of discourse trees. In *Proceedings of the ACL Workshop on Standards and Tools for Discourse Tagging*, page 48–57.

M. P Marcus, B. Santorini, and M. A Marcinkiewicz. 1994. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330.

G. A Miller. 1995. WordNet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

E. Miltsakaki, R. Prasad, A. Joshi, and B. Webber. 2004. The penn discourse treebank. In *Proceedings of the 4th International Conference on Language Resources and Evaluation.*

J. Moeschler, C. Chevallier, T. Castelain, J.B. Van der Henst, and I. Tapiero. 2006. Le raisonnement causal: de la pragmatique du discours à la pragmatique expérimentale. *Nouveaux cahiers de linguistique française*, 27:241–262.

J. Moeschler. 2000a. Le modèle des inférences directionnelles. *Cahiers de Linguistuque Francaise 22*, pages 57–100.

J. Moeschler. 2000b. «l'ordre temporel est-il naturel?». *MOESCHLER J. & BÉGUELIN MJ (éds), Référence temporelle et nominale, Berne, Peter Lang*, pages 71–105.

J. Moeschler. 2003a. Causality, lexicon, and discourse meaning. *Rivista di Linguistica, 15.2*, page 343–369.

J. Moeschler. 2003b. «l'expression de la causalité en français». *Cahiers de linguistique française*, 25:11–42.

J. Moeschler. 2007. Discours causal, chaîne causale et argumentation. *Information temporelle, procédures et ordre discursif*, 18:69–86.

J. MOESCHLER. 2010. Causal, inferential and temporal connectives: Why parce que is the only causal connective in french, ms. *The Role of Affect in Discourse Markers, Rouen, Presses Universitaires de Rouen et du Havre*, pages 125–149.

J. Moeschler. 2011. Causal, inferential and temporal connectives: Why parce que is the only causal connective in french. *Hancil S. (ed.), Marqueurs discursifs et subjectivité*, pages 97–114.

J. Moeschler. in press. *Causalité, chaînes causales et argumentation.*

R. Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2):1–69.

M. Palmer, D. Gildea, and P. Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

C. Pechsiri, A. Kawtrakul, and P. Piriyakul. 2006. Mining causality knowledge from textual data. In *AIA'06: Proceedings of the 24th IASTED international conference on Artificial intelligence and applications*, page 85–90, Anaheim, CA, USA. ACTA Press.

M.P. Péry-Woodley, N. Asher, P. Enjalbert, F. Benamara, M. Bras, C. Fabre, S. Ferrari, L.M. Ho-Dac, A. Le Draoulec, Y. Mathet, et al. 2009. Annodis: une approche outillée de l'annotation de structures discursives.

E. Pitler, A. Louis, and A. Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. In *ACL-IJCNLP '09: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2*, page 683–691, Morristown, NJ, USA. Association for Computational Linguistics.

R. Prasad, E. Miltsakaki, N. Dinesh, A. Lee, A. Joshi, L. Robaldo, and B. Webber. 2007. The penn discourse treebank 2.0 annotation manual.

R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber. 2008. The penn discourse treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, page 2961–2968.

J. Pustejovsky, P. Hanks, R. Sauri, A. See, R. Gaizauskas, A. Setzer, D. Radev, B. Sundheim, D. Day, L. Ferro, et al. 2003. The timebank corpus. In *Corpus Linguistics*, volume 2003, page 40.

J. Pustejovsky, J. Littman, R. Saurí, and M. Verhagen. 2006. *TimeBank 1.2 Documentation, http://www.timeml.org/site \ /timebank/documentation- \ 1.2.html.*

Université de Montréal RALI laboratory. 1997. *corpus de bitextes anglais-français.*

A. Reboul. 2000. La représentation des éventualités dans la théorie des représentations mentales. *Cahiers de Linguistuque Francaise 22*, page 13.

A. Reboul. 2005. Similarities and differences between human and nonhuman causal cognition. *www.interdisciplines.org/causality.*

Dennis Reidsma and Rieks op den Akker. 2008. Exploiting 'subjective' annotations. In *Proceedings of the Workshop on Human Judgements in Computational Linguistics*, HumanJudge '08, page 8–16, Manchester, United Kingdom. Association for Computational Linguistics. ACM ID: 1611631.

M. Riaz and R. Girju. 2010. Another look at causality: Discovering Scenario-Specific contingency relationships with no supervision. In *Semantic Computing (ICSC), 2010 IEEE Fourth International Conference on*, page 361–368.

T.J. Sanders, W.P. Spooren, and L.G. Noordman. 1992. Toward a taxonomy of coherence relations. *Discourse processes*.

T.J.M. Sanders, W.P.M. Spooren, and L.G.M. Noordman. 1993. Coherence relations in a cognitive theory of discourse representation. *Cognitive Linguistics (includes Cognitive Linguistic Bibliography)*, 4(2):93–134.

T. Sanders. 2005. Coherence, causality and cognitive complexity in discourse. In M. Bras In M. Aurnague and L. Vieu, editors, *Proceedings of SEM-05, First International Symposium on the exploration and modelling of meaning*, pages 105–114.

L. de Saussure. 2000. Les « règles conceptuelles » en question. *Nouveaux Cahiers de la Linguistique Française, Département de linguistique, Université de Genève*, 22:147–164.

R. E Schapire and Y. Singer. 2000. BoosTexter: a boosting-based system for text categorization. *Machine learning*, 39(2):135–168.

H. Schmid. 1994. Probabilistic part-of-speech tagging using decision trees.

W. A Scott. 1955. Reliability of content analysis: the case of nominal scale coding. *Public opinion quarterly*.

S. Siegel and N. J Castellan Jr. 1988. *Nonparametric statistics for the behavioral sciences* . McGraw-Hill Book Company.

B. Snyder and M. Palmer. 2004. The english all-words task. In *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, page 41–43.

D. Sperber and D. Wilson. 1995. *Relevance: Communication and cognition*. Wiley-Blackwell.

C. Sporleder and A. Lascarides. 2007. Exploiting linguistic cues to classify rhetorical relations. In Nicolas Nicolov, Kalina Bontcheva, Galia Angelova, and Ruslan Mitkov, editors, *Recent Advances in Natural Language Processing IV: Selected Papers from RANLP 2005*, volume 292 of *Current Issues in Linguistic Theory*, pages

157–166. John Benjamins, Amsterdam & Philadelphia.

C. Sporleder and A. Lascarides. 2008. Using automatically labelled examples to classify rhetorical relations: An assessment. *Nat. Lang. Eng.*, 14(3):369–416.

C. Sporleder. 2007. Manually vs. automatically labelled data in discourse relation classification. effects of example and feature selection. *LDV-Forum, Journal for Computational Linguistics and Language Technology.*

D. R Swanson and N. R Smalheiser. 1997. An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artif. Intell.*, 91(2):183–203.

L. Talmy. 1988a. Force dynamics in language and cognition. *Cognitive Science*, 12(1):49–100.

L. Talmy. 1988b. Force dynamics in language and cognition. *Cognitive science*, 12(1):49–100.

Pang-Ning Tan, Vipin Kumar, and Jaideep Srivastava. 2002. Selecting the right interestingness measure for association patterns. *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, page 32–41. ACM ID: 775053.

A.G.B. Ter Meulen. 1997. *Representing time in natural language: The dynamic interpretation of tense and aspect.* The MIT Press.

R. Vieira. 2002. How to evaluate systems against human judgment on the presense of disagreement? In *workshop on joint evaluation of computational processing of Portugese at PorTAL 2002.*

J. Véronis. 1998. A study of polysemy judgements and inter-annotator agreement. In *Programme and advanced papers of the Senseval workshop*, page 2–4.

W. T Wang, J. Su, and C. L Tan. 2010. Kernel based discourse relation recognition with temporal ordering information. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, page 710–719.

J. M Wiebe, R. F Bruce, and T. P O'Hara. 1999. Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, page 246–253.

D. Wilson and D. Sperber. 1998. Pragmatics and time. volume 37, page 1–22. John Benjamins Publishing Company.

D. Wilson and D. Sperber. 2004. Relevance theory. In L. Horn and G. Ward, editors, *The Handbook of Pragmatics.*, pages 607–632. Oxford , Blackwell.

Z. M Zhou, Y. Xu, Z. Y Niu, M. Lan, J. Su, and C. L Tan. 2010. Predicting discourse connectives for implicit discourse relation recognition. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, page 1507–1514.

G. K Zipf. 1935. The psychobiology of language.

Appendix A

French annotation manual

Note: on parle ici d'évènements de façon générale pour désigner aussi bien des états que des évènements. Nous nous intéressons ici uniquement à des exemples causaux entre évènements représentés par des propositions.

Quelques caractéristiques de la causalité

Une cause se passe toujours avant une conséquence. Dans certains cas il est difficile de juger de la temporalité, comme dans *Jean a tapé le verre contre la table et le verre s'est cassé* ou *Jacques s'est fatigué en conduisant*, mais la cause ne peut en tous cas jamais se passer après la conséquence.

On vous propose trois tests linguistiques avec *C'est parce que*, *parce que* et *donc*. En général,*c'est parce que* est le test le plus fiable, si le test de *c'est parce que* ne fonctionne pas, alors il n'y a pas de relation causale. Les deux autres tests peuvent aider, mais ils peuvent être plus ambigus. Par exemple, dans *Jean est malade parce que je ne l'ai pas vu de la journée* la deuxième partie n'est pas la cause de la première.

Une autre caractéristique de la causalité est la propriété contrefactuelle: si la cause ne s'était pas passée alors la conséquence ne se serait pas passée non plus. Quand on utilise ce test il faut faire attention à ce qui se serait probablement passé. Par exemple, si on se demande si *Jean s'est cassé la jambe en skiant* est causal, il faut se demander se qui se serait probablement passé si Jean n'était pas allé skier (il ne se serait pas cassé la jambe), même si il aurait pu se casser la jambe d'une autre façon, par exemple en glissant sur une peau de banane.

Cas ambigus et quelques trucs supplémentaires

Tests supplémentaires Il est parfois utile d'imaginer une suite de causes et de conséquences directes entre les événements que l'on teste pour voir si il s'agit d'une causalité indirecte (on considère toujours un cas de causalité indirect comme causal). Par exemple si on teste *Jean est à l'hôpital, Jacques l'a poussé* on peut trouver la suite Jacques pousse Jean -> Jean tombe -> Jean se blesse -> Jacques appelle les secours -> les secours constatent que Jean est blessé -> les secours emmènent Jean à l'hôpital -> Jean est à l'hôpital. On parle alors de *chaîne causale.*

La causalité est asymétrique. Si on a deux évènements, l'un peut causer l'autre, ou il peut ne pas y avoir de relation causale, mais les deux ne peuvent pas se causer l'un l'autre. Si on n'arrive pas à déterminer parmi les deux évènements, lequel est la cause et lequel est la conséquence, on a souvent affaire à un cas non causal

Cas non-causaux typiques La causalité n'est pas le si ... alors logique. Par exemple *si George Bush est le président des États Unis alors l'Allemagne est en Europe* est logiquement correct, mais pas causal. On peut le déterminer par l'impossibilité de construire une chaîne causale, ou par le fait qu'il n'y a pas d'asymétrie. *C'est un triangle, il a trois côtés* pose le même problème, même si le raisonnement contrefactuel s'applique est que la phrase passe le test de *parce que* et de *donc.* Le test de c'est parce que (c'est parce qu'il a trois côté que c'est un triangle) est plus difficile. Mais ici, c'est surtout la symétrie qui devrait faire suspecter un cas non causal.

Cas ambigus typiques Le contexte peut être la cause. Par exemple *Il s'est cassé le bras en faisant du patin à roulette..* Si on essaie de faire une chaîne causale, on devra sortir un sous-évènement (par exemple foncer dans un mur, ou tomber) de l'évènement contexte (faire du patin à roulette). Ce cas est ambigu. Pour cette application nous considérerons les évènements contextes comme causaux, et *Il s'est cassé le bras en faisant du patin à roulette.* devrait être positif pour la causalité.

Cas d'actes de langage et épistémologiques. On considérera les cas du type *dépêche-toi parce qu'on va être en retard* comme non-causaux. En fait il y a une relation causale entre *on va être en retard* et *je te demande de te*

dépêcher, mais on ne résoudra pas ce type d'ellipse ici. De même dans *il va faire jour, il est 7h30* l'heure cause la croyance qu'il va faire jour, pas le fait qu'il va faire jour et on ne résout pas non l'ellipse dans ce cas. Par contre, le cas similaire *George est sorti parce que sa veste n'est pas sur la chaise* est causal, mais *dans l'autre sens.* On ne résout toujours pas l'ellipse *je crois que* mais ici la cause de la croyance est la conséquence du fait que l'on croit être vrai, donc on a un cas causal: Jean est sorti *cause* sa veste n'est pas sur la chaise.

Les évènements négatifs et les états sont souvent difficiles. Par exemple *Marjorie est majeure, elle peut voter* ou *Comme aucune étude n'a été faite auparavant sur ce sujet, ce rapport sera fort pertinent.* Dans ce genre de cas il est souvent utile d'essayer de faire une chaîne causale (p.e. il n'y a pas d'études -> les spécialistes n'ont pas de données -> les spécialistes veulent des données -> ce rapport sera pertinent.) ou de chercher à remonter à une loi générale (p.e. si on est majeur, on peut voter). Dans tous les cas, il faut bien vérifier l'asymétrie (*C'est parce qu'elle est majeur que Marjorie peut voter* marche parfaitement alors que *c'est parce que Marjorie peut voter qu'elle est majeur* est bizarre).

Appendix B

Texts to analyse for the annotation experiments

B.1 Eliciting intuitive characteristics of causation

For the experiment, the extracts were given in random order.

B.1.1 Parce que without connector

1. Contexte: Il s'agit d'un dialogue entre un policier et un homme accusé de meurtre. Le policier offre à l'homme d'être jugé pour meurtre au deuxième degré en échange d'informations accessoires sur le meurtre.

   "[Appelant]: Comment pouvez-vous savoir que [l'accusation pour laquelle je serais jugé si je refuse le marché] est plus grave? [Policier no 1]: Eh bien, il s'agit d'une accusation de [meurtre] au premier degré. Nous t'offrons une accusation de meurtre au deuxième degré. C'est, c'est comme eu [. . .] Je ne peux pas croire que tu hésites, parce que parce que ce serait . . ."

2. Contexte: Il s'agit d'un compte rendu de procès. Le ministère public a montré le compte rendu d'une conversation entre l'accusé et des policiers dans laquelle l'accusé indique où se trouve l'arme du crime.

“Il reste que le ministère public a tenté de produire la déclaration au procès. Elle lui permettait de faire indirectement ce que le juge du procès lui avait interdit de faire directement: produire la preuve que l’appelant savait où était cachée l’arme à feu.”

3. Contexte: extrait des initiatives parlementaires du parlement canadien.

   “Même si nous, de ce côté-ci de la Chambre, ne sommes pas toujours d’accord avec l’ensemble des politiques, même s’il peut arriver que nous désapprouvions certains éléments du budget, certaines dépenses ou certaines compressions, je crois qu’il faut accorder du mérite au gouvernement. Il essaie au moins de déterminer ce qui est bon pour l’ensemble du pays. “

4. Contexte: dans le roman de Jules Verne "de la terre à la lune", un personnage argumente qu’il ne doit pas y avoir d’atmosphère sur la lune, car on a jamais pu voir que les rayons lumineux étaient déviés quand ils passent près de la lune.

   "En effet, répondit Michel Ardan, voilà votre meilleur argument, pour ne pas dire le seul, et un savant serait peut-être embarrassé d’y répondre; moi, je vous dirai seulement que cet argument n’a pas une valeur absolue. Il suppose le diamètre angulaire de la Lune parfaitement déterminé, ce qui n’est pas."

5. Contexte: extrait des initiatives parlementaires du parlement canadien.

   "Les députés de l’Assemblée nationale, à Québec, prêtent déjà allégeance à la Constitution et au peuple québécois, ils sentent le besoin d’affirmer leur loyauté à l’égard du peuple qu’ils représentent."

6. Contexte: un travail de linguistique introduit la notion de "faux amis".

   "Ces mots apparentés sont dits des "faux amis". Leur ressemblance morphologique crée une attente sémantique qui peut induire en erreur. "

B.1.2 With parce que

1. Contexte: compte rendu judiciaire.

   "Ainsi donc, le personnel affilié au CUPE a été traité défavorablement par rapport au personnel affilié au SISS, tout simplement parce qu'il se trouvait que les conventions collectives prenaient fin à des dates différentes."

2. Contexte: extrait des initiatives parlementaires du parlement canadien.

   "Il faudrait [...] former un comité de la Chambre, pas un comité mixte de la Chambre et du Sénat, mais un comité des élus du peuple pour que, ensemble, nous puissions regarder, poste par poste, les dépenses du gouvernement et que nous, les élus du peuple–parce que, finalement, quand on s'en retourne dans nos comtés et quand on parle des politiciens du Canada, c'est de nous ici, dans cette Chambre, qu'on parle, et je pense que nous avons une responsabilité–donc, nous du peuple devrions être capables de regarder poste par poste les budgets et être capables de faire les coupures où elles s'imposent pour aller chercher de l'argent et [le] générer pour créer des emplois."

B.1.3 Donc without connector

1. Contexte: introduction d'une étude sur l'aide aux aînés au canada.

   "Aucune étude semblable n'a été faite auparavant, ni au Canada ni à l'étranger et ce, malgré la place des TCI dans tous les aspects de la vie quotidienne. Cette recherche sera fort pertinente tant pour les aînés d'aujourd'hui et de demain, que pour les industries et les gouvernements."

2. Contexte: Compte rendu de procès. L'appelant à avoué le meurtre de Hughes.

   "L'existence de similarités factuelles entre le meurtre de Worms et celui de Hughes a convaincu les policiers que l'appelant était également

responsable de la mort de Worms survenue plus tôt. Ils ont poursuivi leur interrogatoire de l'appelant."

3. Contexte: Roman de Jules Verne. On a changé la position de certains éléments d'un télescope.

   "Cette combinaison avait l'avantage de supprimer le petit miroir destiné à renvoyer l'image à l'oculaire. Celle-ci ne subissait plus qu'une réflexion au lieu de deux. Il y avait un moins grand nombre de rayons lumineux éteints."

4. Contexte: extrait d'un manuel de Xerox pour un scanner et un logiciel de reconnaissance de caractères.

   "Les extensions de fichiers sont ajoutées automatiquement par le système, vous ne devriez jamais avoir à en ajouter une manuellement."

5. Contexte: extrait des initiatives parlementaires du parlement canadien.

   "Le Canada fait partie du Commonwealth britannique et à ce titre, nous acceptons la reine comme chef d'État. Les députés lui prêtent un serment d'allégeance."

6. Contexte: travail de linguistique. A propos des faux amis absolus.

   "Les [faux amis] absolus, comme ceux qui sont cités dans les exemples précédents, se caractérisent par le fait que leurs significations sont complètement disjointes; ils ne peuvent jamais être utilisés comme des traductions réciproques."

B.1.4 With donc

1. Contexte: extrait des initiatives parlementaires du parlement canadien.

   "Le budget qui nous a été présenté il y a trois semaines est déjà sur la mauvaise voie [...]. Je voudrais donc, aujourd'hui, examiner le budget en tenant compte de trois facteurs."

2. Contexte: rapport de l'ONU.

   "J'ai exprimé ma profonde préoccupation devant l'intention manifestée par la République populaire démocratique de Corée de dénoncer le Traité [sur la non-prolifération], et je suis donc encouragé par le fait que son Gouvernement ait reporté l'adoption d'une décision finale sur cette question et engagé des négociations avec les autres parties intéressées, notamment l'Agence internationale de l'énergie atomique"

B.1.5 Mais removed

1. Contexte: travail académique à propos d'un système de traduction automatique.

   "Il n'existe pas de données précises aisément accessibles [au sujet de l'applicabilité de la traduction automatique mixte], on peut dire à coup sûr que la part actuelle de la [traduction automatique], pure ou mixte, se situe bien en deçà de 1 p. cent du marché global de la traduction."

2. Contexte: travail académique à propos d'un outil d'aide à la traduction.

   "Dans quelle mesure nous parviendrons à [assouplir certaines conditions] sans que nous échappe aucune des incohérences terminologiques réelles présentement détectées par le système, cela reste à voir. Il serait beaucoup plus difficile de concevoir et de rendre opérationnelles des stratégies permettant au système de distinguer entre les cas acceptables et inacceptables de termes cibles qui ont été omis, ou encore remplacés par un autre terme ou une périphrase."

B.1.6 With mais

1. Contexte: Roman de Jules Verne. A propos du voyage vers la lune du héros d'une nouvelle de Poe.

"Ce voyage, comme les tentatives précédentes, était simplement imaginaire, mais ce fut l'oeuvre d'un écrivain populaire en Amérique, d'un génie étrange et contemplatif."

2. Contexte: extrait d'un manuel de Xerox pour un scanner et un logiciel de reconnaissance de caractères.

"La précision a certes un rôle important à jouer dans la reconnaissance de texte mais elle ne représente qu'une partie de l'équation."

B.2 Example of a file that was printed and given to a subject for this experiment

Section/année: Sexe: masculin/féminin Langue maternelle:

Merci d'avoir accepté de répondre à ce questionnaire. Nous allons vous demander d'indiquer, pour des passages de textes, si, à votre avis, une relation causale est exprimée entre certains éléments du texte. Si vous pensez que c'est le cas, identifiez, s'il vous plaît, le passage du texte qui représente la cause et celui qui représente l'effet. Dans tous les cas, justifiez votre réponse.

1) Contexte: travail académique à propos d'un système de traduction automatique.

"Il n'existe pas de données précises aisément accessibles [au sujet de l'applicabilité de la traduction automatique mixte], on peut dire à coup sûr que la part actuelle de la [traduction automatique], pure ou mixte, se situe bien en deçà de 1 p. cent du marché global de la traduction."

Une relation causale est-elle exprimée entre certains éléments de ce passage ? Si vous ne savez pas si une relation causale est exprimée ou non, indiquez, s'il vous plaît, les informations qu'il vous manque pour décider. Barrez, s'il vous plaît, la mention inutile et justifiez votre réponse:

Je pense qu'une/aucune relation causale est exprimée dans ce passage parce que ...

Si vous avez répondu oui, identifiez, s'il vous plaît, précisément le segment de texte qui représente la cause et celui qui représente l'effet:

2) Contexte: extrait d'un manuel de Xerox pour un scanner et un logiciel de reconnaissance de caractères.

"La précision a certes un rôle important à jouer dans la reconnaissance de texte mais elle ne représente qu'une partie de l'équation."

Une relation causale est-elle exprimée entre certains éléments de ce passage ? Si vous ne savez pas si une relation causale est exprimée ou non, indiquez, s'il vous plaît, les informations qu'il vous manque pour décider. Barrez, s'il vous plaît, la mention inutile et justifiez votre réponse:

Je pense qu'une/aucune relation causale est exprimée dans ce passage parce que ...

Si vous avez répondu oui, identifiez, s'il vous plaît, précisément le segment de texte qui représente la cause et celui qui représente l'effet:

3) Contexte: rapport de l'ONU.

"J'ai exprimé ma profonde préoccupation devant l'intention manifestée par la République populaire démocratique de Corée de dénoncer le Traité [sur la non-prolifération], et je suis donc encouragé par le fait que son Gouvernement ait reporté l'adoption d'une décision finale sur cette question et engagé des négociations avec les autres parties intéressées, notamment l'Agence internationale de l'énergie atomique"

Une relation causale est-elle exprimée entre certains éléments de ce passage ? Si vous ne savez pas si une relation causale est exprimée ou non, indiquez, s'il vous plaît, les informations qu'il vous manque pour décider. Barrez, s'il vous plaît, la mention inutile et justifiez votre réponse:

Je pense qu'une/aucune relation causale est exprimée dans ce passage parce que ...

Si vous avez répondu oui, identifiez, s'il vous plaît, précisément le segment de texte qui représente la cause et celui qui représente l'effet:

4) Contexte: extrait des initiatives parlementaires du parlement canadien.

"Les députés de l'Assemblée nationale, à Québec, prêtent déjà allégeance à la Constitution et au peuple québécois, ils sentent le besoin d'affirmer leur loyauté à l'égard du peuple qu'ils représentent."

Une relation causale est-elle exprimée entre certains éléments de ce passage ? Si vous ne savez pas si une relation causale est exprimée ou non, indiquez, s'il vous plaît, les informations qu'il vous manque pour décider. Barrez, s'il vous plaît, la mention inutile et justifiez votre réponse:

Je pense qu'une/aucune relation causale est exprimée dans ce passage parce que ...

Si vous avez répondu oui, identifiez, s'il vous plaît, précisément le segment de texte qui représente la cause et celui qui représente l'effet:

5) Contexte: Il s'agit d'un dialogue entre un policier et un homme accusé de meurtre. Le policier offre à l'homme d'être jugé pour meurtre au deuxième degré en échange d'informations accessoires sur le meurtre.

"[Appelant]: Comment pouvez-vous savoir que [l'accusation pour laquelle je serais jugé si je refuse le marché] est plus grave?

[Policier no 1]: Eh bien, il s'agit d'une accusation de [meurtre] au premier degré. Nous t'offrons une accusation de meurtre au deuxième degré. C'est, c'est comme eu [. . .] Je ne peux pas croire que tu hésites, parce que parce que ce serait . . ."

Une relation causale est-elle exprimée entre certains éléments de ce passage ? Si vous ne savez pas si une relation causale est exprimée ou non, indiquez, s'il vous plaît, les informations qu'il vous manque pour décider. Barrez, s'il vous plaît, la mention inutile et justifiez votre réponse:

Je pense qu'une/aucune relation causale est exprimée dans ce passage parce que ...

Si vous avez répondu oui, identifiez, s'il vous plaît, précisément le segment de texte qui représente la cause et celui qui représente l'effet:

6) Contexte: travail de linguistique. A propos des faux amis absolus.

"Les [faux amis] absolus, comme ceux qui sont cités dans les exemples précédents, se caractérisent par le fait que leurs significations sont complètement disjointes; ils ne peuvent jamais être utilisés comme des traductions réciproques."

Une relation causale est-elle exprimée entre certains éléments de ce passage ? Si vous ne savez pas si une relation causale est exprimée ou non, indiquez, s'il vous plaît, les informations qu'il vous manque pour décider. Barrez, s'il vous plaît, la mention inutile et justifiez votre réponse:

Je pense qu'une/aucune relation causale est exprimée dans ce passage parce que ...

Si vous avez répondu oui, identifiez, s'il vous plaît, précisément le segment de texte qui représente la cause et celui qui représente l'effet:

7) Contexte: Roman de Jules Verne. On a changé la position de certains éléments d'un télescope.

"Cette combinaison avait l'avantage de supprimer le petit miroir destiné à renvoyer l'image à l'oculaire. Celle-ci ne subissait plus qu'une réflexion au lieu de deux. Il y avait un moins grand nombre de rayons lumineux éteints."

Une relation causale est-elle exprimée entre certains éléments de ce passage ? Si vous ne savez pas si une relation causale est exprimée ou non, indiquez, s'il vous plaît, les informations qu'il vous manque pour décider. Barrez, s'il vous plaît, la mention inutile et justifiez votre réponse:

Je pense qu'une/aucune relation causale est exprimée dans ce passage parce que ...

Si vous avez répondu oui, identifiez, s'il vous plaît, précisément le segment de texte qui représente la cause et celui qui représente l'effet:

8) Contexte: Compte rendu de procès. L'appelant à avoué le meurtre de Hughes.

"L'existence de similarités factuelles entre le meurtre de Worms et celui de Hughes a convaincu les policiers que l'appelant était également responsable de la mort de Worms survenue plus tôt. Ils ont poursuivi leur interrogatoire de l'appelant."

Une relation causale est-elle exprimée entre certains éléments de ce passage ? Si vous ne savez pas si une relation causale est exprimée ou non, indiquez, s'il vous plaît, les informations qu'il vous manque pour décider. Barrez, s'il vous plaît, la mention inutile et justifiez votre réponse:

Je pense qu'une/aucune relation causale est exprimée dans ce passage parce que ...

Si vous avez répondu oui, identifiez, s'il vous plaît, précisément le segment de texte qui représente la cause et celui qui représente l'effet:

9) Contexte: dans le roman de Jules Verne "de la terre à la lune", un personnage argumente qu'il ne doit pas y avoir d'atmosphère sur la lune, car on a jamais pu voir que les rayons lumineux étaient déviés quand ils passent près de la lune.

"En effet, répondit Michel Ardan, voilà votre meilleur argument, pour ne pas dire le seul, et un savant serait peut-être embarrassé d'y répondre; moi, je vous dirai seulement que cet argument n'a pas une valeur absolue.

Il suppose le diamètre angulaire de la Lune parfaitement déterminé, ce qui n'est pas."

Une relation causale est-elle exprimée entre certains éléments de ce passage ? Si vous ne savez pas si une relation causale est exprimée ou non, indiquez, s'il vous plaît, les informations qu'il vous manque pour décider. Barrez, s'il vous plaît, la mention inutile et justifiez votre réponse:

Je pense qu'une/aucune relation causale est exprimée dans ce passage parce que ...

Si vous avez répondu oui, identifiez, s'il vous plaît, précisément le segment de texte qui représente la cause et celui qui représente l'effet:

10) Contexte: extrait des initiatives parlementaires du parlement canadien.

"Il faudrait [...] former un comité de la Chambre, pas un comité mixte de la Chambre et du Sénat, mais un comité des élus du peuple pour que, ensemble, nous puissions regarder, poste par poste, les dépenses du gouvernement et que nous, les élus du peuple–parce que, finalement, quand on s'en retourne dans nos comtés et quand on parle des politiciens du Canada, c'est de nous ici, dans cette Chambre, qu'on parle, et je pense que nous avons une responsabilité–donc, nous du peuple devrions être capables de regarder poste par poste les budgets et être capables de faire les coupures où elles s'imposent pour aller chercher de l'argent et [le] générer pour créer des emplois."

Une relation causale est-elle exprimée entre certains éléments de ce passage ? Si vous ne savez pas si une relation causale est exprimée ou non, indiquez, s'il vous plaît, les informations qu'il vous manque pour décider. Barrez, s'il vous plaît, la mention inutile et justifiez votre réponse:

Je pense qu'une/aucune relation causale est exprimée dans ce passage parce que ...

Si vous avez répondu oui, identifiez, s'il vous plaît, précisément le segment de texte qui représente la cause et celui qui représente l'effet:

B.3 First annotation experiment

The potential cause is given in boldfont and the potential effect in italics.

1. **Paul a pris ses médicaments**, *il va guérir.*
2. *Il fait jour maintenant*, **il faisait nuit il y a quelques heures**.
3. Si **George Bush est président des états unis en 2007** alors *l'Allemagne est en Europe.*
4. *J'ai mis le poulet dans le four*, **je suis en train de cuisiner**.
5. **En signant l'accord**, *il a pris de gros risques.*
6. *Il s'est cassé la jambe* **en skiant**.
7. *Il s'est fatigué* **en roulant toute la journée**.
8. **Frédérique est majeure**, *elle peut voter.*
9. **Le gendarme court**, *il est essoufflé.*
10. Après qu'**il a fait un long voyage au canada**, *il s'est marié.*
11. **Il a ouvert son courrier** puis *il est sorti..*
12. *George est sorti*, **sa veste n'est pas sur le porte manteaux**.
13. *Dépêche-toi*! **On va être en retard**.
14. *Le Concorde ne s'arrêtera pas vraiment*, **il ne sortira jamais de l'imaginaire des hommes.**.
15. *C'est un triangle*, **il a trois côtés**.
16. **J'ai mis le champagne au frigo** pour qu' *il soit bien frais pour le boire.*
17. *Il faut accorder du mérite au gouvernement*, **il essaie de déterminer ce qui est bon pour le pays.**.
18. **Aucune étude semblable n'a été faite auparavant**. *Cette recherche sera fort pertinente.*
19. **Véronique s'est coincé le doigt**, *elle a mal au pied.*
20. *Le poulet a cuit*, **je l'avais mis dans le frigo**.
21. *L'oreille de Joel est tombée*, **il a pris une douche**.

22. **George est allé dans le désert**, *il a été pris dans une tempête de neige.*

23. **Marjorie est allée à la piscine**, *elle a perdu la tête.*

24. *Marjorie est allée à la piscine*, **elle a perdu la tête**.

B.4 Second annotation experiment

The clauses that are to be analysed are given in boldfont and in italics, but this time there are no indication as to which is the potential cause and which is the potential effect

1. Bref, **il ne fut plus permis, même au moins lettré des Yankees, d'ignorer un seul des faits relatifs à son satellite, ni à la plus bornée des vieilles mistress d'admettre encore de superstitieuses erreurs à son endroit.** *La science leur arrivait sous toutes les formes;* elle les pénétrait par les yeux et les oreilles; impossible d'être un âne...en astronomie.

2. Bref, il ne fut plus permis, même au moins lettré des Yankees, d'ignorer un seul des faits relatifs à son satellite, ni à la plus bornée des vieilles mistress d'admettre encore de superstitieuses erreurs à son endroit. **La science leur arrivait sous toutes les formes;** *elle les pénétrait par les yeux et les oreilles;* impossible d'être un âne...en astronomie.

3. Bref, il ne fut plus permis, même au moins lettré des Yankees, d'ignorer un seul des faits relatifs à son satellite, ni à la plus bornée des vieilles mistress d'admettre encore de superstitieuses erreurs à son endroit. La science leur arrivait sous toutes les formes; **elle les pénétrait par les yeux et les oreilles;** *impossible d'être un âne...en astronomie.*

4. **Jusqu'alors, bien des gens ignoraient comment on avait pu calculer la distance qui sépare la Lune de la Terre.** *On profita de la circonstance pour leur apprendre que cette distance s'obtenait par la mesure de la parallaxe de la Lune.* Si le mot parallaxe semblait les étonner, on leur disait que c'était l'angle formé par deux lignes droites menées de chaque extrémité du rayon terrestre jusqu'à

la Lune. Doutaient-ils de la perfection de cette méthode, on leur prouvait immédiatement que, non seulement cette distance moyenne était bien de deux cent trente-quatre mille trois cent quarante-sept milles (– 94,330 lieues), mais encore que les astronomes ne se trompaient pas de soixante-dix milles (– 30 lieues).

5. Jusqu'alors, bien des gens ignoraient comment on avait pu calculer la distance qui sépare la Lune de la Terre. **On profita de la circonstance pour leur apprendre que cette distance s'obtenait par la mesure de la parallaxe de la Lune.** *Si le mot parallaxe semblait les étonner, on leur disait que c'était l'angle formé par deux lignes droites menées de chaque extrémité du rayon terrestre jusqu'à la Lune.* Doutaient-ils de la perfection de cette méthode, on leur prouvait immédiatement que, non seulement cette distance moyenne était bien de deux cent trente-quatre mille trois cent quarante-sept milles (– 94,330 lieues), mais encore que les astronomes ne se trompaient pas de soixante-dix milles (– 30 lieues).

6. Jusqu'alors, bien des gens ignoraient comment on avait pu calculer la distance qui sépare la Lune de la Terre. On profita de la circonstance pour leur apprendre que cette distance s'obtenait par la mesure de la parallaxe de la Lune. **Si le mot parallaxe semblait les étonner,** *on leur disait que c'était l'angle formé par deux lignes droites menées de chaque extrémité du rayon terrestre jusqu'à la Lune.* Doutaient-ils de la perfection de cette méthode, on leur prouvait immédiatement que, non seulement cette distance moyenne était bien de deux cent trente-quatre mille trois cent quarante-sept milles (– 94,330 lieues), mais encore que les astronomes ne se trompaient pas de soixante-dix milles (– 30 lieues).

7. Jusqu'alors, bien des gens ignoraient comment on avait pu calculer la distance qui sépare la Lune de la Terre. On profita de la circonstance pour leur apprendre que cette distance s'obtenait par la mesure de la parallaxe de la Lune. Si le mot parallaxe semblait les étonner, on leur disait que c'était l'angle formé par deux lignes droites menées de chaque extrémité du rayon terrestre jusqu'à la Lune. **Doutaient-ils de la perfection de cette méthode,** *on leur prouvait immédiatement que, non seulement cette distance moyenne était bien de deux cent*

trente-quatre mille trois cent quarante-sept milles (– 94,330 lieues), mais encore que les astronomes ne se trompaient pas de soixante-dix milles (– 30 lieues).

8. Jusqu'alors, bien des gens ignoraient comment on avait pu calculer la distance qui sépare la Lune de la Terre. On profita de la circonstance pour leur apprendre que cette distance s'obtenait par la mesure de la parallaxe de la Lune. Si le mot parallaxe semblait les étonner, on leur disait que c'était l'angle formé par deux lignes droites menées de chaque extrémité du rayon terrestre jusqu'à la Lune. Doutaient-ils de la perfection de cette méthode, on leur prouvait immédiatement que, **non seulement cette distance moyenne était bien de deux cent trente-quatre mille trois cent quarante-sept milles (– 94,330 lieues),** mais encore que *les astronomes ne se trompaient pas de soixante-dix milles (– 30 lieues).*

9. A ceux qui **n'étaient pas familiarisés avec les mouvements de la Lune,** *les journaux démontraient quotidiennement qu'elle possède deux mouvements distincts, le premier dit de rotation sur un axe, le second dit de révolution autour de la Terre, s'accomplissant tous les deux dans un temps égal, soit vingt-sept jours et un tiers* [C'est la durée de la révolution sidérale, c'est-à-dire le temps que la Lune met à revenir à une même étoile.].

10. **A ceux qui n'étaient pas familiarisés avec les mouvements de la Lune, les journaux démontraient quotidiennement qu'elle possède deux mouvements distincts, le premier dit de rotation sur un axe, le second dit de révolution autour de la Terre, s'accomplissant tous les deux dans un temps égal, soit vingt-sept jours et un tiers** [*C'est la durée de la révolution sidérale, c'est-à-dire le temps que la Lune met à revenir à une même étoile.*].

11. **Le mouvement de rotation est celui qui crée le jour et la nuit à la surface de la Lune**; *seulement il n'y a qu'un jour, il n'y a qu'une nuit par mois lunaire, et ils durent chacun trois cent cinquante-quatre heures et un tiers.* Mais, heureusement pour elle, la face tournée vers le globe terrestre est éclairée par lui avec une intensité égale à la lumière de quatorze Lunes. Quant à l'autre face, toujours invisible, elle

a naturellement trois cent cinquante-quatre heures d'une nuit absolue, tempérée seulement par cette «pâle clarté qui tombe des étoiles». Ce phénomène est uniquement dû à cette particularité que les mouvements de rotation et de révolution s'accomplissent dans un temps rigoureusement égal, phénomène commun, suivant Cassini et Herschell, aux satellites de Jupiter, et très probablement à tous les autres satellites.

12. Le mouvement de rotation est celui qui crée le jour et la nuit à la surface de la Lune; **seulement il n'y a qu'un jour, il n'y a qu'une nuit par mois lunaire,** et *ils durent chacun trois cent cinquante-quatre heures et un tiers.* Mais, heureusement pour elle, la face tournée vers le globe terrestre est éclairée par lui avec une intensité égale à la lumière de quatorze Lunes. Quant à l'autre face, toujours invisible, elle a naturellement trois cent cinquante-quatre heures d'une nuit absolue, tempérée seulement par cette «pâle clarté qui tombe des étoiles». Ce phénomène est uniquement dû à cette particularité que les mouvements de rotation et de révolution s'accomplissent dans un temps rigoureusement égal, phénomène commun, suivant Cassini et Herschell, aux satellites de Jupiter, et très probablement à tous les autres satellites.

13. **Le mouvement de rotation est celui qui crée le jour et la nuit à la surface de la Lune; seulement il n'y a qu'un jour, il n'y a qu'une nuit par mois lunaire, et ils durent chacun trois cent cinquante-quatre heures et un tiers.** Mais, heureusement pour elle, *la face tournée vers le globe terrestre est éclairée par lui avec une intensité égale à la lumière de quatorze Lunes.* Quant à l'autre face, toujours invisible, elle a naturellement trois cent cinquante-quatre heures d'une nuit absolue, tempérée seulement par cette «pâle clarté qui tombe des étoiles». Ce phénomène est uniquement dû à cette particularité que les mouvements de rotation et de révolution s'accomplissent dans un temps rigoureusement égal, phénomène commun, suivant Cassini et Herschell, aux satellites de Jupiter, et très probablement à tous les autres satellites.

14. Le mouvement de rotation est celui qui crée le jour et la nuit à la surface de la Lune; seulement il n'y a qu'un jour, il n'y a qu'une nuit par mois lunaire, et ils durent chacun trois cent cinquante-quatre heures et un tiers. Mais, heureusement pour elle, **la face tournée vers le**

globe terrestre est éclairée par lui avec une intensité égale à la lumière de quatorze Lunes. *Quant à l'autre face, toujours invisible, elle a naturellement trois cent cinquante-quatre heures d'une nuit absolue, tempérée seulement par cette «pâle clarté qui tombe des étoiles».* Ce phénomène est uniquement dû à cette particularité que les mouvements de rotation et de révolution s'accomplissent dans un temps rigoureusement égal, phénomène commun, suivant Cassini et Herschell, aux satellites de Jupiter, et très probablement à tous les autres satellites.

15. **Quelques esprits bien disposés, mais un peu rétifs, ne comprenaient pas tout d'abord que, si la Lune montrait invariablement la même face à la Terre pendant sa révolution, c'est que, dans le même laps de temps, elle faisait un tour sur elle-même.** *A ceux-là on disait: «Allez dans votre salle à manger, et tournez autour de la table de manière à toujours en regarder le centre; quand votre promenade circulaire sera achevée, vous aurez fait un tour sur vous-même, puisque votre oeil aura parcouru successivement tous les points de la salle.* Eh bien! la salle, c'est le Ciel, la table, c'est la Terre, et la Lune, c'est vous!» Et ils s'en allaient enchantés de la comparaison.

16. Quelques esprits bien disposés, mais un peu rétifs, ne comprenaient pas tout d'abord que, si **la Lune montrait invariablement la même face à la Terre pendant sa révolution**, c'est que, *dans le même laps de temps, elle faisait un tour sur elle-même.* A ceux-là on disait: «Allez dans votre salle à manger, et tournez autour de la table de manière à toujours en regarder le centre; quand votre promenade circulaire sera achevée, vous aurez fait un tour sur vous-même, puisque votre oeil aura parcouru successivement tous les points de la salle. Eh bien! la salle, c'est le Ciel, la table, c'est la Terre, et la Lune, c'est vous!» Et ils s'en allaient enchantés de la comparaison.

17. Quelques esprits bien disposés, mais un peu rétifs, ne comprenaient pas tout d'abord que, si la Lune montrait invariablement la même face à la Terre pendant sa révolution, c'est que, dans le même laps de temps, elle faisait un tour sur elle-même. A ceux-là on disait: «Allez dans votre salle à manger, et tournez autour de la table de manière à

toujours en regarder le centre; quand **votre promenade circulaire sera achevée**, *vous aurez fait un tour sur vous-même*, puisque votre oeil aura parcouru successivement tous les points de la salle. Eh bien! la salle, c'est le Ciel, la table, c'est la Terre, et la Lune, c'est vous!» Et ils s'en allaient enchantés de la comparaison.

18. Quelques esprits bien disposés, mais un peu rétifs, ne comprenaient pas tout d'abord que, si la Lune montrait invariablement la même face à la Terre pendant sa révolution, c'est que, dans le même laps de temps, elle faisait un tour sur elle-même. A ceux-là on disait: «Allez dans votre salle à manger, et tournez autour de la table de manière à toujours en regarder le centre; quand votre promenade circulaire sera achevée, **vous aurez fait un tour sur vous-même**, puisque *votre oeil aura parcouru successivement tous les points de la salle.* Eh bien! la salle, c'est le Ciel, la table, c'est la Terre, et la Lune, c'est vous!» Et ils s'en allaient enchantés de la comparaison.

19. Quelques esprits bien disposés, mais un peu rétifs, ne comprenaient pas tout d'abord que, si la Lune montrait invariablement la même face à la Terre pendant sa révolution, c'est que, dans le même laps de temps, elle faisait un tour sur elle-même. A ceux-là on disait: «Allez dans votre salle à manger, et tournez autour de la table de manière à toujours en regarder le centre; quand votre promenade circulaire sera achevée, vous aurez fait un tour sur vous-même, puisque votre oeil aura parcouru successivement tous les points de la salle. Eh bien! **la salle, c'est le Ciel, la table, c'est la Terre, et la Lune, c'est vous!**» Et *ils s'en allaient enchantés de la comparaison.*

Appendix C

Implicit reason occurrences in children tales

1. Les paysans étaient désespérés. Toutes les graines qu'ils plantaient pourrissaient les pieds dans l'eau.

2. il s'ennuyait très souvent. Dans ce pays-là, l'empereur n'avait pas le droit d'inviter des amis pour s'amuser.

3. À ce moment, on entend un air de trompettes. La mère de l'empereur vient d'arriver dans le palais !

4. la maman, assise près de la cheminée, reste là à pleurer. Elle est si fatiguée qu'elle ne peut plus travailler, et elle n'a plus du tout d'argent pour acheter à manger.

5. ils poussent un cri d'étonnement. Ce bébé a un oeil noir et un oeil tout gris.

6. Un paysan très riche et très vieux gardait cent chats dans sa maison, pas très loin d'un petit village. Il n'oubliait pas qu'autrefois ces chats avaient sauvé ses récoltes en faisant la chasse aux mulots, aux rats et aux souris des champs qui dévastaient tout le pays.

7. Justement, la servante s'en va, pleine de griffures et d'égratignures: elle a lancé de l'eau bouillante dans les pattes de trois chats tigrés, et ils se sont vengés.

8. Ils ronronnent doucement en la voyant si travailleuse.

9. Lisa est bien plus heureuse qu'avant, mais en pensant à sa mère et à Lola elle a tout de même le coeur serré. Elle aimerait bien les revoir.

10. Le tailleur de pierre était content, il aimait sa vie.

11. Mais ni Barbebouc ni la princesse Minuscule ne s'ennuient. Ils ont tellement de choses à se raconter !

12. je n'en veux plus de ta vie en or, moi, j'aime la vraie vie bien vivante, avec des fruits qui sentent bon, de la vaisselle qui sonne et des chaises en bois léger.

13. Arrivé au pays de l'Abracabizcorne, le prince n'en croit pas ses yeux : les hommes lavent leur barbe dans du vin, les escaliers montent vers rien et les chats chantent tralalalila.

14. Petit à petit, il comprend pourquoi les rois épousent les princesses. Le roi Léon est amoureux !

15. Diamant ne répond rien. Il est bien trop occupé à fabriquer un théâtre de marionnettes dans une noix de coco.

16. le Thon se plaint d'avoir mal à la gorge depuis une semaine. Quelque chose le gêne pour avaler.

17. Mais, bientôt, on ne trouva plus de barbier. Ils étaient tous en prison !

18. Pourtant, il hésite, il n'est pas sûr d'être amoureux.

19. Pourtant, il hésite encore, il n'est pas sûr d'être amoureux.

20. Le soir même, la mère envoie Ludivine chez la sorcière. Elle se dit qu'avec un peu de magie, on ne sait jamais, après tout !

Appendix D

A small sample of the feature extraction programme

D.1 Lemmatizing

Note that lemmatization can also be done in advance and lemmas loaded from a file, which is much faster.

```
1
2
3  sub lemmatize
4  {
5    #input: a string
6    #output: an array of words, lemmas and POS
7    my $sentence=shift;
8
9    $sentence=~s/"/\\"/g;#escape quotes
10   $sentence=~s/'/\\'/g;
11   $sentence=~s/!/\\!/g;#bash doesn't like them.
12   $sentence=~s/;/\\;/g;#bash doesn't like them.
13   $sentence=~s/`/\\'/g;#bash doesn't like them.
14   my @res;
15   my $cmdOut=`echo "$sentence" | /home/cecile/data/
         these/outils/TT/cmd/tree-tagger-english`;
16   my @lines=split("\n", $cmdOut);
```

```
17    foreach my $l (@lines)
18    {
19      if($l=~/^(.+)\t(.+)\t(.+)$/)
20      {
21        my @tmpArr=($1,$2,$3);#word tag verb
22        push (@res, \@tmpArr);
23      }
24      elsif($l =~ /^<.+>$/)
25      {
26        my @tmpArr=($l,"","");
27        push (@res, \@tmpArr);
28      }
29      else {die "wrong input format: $l"}
30    }
31    return @res;
32  }
```

D.2 Getting hypernyms for content words

```
1   sub getContentWordsHype
2   {
3   #in: a string of words, seperated by spaces.
4   #out: a string of hyperonymes
5     my $le=shift;
6     my $wn = shift;
7     my @lems=@$le;
8     my @pair;
9     my $cl1="";
10    my $cl2="";
11    my $state=0;
12    foreach my $words (@lems)
13    {
14      if($state==0 && @$words[0] eq "<cl1>"){$state=1;}
15      elsif($state==0 && @$words[0] eq "<cl2>"){$state
            =2;}
```

```
16      elsif(($state==1 || $state==2) && (@$words[0] eq "
          </cl1>" || @$words[0] eq "</cl2>")){$state=0;}
17      elsif(@$words[1] eq "JJ" || @$words[1] eq "JJR" ||
          @$words[1] eq "JJS" || @$words[1] eq "NN" ||
          @$words[1] eq "NNS" || @$words[1] eq "NNP" ||
          @$words[1] eq "NNPS" || @$words[1] eq "RBR" ||
          @$words[1] eq "RBS" || @$words[1] eq "RP" ||
          @$words[1] eq "VB" || @$words[1] eq "VBD" ||
          @$words[1] eq "VBG" || @$words[1] eq "VBN" ||
          @$words[1] eq "VBP" || @$words[1] eq "VBZ")
18      {
19        if(@$words[2] ne "%" && !(@$words[2]=~/\./))#
            boostexter doesn't like them.
20        {
21          my $pos;
22          if(@$words[1] eq "JJ" || @$words[1] eq "JJR" ||
              @$words[1] eq "JJS")
23          {$pos="a";}
24          elsif(@$words[1] eq "NN" || @$words[1] eq "NNS"
              || @$words[1] eq "NNP" || @$words[1] eq "
              NNPS")
25          {$pos="n";}
26          elsif(@$words[1] eq "RBR" || @$words[1] eq "RBS
              " || @$words[1] eq "RP")
27          {$pos="r";}
28          else
29          {$pos="v";}
30          my $hype;
31          if(@$words[2] eq "<unknown>"){$hype="<unknown>"
              ;}
32          else
33          {
34            my $exist = join(",", $wn->queryWord(@$words
                [2]."#".$pos));
35            if($exist eq ""){$hype=@$words[2];}
36            else
```

```
37              {
38                  $hype= join("␣", $wn->querySense(@$words
                        [2]."#".$pos."#1", "hypes"));
39                  if ($hype eq ""){$hype = @$words[2]}
40              }
41          }
42          if($state == 1){$cl1.="␣".$hype;}
43          elsif($state == 2){$cl2.="␣".$hype;}
44        }
45      }
46    }
47    push (@pair, $cl1);
48    push (@pair, $cl2);
49    #return string of correct lemms
50    return @pair;
51  }
```