Human judgements on causation in French texts

Cécile Grivaz

Département de Linguistique
University of Geneva, Switzerland
cecile@grivaz.net

Abstract

The annotation of causal relations in natural language texts can lead to a low inter-annotator agreement. A French corpus annotated with causal relations would be helpful for the evaluation of programs that extract causal knowledge, as well as for the study of the expression of causation. As previous theoretical work provides no necessary and sufficient condition that would allow an annotator to easily identify causation, we explore features that are associated with causation in human judgements. We present an experiment that allows us to elicit intuitive features of causation. We test the statistical association of features of causation from theoretical previous work with causation itself in human judgements in an annotation experiment. We then establish guidelines based on these features for annotating a French corpus. We argue that our approach leads to coherent annotation guidelines, since it allows us to obtain a $\kappa = 0.84$ agreement between the majority of the annotators answers and our own educated judgements. We present these annotation instructions in detail.

1. Introduction

Although causation has been extensively studied since Hume's fundamental work (Hume, 1740), there are no consensus tests that would allow an annotator to identify easily a segment of text as causal, and human judgement tends to be inconsistent. For example, in the absence of a context it is unclear whether or not sentence 1 is causal. Simply reformulating sentence 1 as sentence 2 does not help to decide, because knowing whether or not sentences 1 and 2 are synonymous requires an a priori knowledge of the causation in 1. Even in less extreme cases, such as 3, human judgement is not always coherent because some would argue that his arm was broken by falling rather than by skiing. Moreover, naive annotators are inconsistent in their judgement of 4, even tough experts regard it as non-causal.

(1) John got sick; he took a shower.

(2) Taking a shower caused John to get sick.

(3) He broke his arm while skiing.

(4) It is triangle; it has three sides.

Previous work on the annotation of causal relations based the annotation instructions either solely on the annotator's intuition (Carlson et al., 2001) or on linguistic tests (Inui, 2005; Bethard et al., 2008). The latter can be ambiguous in regard to causation, due to such factors as non-causal use of causal connectors (e.g. speech act or epistemic use of *because*). Previous theoretical work rarely gives other tests for causation than reformulation such as *x is the cause of y*. However, it extensively describes features of causation, providing necessary conditions, but, to our knowledge, no sufficient condition. We address these shortcomings by studying systematically the features of causation that allow annotators to clarify their intuitions. We then demonstrate that our approach is successful in an annotation study (section 4.).

To the best of our knowledge, there is no French corpora annotated with causal relations. Such a corpus would be useful for the general study of causation, and particularly for the evaluation of systems that find these relations automatically. This paper, set in a French framework, describes annotation instructions for causal relations that are expressed between clauses. These instructions allow annotators to give answers closely resembling our educated intuitions. By using the identifying features of causation, these rules help to remove several difficult ambiguities. In this paper, we test the following hypotheses.

Hyp1 Human reasoning consciously makes use of several intuitive tests of causation.

Hyp2 Several features of causation are statistically associated with the annotators' recognition of causality.

Hyp3 Our annotation rules allow annotators to coherently identify causation.

We give evidence against Hyp1 by examining such features in section 2. We provide evidence for Hyp2 by identifying features that are associated, in human judgements, with causation (section 3.) Finally, we present our annotation instructions and give evidence for Hyp3 by showing that these rules are coherent since they lead to a high agreement ($\kappa = 0.84$) between the majority of annotators and our predictions (section 4.)

2. Intuitive Features of Causation

We describe an experiment that elicits the intuitive features or tests of causation that are consciously used in causal reasoning. We wanted to test hypothesis Hyp1, which states that such features exist. Moreover, we wanted to know if these features are different from previous theoretical work, and if they can be integrated into our annotation instructions. We found evidence against Hyp1 through this experiment.

We asked nine naive subjects to decide if a causal relation was expressed in ten short segments of French texts. We asked them to justify their answers systematically. We wanted to discover what justification would be given in the

absence of specific rules or of an annotation manual. We then analysed the features of causation that were used in the justifications.

The segments were extracted from the French part of the BAF corpus (RALI laboratory, 1997), which presents several types of texts: institutional text, academic writings, and a novel. The segments consisted of up to four sentences, but mostly of one or two. We also gave some elements of context for each segment. There were 6 types of segments: texts containing *parce que/because*, containing *donc/so*, containing *mais/but*, and texts that had contained one of those connectives, but in which we removed the connective before giving it to the subjects. The connective *mais* was chosen as a canonical non-causal connective. We wanted to study the answers in the presence of causal or non-causal connectives, and with or without the aid of the connectives, in order to elicit the largest possible number of justifications.

Most (72.2%) of the segments were judged as causal. This is not surprising since, most (80.0%) of the segments either had a causal connective, or had one that was removed.

We classified the justifications into five types: rewording, linguistic tests, presence of an explicit marker, presence of a non-causal relation and others.

Rewording includes justifications that are a rewording of the instructions (*Is a causal relation expressed?*) and that don't contain information. For example, subjects responded 5, 6 or 7, for a negative case. This justification was the most common for negative cases.

(5) Une explication est donnée. (There is an explanation.)

(6) Il donne des raisons. (He gives reasons.)

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

Linguistic tests consist of placing an explicit marker in the text and then deciding if the resulting text is a rephrasing of the original wording. Subjects used the connectives *c'est parce que/it's because*, *parce que/because* and *donc/so*. When building a test sentence with an explicit causal connective, they often replaced the connectives that we took out of the text segment. They also used many different verbs such as *est le fruit de/is the result of*, *entraîne/leads to* or *permet/allows*. When doing so, they nominalised the clauses they were testing, or used the phrase *le fait que/the fact that*. We also classify uses of *est la cause de/is the cause of* in this category. The difference between the rewording and the linguistic test categories is the use of the clauses that subjects were testing in the justification. For example, 8 is classified as rewording, while 9 is a linguistic test. This justification was the most common for causal cases.

(8) Une cause est exprimée. (There is a cause expressed.)

(9) Le fait que *clause1* est la cause du fait que *clause2*. (The fact that *clause1* is the cause of the fact that *clause2*.)

Some justifications point to the presence of an **explicit marker** in the text, such as *parce que/because* or *mais/but*. We find for example 10 or 11. 66.7% of the segments containing a causal connective led to this type of justification, and only one (11.1%) of the segments bearing a non-causal connective *mais/but* was justified this way.

(10) *Donc* apparaît. *Donc* introduit une conséquence. (*So* appears. *So* signals a consequence.)

(11) Le *mais* exprime une nuance, une restriction dans ce cas précis. (The *but* expresses a nuance, a restriction in this precise case.)

The presence of a **non-causal relation** was sometimes used to justify negative answers. Annotators would argue that the text could not be causal, because they could identify a non-causal relation in the segment. For example, they would write 12.

(12) C'est une description, la seconde phrase apporte seulement une précision. (It is a description, the second sentence only further refines the first one.)

Finally, the **other** category contained justifications that could not be classified into any of the other four groups and mainly included instances where the subject had drawn a question mark instead of writing a justification.

Figure 1 shows the number of justifications for each type of answer that is positive or negative in regard to causation. Note the high number of rewordings, which do not explain a causal judgement further. Rewording is the most common justification for negative cases.

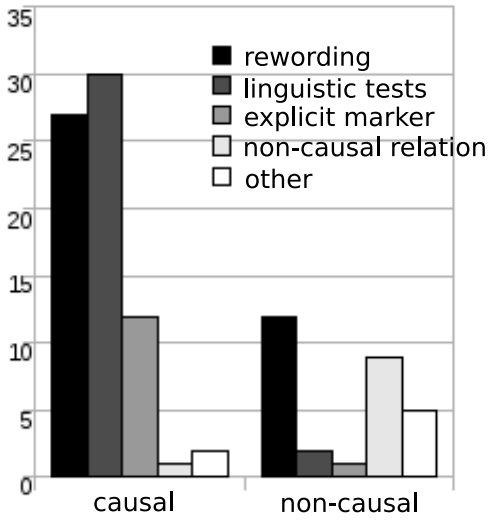

Figure 1: Types of justifications for causal intuitions. This graph shows the number of each type of justification for texts that were considered causal on the left, and non-causal on the right.

The variety of justifications is surprisingly low. Remarkably, there were no reference to the real-world events that the clauses described, but only to the text itself. For instance, no one mentioned the counterfactual argument "had

the cause not happened the effect would not have happened either", although such an argument is common in theoretical work. The high number of linguistic tests and the absence of real-world references might be explained by the fact that the analysed occurrences were textual, and that the subjects were linguistics students. A video input would probably lower the amount of linguistic tests in favour of real-world reasoning.
Preliminary tests also showed that given a single instruction such as 13, experimental subjects tended to omit justifications altogether, or gave justifications only when they could find a causal relation, and not on negative cases. In the final experiment, we repeated the instructions after each segment of text, and subjects were asked to cross out the unnecessary part and finish the sentence 14.

(13) Y a t'il une relation causale exprimée dans ce texte ? Justifiez votre réponse. (Is there a causal relation expressed in this text ? Justify your answer.)

(14) Je pense qu'une/aucune relation causale est exprimée dans ce texte parce que... (I believe that a/no causal relation is expressed in this text because...)

We believe that the difficulty that we encountered in obtaining systematic justifications, the lack of variety in justifications and the high amount of rewording provide evidences that hypothesis Hyp1 is invalid. Human reasoning does not consciously make use of intuitive tests of causation. If people are sensitive to features of causation, these features are not directly accessible for causal reasoning. Subjects can decide if something is causal or not, but they cannot describe how they know it.

3. Association of the judgement of causation and its features

Hypothesis Hyp2 states that some features of causation are statistically associated with the readers' recognition of causality. We verified this hypothesis by asking six naive subjects to determine if causation as well as a number of its features were present in 24 short French texts. The texts were artificial and ambiguous in regard to causation.
We chose the tested features based on the results of the previous experiment as well as on theoretical work. The features are: temporal order, ability to build a causal chain, linguistic test by using *parce que/because* and *donc/so*, counterfactuality and paraphrases.
Temporal asymmetry is necessary to causation, as the cause cannot happen after the effect (Hume, 1740). In many cases, the cause happen before the consequence as in 15. In cases of direct causation, it happens immediately before the consequence as in 16. The cause and the effect can also happen at least partially simultaneously as in 17. To assess the effect of temporal order, we asked the annotators to identify whether the potential cause was before, right before, or at least partially simultaneous to the effect.

(15) John is cold, he went out without his coat.

(16) The glass reached the floor and broke.

(17) He got tired driving.

A **causal chain** is a chain of direct causes and consequences that can be associated with any causal relation (Moeschler, 2003). For example 18 can be associated with the causal chain 19. We asked the subjects if they could build such a chain between the events.

(18) John fell, Mary had pushed him.

(19) Mary pushes John → John is off balance → John falls.

We also asked the annotators to decide the **counterfactual** case, that if the potential cause had not happened whether or not the potential effect would have happened anyway. This classical property of causation (see, for example (Reboul, 2005)), is one of the features that allows an annotator to differentiate causation from logical implications.
Finally, since we believe that an event cannot be its own cause, we asked them if the two clauses were **paraphrases** referring to the same event. We expected this last feature to be negatively associated with causation.
Figure 2 shows the amount of positive answers for each feature for cases that were analysed as causal or non-causal. Some features are indeed often associated with causation, particularly, causal chains, the linguistic test with *parce que* and conterfactuality are most often identified in the presence than in the absence of causation. We tested this associations with a Fisher's exact test.
Figure 3 shows the *p*-values of Fisher's exact test of association between each feature and causation in the annotations. A smaller *p*-value indicates a higher association. Partial simultaneity of the events, the ability to build a causal chain, linguistic tests and counterfactuality were statistically associated with causation in human judgements, confirming our hypothesis Hyp2.

4. Annotation instructions

Annotators must rely on intuition in order to recognise causality, as no easily usable test conditions exist. However our results from the previous section suggest that features that are associated with causation can be identified to allow annotators to clarify their intuition. Our instructions consist of a number of such features and resolve the ambiguities of several difficult cases. Some of the features are typical of causation while some allow annotators to rule out non-causal cases.
The instructions are based on intuitive features of causation and on features that were associated with causation in the previous experiment. Moreover, the rules are based on disambiguation tests for difficult cases. We identified these cases in two manners. First, we looked for examples in which the majority of annotators did not respond as expected in the previous experiment. Second, we asked two linguists that study causation to annotate a corpus of ambiguous examples, resulting in a low inter-annotator agreement ($\kappa = 0.32$). Other difficult cases could then be found in the divergent examples. For each divergent case, we developed, together with the linguists, a test that allows annotators to remove ambiguity in similar cases. In some cases no such test could be found, and such cases could only be used to draw the attention of the annotators to the fact that

special care should be taken in such cases. We will detail the features in the next paragraphs.

The features that helped to rule out non-causal occurrences were temporal order, counterfactuality and ontological asymmetry. The test of **temporal order** states that if the potential cause occurs after the potential effect then the example is not causal. We believe that a precise analysis of temporality is not necessary, as only the precedence of the potential consequence on the potential cause helps ruling out non causal cases.

The **counterfactuality** test is the following: would the potential effect have probably happened even in the absence of the potential cause? If so, the example is not causal. This test allows annotators to rule out non causal cases such as 20 that are not causal and display no conterfactuality.

(20) My bus will leave soon, I just finished my breakfast.

Finally, there should be an **ontological asymmetry** in causation (Hume, 1740). If a first event is the cause of a second event, then the second event can only very rarely be the cause of the first event. This feature holds even if the two events happen simultaneously. For example, in 21 *It is John's birthday* can be the cause of *John is happy*, but *John is happy* cannot be the cause of *it is John's birthday*. One could imagine some cases of circular causation, but we believe that those cases are rare, and both ways would be very explicit separately in the text, because of the strangeness of the phenomenon (in which case both occurrences should be annotated as causal separately, and each in only one direction). We tested this by stating that if it is difficult to choose which event is the cause and which is the effect then the example is not causal. This last test allows annotators to rule out cases such as 22.

(21) John was happy because it was his birthday.

(22) It is a triangle; it has three sides.

Features that helped to clarify intuition were the ability to build causal chains and the linguistic tests.

In difficult cases the instructions ask the annotator to try to build a **chain** of direct causes and effects between the events. Although it is possible that building a causal chain requires a prior intuition of whether the occurrence is causal or not, we believe that trying to build causal chains can help annotators clarify their intuition, especially in cases of events that are far from each other in time and where the causation is very indirect.

Linguistic tests were very present in the first experiment that elicited intuitive features of causation. They were also statistically associated with causation in the previous experiment. They present two drawbacks. First, it is sometimes difficult to know if an explicitly causal sentence is synonymous with its implicit counterpart and it might require prior knowledge of whether the occurrence is causal or not. Second, *parce que/because* and *donc/so* can be non-causal in their epistemic or speech act usage. In a sentence such as 23 the second clause is not the cause of the first clause but the cause of the belief that the sun is about to rise, and we did not want this kind of ellipses to be annotated as causal. In a similar fashion, we wanted 24 to be annotated as causal, but the other way around. We did not want the annotator to annotate the ellipsed *George's jacket is not on the chair* cause *my belief that George is out*, but we wanted them to annotate the causal relation *George is out* cause *his jacket is not on the chair*. In a similar fashion, we did not want the annotators to identify speech acts as 25 as causal, either.

(23) The sun is about to rise because it's 7.30 a.m.

(24) George is out because his jacket is not on the chair.

(25) Hurry up because we're going to be late.

Despite these drawbacks, linguistic tests provide an intuitive and easy way to clarify one's intuition on causation.

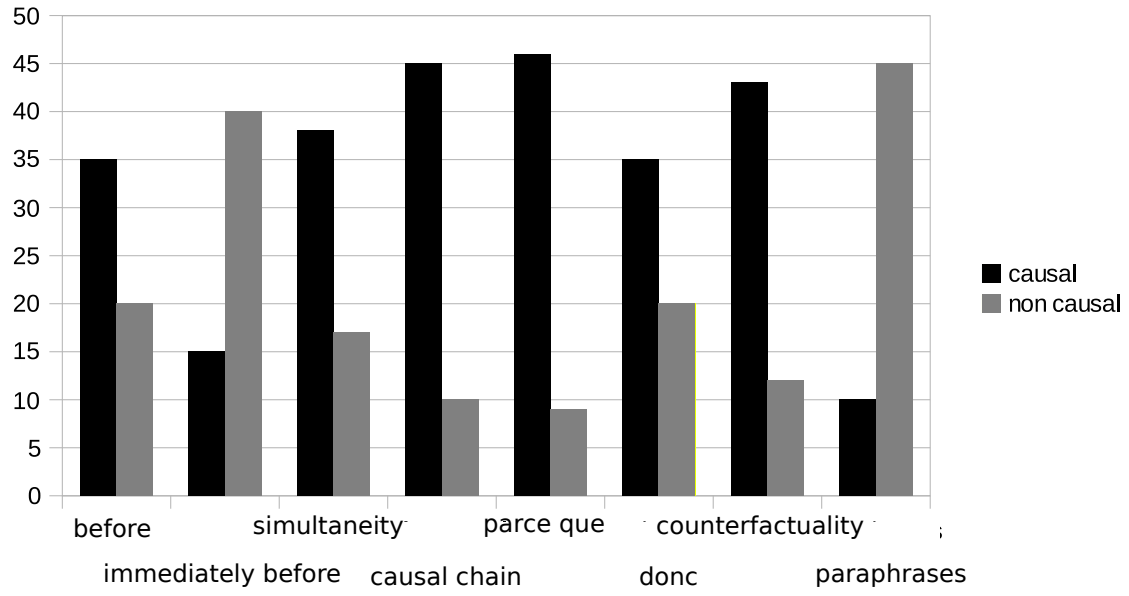

Figure 2: Amount of positive answers for each feature for cases that were analysed as causal or non-causal

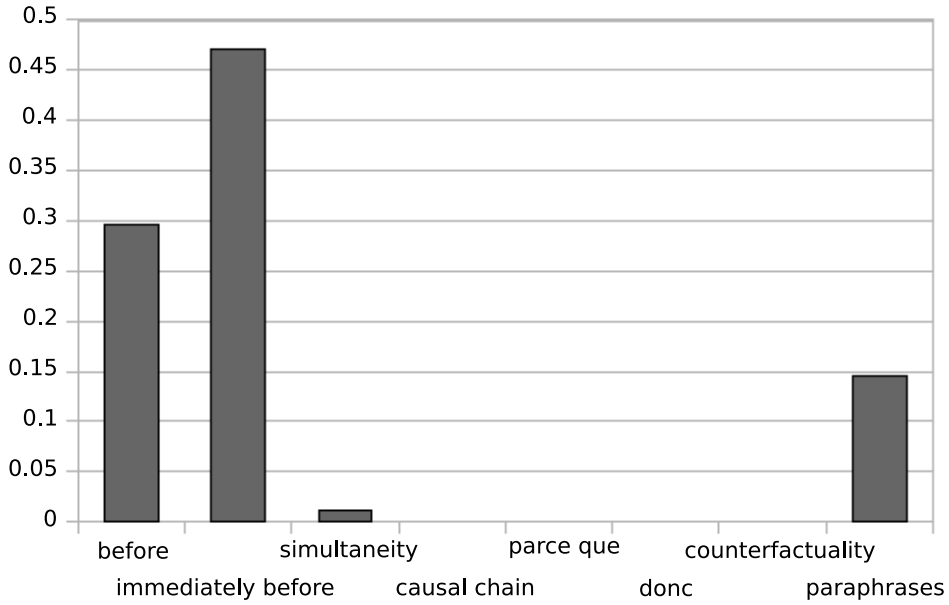

Figure 3: The *p*-value of Fisher's exact test of association for each feature and the presence of causation. Smaller values indicate a higher association.

We explained the non-causal uses in detail in the annotation manual, and proposed the linguistic tests with *parce que* and *donc*. We also added a linguistic test with *c'est parce que/it is because*, which does not, as far as we know, have a speech act or epistemic usage, and thus is much more discriminative.
Finally we detailed in the manual two difficult types of occurrences. First, cases where the cause event is not explicit but is happening inside the context of an explicit event such as in 26. Here, the cause of breaking one's leg is not skiing but an event inside the context of skiing, such as falling. In such cases, linguists disagree as to whether the case is causal or not. In this study, wanting a broad definition of causation, we asked the annotators to mark the occurrence as causal.

(26) He broke his leg while skiing.

Second, judgement on causation become more difficult, if the potential cause is negated such as in 27, or if it's aspectual class is that of a state as in 28. We drew the annotator's attention to these cases, and advised them to try to clarify their intuition by building a causal chain or by searching a more general law from which an inference could be drawn such as *Raphaelle likes tall men*. We also asked them to be very cautious about verifying the ontological asymmetry in such cases.

(27) Since no other study of this type have been done before, this one will be very interesting.

(28) Raphaelle finds John attractive, because he is very tall.

We tested these instructions by using four annotators on 15 text segments. The text segments were from a novel (*De la terre à la lune*, by Jules Verne). Annotators could see whole paragraphs in which the text to be analysed appeared, so that they had sufficient contextual information. We selected text segments discarding those that were obviously non-causal, so as to spare annotation time and get more significant results. Each text was analysed by each of the four annotators.
The annotators did not receive any training besides the instructions. The annotations were noisy and led to a mediocre agreement between pairs of annotators. However, we could reduce the noise by selecting the majority annotations. We obtained a very high $\kappa = 0.84$ between the majority of the annotators answers and our own educated judgements. By using our instructions, we were mostly able to communicate our causation criteria to the annotators.

5. Previous Works

To the best of our knowledge, there is no French corpus annotated with causal relations. In English, however, Carlson and colleagues (Carlson et al., 2001) annotated a large corpus with several relations in the framework of Rhetorical Structure Theory, which contains causal relations.
Carlson and colleagues annotated 385 documents of the Penn Treebank (Marcus et al., 1994). In their work, the cause relation consists of three subcategories: *cause*, *result*, and *consequence*. The cause and result categories are differentiated by the relative importance given in the text to either the cause or the effect (Carlson and Marcu, 2001). The difference between the cause-result type of relation and the consequence type is that consequence is a less direct causal link. We also consider the *reason* relation as causal. The difference between this relation and the others is that here the result is carried out by an animate agent. The guidelines do not further define the concept of cause. However, in case of ambiguity, the annotators are instructed to select the less general relation that applies.

In Japanese, Inui (Inui, 2005) annotated 750 social domain newspaper articles with causal relations signalled or not by a causal marker. Inui not only annotated causal relations that held between two clauses such as *John fell because Mark pushed him* but also between noun phrases such as *the lack of rain caused a drought*. Inui used several linguistic tests to identify causal relations. If the sentence resulting from applying the linguistic test was semantically and syntactically correct, then the sentence was annotated as causal. If it was not, then another test would be used until the text passed a test or failed all tests. Inui also added a necessity tag to each causal relation. This tag would indicates if the relations usually held or not.

Compared to Carlson and colleagues work, this paper is more specific. It focuses on only one relation, not all relations defined by RST, and we do not subcategorise the cause relation further. We do not make a difference between cause and reason, for example.

A difference between our work and both previous papers presented here, is that the focus of our work is to precisely define guidelines for identifying causal relations. We used features of causation that are not only rewording or linguistic tests, and we believe those features helped make the annotations more coherent.

6. Conclusion

In this paper, we described an experiment to discover intuitive features of causation. With this experiment we provided evidences against our hypothesis Hyp1, which stated that intuitive consciously used tests of causation exist.

We provided evidence for hypothesis Hyp2, which states that there exist features of causation that are statistically associated with causality by calculating this association for a number of features with an annotation experiment on causaly ambiguous texts.

We used these features as well as features from previous theoretical work to write an annotation manual for causation. We presented these annotation instructions which lead to a high agreement between our answers and the majority of the annotators. We believe that this shows that our understanding of causation is coherent, and can be effectively transmitted through our instructions. We thus provided evidence for hypothesis Hyp3.

We believe that a similar methodology can be used for other difficult annotation tasks. We also believe that discovering intuitive features of a task, as well as discussions resulting from different annotations by several experts, can lead to a better modelling and comprehension of complex linguistic features.

Our annotation manual does not lead to a sufficiently high kappa score between non-expert annotators. We believe it should be further disambiguated until the kappa score is high enough to annotate a useful French corpus. Particularly, we would like to explore the subjectivity of causation, and find ways to identify the point of view of the text segment itself, and not of the annotators. We also believe annotations would benefit from more training of the annotators. Finally, we plan to experiment with the annotation of the textual boundaries of the cause and the effect events, which is essential for annotating a corpus.

Finally, this work is part of a project that aims at developing a computer program capable of doing this annotation task automatically. We believe that this work helped much in clarifying the computer task, and that our annotation manual will be very useful to evaluate the program results.

7. Acknowledgement

This work was financed by a Swiss National Foundation project (100012-113382) and by a scholarship from the Ernest Boninchi Foundation.

8. References

Steven Bethard, William Corvey, Sara Klingenstein, and James H. Martin. 2008. Building a corpus of temporal-causal structure. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Lynn Carlson and Daniel Marcu. 2001. Discourse tagging reference manual. Technical report, Univ. of Southern California / Information Sciences Institute.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2001. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*, pages 1–10, Morristown, NJ, USA. Association for Computational Linguistics.

David Hume. 1740. *A Treatise of Human Nature*. London, Millar.

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

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

Jacques Moeschler. 2003. Causality, lexicon, and discourse meaning. *Rivista di Linguistica, 15.2*, pages 343–369.

Université de Montréal RALI laboratory. 1997. Corpus de bitextes anglais-français.

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