Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) @LREC 2022, pages 151–159
Marseille, 24 June 2022

A Cognitive Approach to Annotating Causal Constructions in a Cross-Genre Corpus

Angela Cao, Gregor Williamson, Jinho D. Choi
Emory University
Atlanta, GA 30322, USA
Department of Computer Science
{angela.yuan.cao, gregor.jude.williamson, jinho.choi}@emory.edu

Abstract

We present a scheme for annotating causal language in various genres of text. Our annotation scheme is built on the popular categories of CAUSE, ENABLE, and PREVENT. These vague categories have many edge cases in natural language, and as such can prove difficult for annotators to consistently identify in practice. We introduce a decision based annotation method for handling these edge cases. We demonstrate that, by utilizing this method, annotators are able to achieve inter-annotator agreement which is comparable to that of previous studies. Furthermore, our method performs equally well across genres, highlighting the robustness of our annotation scheme. Finally, we observe notable variation in usage and frequency of causal language across different genres.

Keywords: causal annotation, cross-genre annotation, manual annotation, semantic relations

1. Introduction

The way we comprehend the world through notions of *causer* and *caused* dominates how we form notions of responsibility, make decisions based on world knowledge, and relate events to one another. For example, are the addictive properties of nicotine or genetics to blame for the correlation between lung cancer and smoking (Gundle et al., 2010)? Do language patterns limit channels of thought, or do channels of thought limit language patterns (Whorf, 1956)? Did Eve make Adam eat the apple (Pearl, 2009)? In line with previous work on annotating causal relations in text, which makes the author's internal causal reasoning primed for the purpose of analysis, this paper presents the Constructions of CAUSE, ENABLE, and PREVENT (CCEP) corpus. This project builds mainly upon the Bank of Effects and Causes Stated Explicitly (BECauSE) of Dunietz (2018), Dunietz et al. (2017b), and Dunietz et al. (2015) while incorporating a force dynamics approach to causation categorization first introduced by Wolff et al. (2005) and defined in Table 1. We provide a multi-test approach for annotators in order to ground intuitions about the vague concepts of CAUSE, ENABLE and PREVENT (abbreviated as C, E, and P, respectively) in a straightforward and accurate manner. Unlike the majority of previous annotation studies on causal language, which typically work with news data, the CCEP is annotated on a cross-genre dataset including short stories, Reddit posts, in addition to news data, to provide insights into how causal relations are described differently across genres. In the next section, we provide a brief overview of the theoretical motivation behind the categories of CAUSE, ENABLE, and PREVENT. Following this, in section 3, we provide an overview of related causal annotation research in order to contextualize the present study. Next, in section 4, we provide a description of our annotation guidelines and supporting materials[1]. In section 5, we describe the training methods and tools used during annotation. Section 6 presents our IAA scores, comparing them to other causal annotation projects, which demonstrates the robustness and reliability of the present scheme. Finally, we discuss future directions for research as well as outstanding practical and theoretical issues in section 7, before concluding in section 8.

2. Theoretical motivation

The force dynamics theory of causation (Wolff et al., 2005; Wolff, 2007) is an approach to knowledge representation that encodes how causal judgements may be formed in human cognition (Wolff and Thorstad, 2017). The concepts of CAUSE, ENABLE, and PREVENT are distinguished according to "various patterns of tendency, relative strength, rest, and motion between an *affector* and a *patient*" (Wolff and Zettergren, 2002, p.2). More specifically, these notions are defined in terms of whether the affector and the patient act in concordance, whether there is a tendency for the patient toward the result, and whether the result occurs or not. The specific attributes of each category are given in Table 1.

	Patient tendency toward result	Affector-Patient Concordance	Occurrence of result
CAUSE	N	N	Y
ENABLE	Y	Y	Y
PREVENT	Y	N	N

Table 1: Wolff et al.'s (2005) force dynamics theory of causation.

While useful, this table is somewhat misleading, as boundaries between the three classes are often unclear.

[1]Publicly available at: https://github.com/emorynlp/LAW-2022-Causal

A more appropriate way of understanding these classes is as products of various force vectors, as in Figure 1.

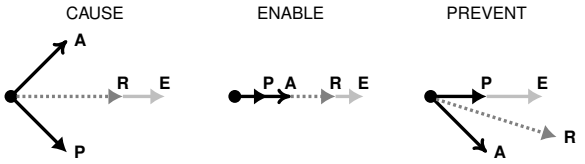

Figure 1: Representation of CAUSE, ENABLE, and PREVENT from Wolff (2007), where forces associated with the affector (A), forces associated with the patient (P) combine to form the resultant force (R) that may or may not be directed towards the endstate (E).

These vector diagrams represent the various forces at play in a causal relation. The patient is viewed as having a *tendency* for the endstate when the force associated with the patient is in the same direction as the endstate. Furthermore, the patient and affector act in *concordance* when the patient's force is in the same direction as the affector's force. The endstate may only *occur* when both the resultant's force and the force of the endstate are collinear. In PREVENT relations, the resultant force and the endstate are not collinear, and so the endstate that the patient tends toward does not occur. Understood as complex interactions of various factors, it is clear that there are numerous edge cases where affector and patient work more or less in concordance. As Wolff (2007) observes, people use qualitative assessments when deciding whether the resultant force could have been produced from the affector and patient forces. Accordingly, it would be unreasonable to ask annotators to consider complex vector operations when annotating text. With this in mind, two questions arise. Firstly, how can we enable annotators to resolve instances which lie at the edges of these categories? And secondly, how can we design intuitive guidelines to aid annotators in recognizing these relations, helping them identify the appropriate category when annotating causal language?

3. Related Research

Table 2 summarizes a number of influential studies on causal annotation. Among these works there are those in which annotations are performed manually (Mostafazadeh et al., 2016b; Caselli and Vossen, 2017; Dunietz et al., 2017a; Dunietz, 2018), those in which events are pre-identified (Mirza et al., 2014; Mirza and Tonelli, 2016; Caselli and Vossen, 2017), those in which additional temporal relations are annotated (Mirza et al., 2014; Mirza and Tonelli, 2016; Mostafazadeh et al., 2016b; Caselli and Vossen, 2017; Dunietz, 2018), as well as those that categorize the causal relation into the three CEP categories (Mirza et al., 2014; Mirza and Tonelli, 2016; Mostafazadeh et al., 2016b; Caselli and Vossen, 2017).

We identify three improvements that could be implemented in annotation schemes of causal relations. Firstly, most of the previous annotation schemes that aim to implement the CEP categories use simple counterfactual tests to discern between them. However, counterfactual reasoning by itself is often cognitively taxing and these rather simplistic counterfactual tests are not always ideal since, as mentioned in section 2, there are many edge cases which are hard to reason about. For example, consider the Causal and Temporal Relation Scheme's (CaTeRS) definitions of A CAUSE B, which is: *In the textual context, if A occurs, B most probably occurs as a result*, and A ENABLE B, which is: *In the textual context, if A does not occur, B most probably does not occur*. These definitions are concerned with only one facet of the CEP relations—namely, necessity and sufficiency. However, Wolff et al. (2005) does not define *necessity* as an attribute of ENABLE nor *sufficiency* for CAUSE or PREVENT. Not only are the notions of sufficiency and necessity a point of contention in literature (Lauer and Nadathur, 2020; Baglini and Siegal, 2020; Bar-Asher Siegal and Boneh, 2019), but these characteristics of CEP arguably arise as a byproduct of the core attributes of CAUSE, ENABLE, and PREVENT as shown in Figure 1.

Secondly, causal language encompasses a wide variety of lexical items. Much previous work in annotation of causal language ties causal meaning to a closed class of *triggers*. For example, the Penn Discourse Treebank's (PDTB) triggers are limited to conjunctions and adverbials, while PropBank limits its annotation of causal language to arguments of verbs. Furthermore, since the arguments of causal relations are usually taken to be events, as in Mostafazadeh et al. (2016b), some schemes do not annotate causal relations where only the agent in the Cause is specified. Thus, a richer representation of causal language enabled by a wide variety of identified triggers would improve the field's understanding of causal language.

Finally, the majority of causal annotation has been carried out on data from news sources. As such, there is a clear need for causal annotation of different genres and text types.

3.1. BECauSE

Of most relevance to the present study is the BECauSE corpus of causal relations developed in Dunietz et al. (2015), Dunietz et al. (2017b) and Dunietz (2018). The causal relations in this corpus are annotated based on pre-identified connectives between a Cause argument and an Effect argument listed in the Constructicon, a spreadsheet containing 191 pre-identified causal constructions and other relevant information. The causal relations are identified in 3x2 dimensions, including `Purpose`, `Motivation`, `Consequence` and `Facilitate` vs. `Inhibit`. However, he notes that the combination of both `Inhibit` and `Purpose` is not possible. Furthermore, since the identification choice between `Inhibit` and `Facilitate` relationships were pre-identified in Dunietz's Constructicon, the

Annotation scheme	Manual annotation	Pre-identified events	Temporal relations	Discourse relations	CEP
PDTB (Prasad et al., 2008; Prasad et al., 2006)	✓			✓	
PropBank (Kingsbury and Palmer, 2003; Bonial et al., 2014)	✓			✓	
Causal TempEval-3 (Mirza et al., 2014)		✓	✓		✓
CATENA (Mirza and Tonelli, 2016)		✓	✓		✓
CaTeRS (Mostafazadeh et al., 2016b)	✓		✓		✓
Storyline Extraction (Caselli and Vossen, 2017)	✓	✓	✓		✓
BECauSE 2.1 (Dunietz et al., 2017b; Dunietz, 2018)	✓		✓		*

* BECauSE uses `Facilitate` and `Inhibit`, where `Facilitate` maps onto CAUSE/ENABLE and `Inhibit` to PREVENT.

Table 2: Previous causal annotation schemes.

project's annotators' decision-making was constrained to the dimension of `Purpose`, `Motivation`, and `Consequence`. Notably, Dunietz expresses a desire to attempt more fine-grained distinctions based on Wolff et al. (2005)'s aforementioned CEP categories, although he is unable to achieve sufficiently stable inter-annotator agreement.

4. The CCEP Annotation Scheme

The Constructions of CAUSE, ENABLE, and PREVENT (CCEP) annotation scheme includes the annotation guidelines which utilizes the Constructicon as an annotation tool. Included in the annotation guidelines is a flowchart (named the Causal Relation Decision Tree abbreviated as CRDT, presented as Figure 2) designed to guide the annotators' decision process. These three components are adapted from Dunietz (2018).

In this section we describe the main features of both the Constructicon and the Annotation Scheme. Annotating instances of "causal language" within the CCEP scheme consists of labelling clauses or phrases which denote an event, state, action, or entity, the Cause, which is *explicitly presented as* promoting or hindering another, the Effect. The Cause and Effect must be textually connected through an explicit trigger, referred to as the "connective".

4.1. Parts of an annotatable causal instance

Annotation of an instance is prompted by the appearance of a causal connective, which can be related with up to three other spans of text of which any may be disjoint. Annotation spans are thus one of four types: (i) The Causal Connective which functions as the basis of all annotation instances and signifies the possibility of a causal construction (e.g. *for...to*, *because*), (ii) The Cause span which is generally an event or state involving an entity and is ideally expressed as a propositional clause or phrase, (iii) The Effect span which is also generally an event or state, ideally expressed as a propositional clause or phrase, and (iv) The Means span which includes an action that serves the purpose of differentiating between the agent of the Cause and the action by which that agent induces the Effect.

4.2. The Constructicon

Causal connectives are pre-identified in the Constructicon which is provided to annotators to actively use as they annotate. It is adapted from Dunietz (2018) with the addition of three causal connectives identified during annotation (*'due to'*, *'stop'*, and *'caused by'*). We also deleted six columns containing information which is not pertinent to the CEP classification task, including 'WordNet senses included', 'Type', 'Degree', 'Notable restrictions on type', 'Possible overlapping categories' (since these are only relevant with Dunietz's roles), and 'Number of distinct construction variants' (which was deemed unimportant for annotators). The Constructicon grounds the backbone of this scheme in Construction Grammar, meaning that *constructions* are taken as the fundamental units of language. On this account, constructions pair directly with meanings. As such, causal relations should be easily observable in specific lexical constructions, following the surface construction labeling approach. The Constructicon is provided as a searchable spreadsheet of 194 causal connective patterns, and was designed to minimize the decision-making burden placed on annotators. Examples of constructions include *for* <*Effect*> *to* <*Effect*>, <*Cause*> and <*Effect*> *because* <*Cause*>.

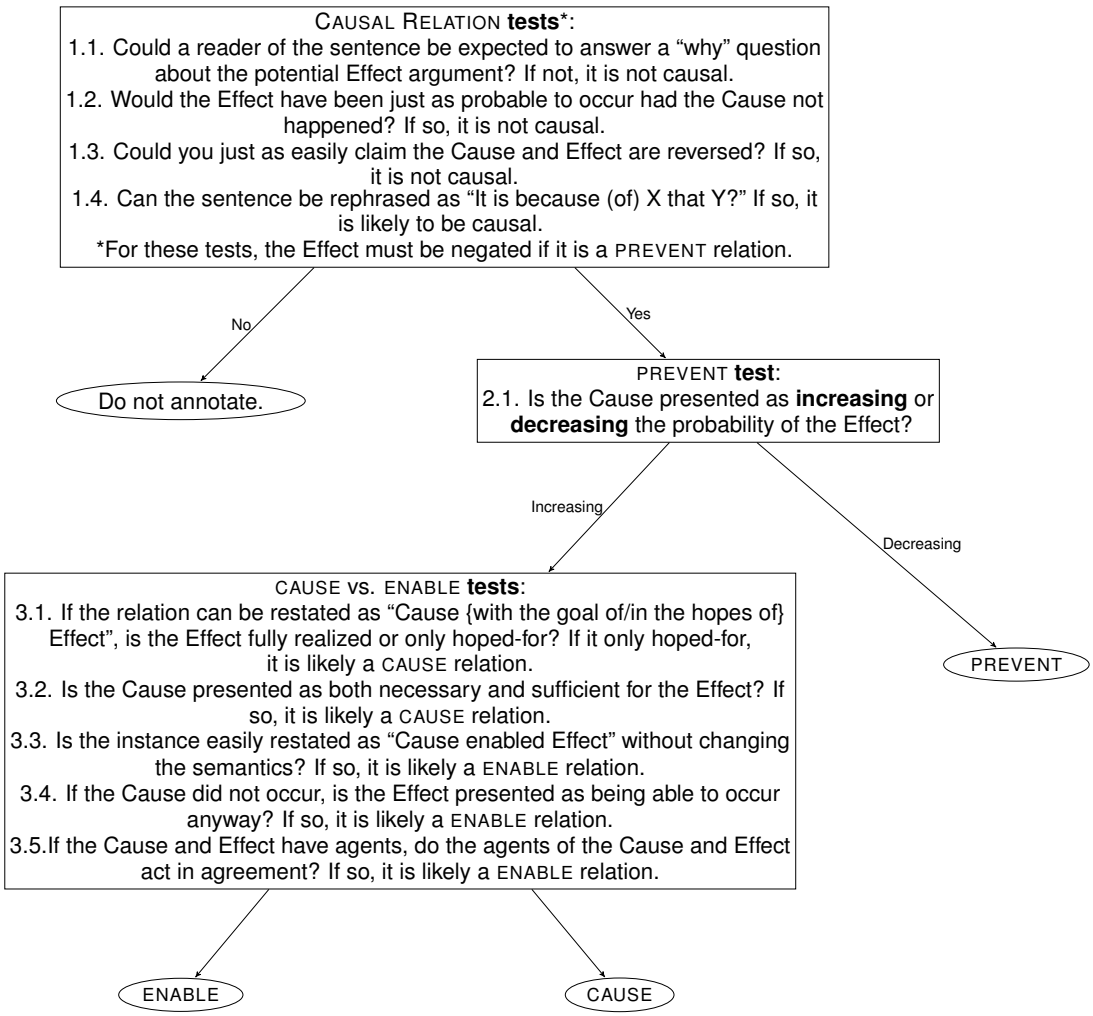

Figure 2: Decision tree for causation categorization (the CRDT).

4.3. Causation in CCEP

While Dunietz focuses on causal categories of `Purpose`, `Motivation`, and `Consequence`, as well as `Facilitate` and `Inhibit`, we aim to extend the applicability of his tools to categorize CAUSE, ENABLE, and PREVENT, which is a more nuanced exploration of his second dimension. Dunietz (2018) discusses a preliminary attempt to have a 3x3 categorization including CEP; unfortunately, he is unable to reach satisfactory IAA scores. His solution is to collapse CAUSE and ENABLE into `Facilitate`, leaving PREVENT to map to `Inhibit`, where in the 3x2 combination of possible relations, relations of both `Inhibit` and `Purpose`-types were not possible.

As discussed above, the CCEP scheme is built on the force dynamics model of causation from Wolff and Song (2003). Consequently, annotators are tasked with identifying causal relations as CAUSE, ENABLE, or PREVENT-type. Since the Constructicon specifies when a connective is PREVENT-type, the core task for annotators of the CCEP scheme is to distinguish between instances of CAUSE and ENABLE. To this end, we provide the following tests presented in the annotator's decision flow as depicted in the CRDT in Figure 2.

Test 3.1. If the relation can be restated as "⟨Cause⟩ {with the goal of / in the hopes of} ⟨Effect⟩", is the Effect fully realized or only hoped-for? If it is only hoped-for, it is likely a CAUSE relation.

Test 3.2. Is the Cause presented as both necessary and sufficient for the Effect? If so, it is likely a CAUSE relation.

Test 3.3. Is the instance easily restated as "⟨Cause⟩ enabled ⟨Effect⟩" without changing the meaning? If so, it is likely a ENABLE relation.

Test 3.4. If the Cause did not occur, is the Effect presented as being able to occur anyway? If so, it is likely a ENABLE relation.

Test 3.5. If the Cause and Effect have agents, do the agents of the Cause and Effect act in agreement? If so, it is likely an ENABLE relation.

These tests are ordered hierarchically, so passing test 3.1 holds more weight than passing test 3.5. However, tests are not necessarily definitive. For instance, if a relation does not pass test 3.1, this does not guarantee it is an

ENABLE relation. As such, annotators are instructed to work through each test and make a judgement that takes into account the greater weight of the earlier tests over the later tests.

Test 3.1 is intended to capture causal relations of purpose. Specifically, when an agent acts in a way to bring about a desired state of affairs, that desire causes the agent to act.

Test 3.2 reflects the fact that Causes of ENABLE are not sufficient alone for the Effect to occur given the patient tendency towards the endstate. Therefore, if the Cause is presented as necessary and sufficient, it must be a Cause of a CAUSE relation (by contraposition). For example, if the author writes, *'I failed the test only because the professor dislikes me'*, the span of *'the professor dislikes me'* is to be interpreted as the sole Cause, sufficient for bringing about the author's failure, and should thus be annotated as a CAUSE relation.

Test 3.3 is motivated by the observation that while not all instances of the use of lexical *cause* are of CAUSE-type (e.g., *'a cause of her death were her poor eating habits'*), uses of *enable* are generally of ENABLE-type.

Test 3.4. is grounded in similar reasoning to the point made for Test 3.3, but holds for cases where a force relevant to the causal relation is not captured within the span of the Cause or Effect, but may or may not be mentioned elsewhere in the document. If all relevant forces act toward the same endstate, it may be possible for one of the forces to compensate for the lack of an alternate force moving in the same direction.

Finally, test 3.5 is designed to determine the cases in which the affector and patient act in concordance, tracking Wolff's notion of ENABLE.

To conclude, these diagnostics aid in clarifying the vague notions of CEP for annotators in a way that sufficiently retains the original prototypical notions of CAUSE, ENABLE, and PREVENT characterized by Wolff and Song (2003).

5. Methodology

5.1. Data

The CCEP is a corpus of 150 documents (totalling 22,558 tokens) taken from three different sources: Aesops Fables[2], CNN newswire from the cnn_dailymail corpus[3], and Reddit posts taken from popular college subreddits[4]. Posts are filtered using the Profanity-Check Python library[5]. All data from these sources are tokenized using the ELIT Tokenizer[6] and then filtered to a length between 100 and 200 tokens.

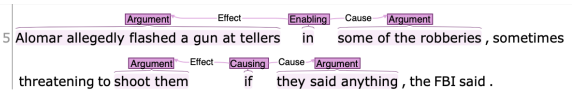

Figure 3: A sample annotation instance in INCEpTION.

5.2. Training

To guarantee that annotators understand the guidelines and meet a standard of performance, they undergo extensive training prior to undertaking annotation. The training consists of three stages: (i) annotators read the guidelines and view an instructional video, (ii) they take 10 online quizzes[7] consisting of 10 questions each on span identification, argument labelling, and relation labelling, and (iii) they must achieve a satisfactory inter-annotator agreement (IAA) score with gold-standard annotation of 10 practice documents. We began the training process with four annotators, consisting of three undergraduate students and a postdoctoral researcher who are all experienced annotators. Of these four, two progressed into the annotation process. Annotators are instructed to rotate through the various data sources in batches of 5 to ensure that any difference in IAA scores is not a result of familiarity with the annotation tool or experience following the annotation scheme.

5.3. Annotation Tool

Annotation was performed using the INCEpTION tool[8] (illustrated in Figure 3) developed by Technische Universität Darmstadt (Klie et al., 2018). This tool enabled the coordination of CCEP with two other parallel annotation projects in multiple layers including coreference and temporal relation annotation.

6. Results from the CCEP corpus

6.1. Inter-Annotator Agreement

We used F_1 to measure span agreement and Cohen's Kappa to measure causation type and argument labels in order to be able to compare our performance to Dunietz (2018)'s, as shown in Table 3. As demonstrated in Table 4, our overall corpus of causal annotations yields an F_1 score of 0.77 for connective identification, which is an improvement on the 0.70 of Dunietz (2018). Allowing for partial overlap, our F_1 score of 0.83 also improves upon Dunietz's 0.78. For agreed connective spans, the corpus also yielded a κ score of 0.83 for types of causation. This is similar to Dunietz's 0.80 for the causation categories of `Purpose`, `Motivation`, and `Consequence`. However, our argument span score of 0.71 was lower than Dunietz's at 0.86 (excluding overlap) and his 0.96 compared to our 0.86 including overlap. This was likely due to argument length disagreement, as all three document types contained very

[2] `https://www.gutenberg.org/cache/epub/21/pg21.txt`
[3] `https://huggingface.co/datasets/cnn_dailymail`
[4] `https://github.com/emorynlp/RedditData` accessed on 14th February 2022.
[5] `https://github.com/vzhou842/profanity-check`
[6] `https://github.com/emorynlp/elit-tokenizer`
[7] Training quizzes were created using Google Forms.
[8] `https://inception-project.github.io/`

Annotation scheme	Relation types	Arguments IAA	Arguments metric	Connectives IAA	Connectives metric	Relation IAA	Relation metric	Corpus size
PDTB	1	0.90* (Miltsakaki et al., 2004)	Percent	n/a	n/a	0.53† (Pitler et al., 2008)	F_1	2499 (news) (Prasad et al., 2019)
PropBank	1	0.93	Cohen's Kappa	0.93	Cohen's Kappa	0.91	Cohen's Kappa	2499 (news) (Palmer et al., 2005)
Causal TimeEval-3	3	n/a	n/a	0.55	F_1	0.3	F_1	20 (news)
CATENA	3	n/a	n/a	n/a	n/a	0.622	F_1	276 (news) (Pustejovsky et al., 2006) (Graff, 2002) (UzZaman et al., 2012)
CaTeRS	9**	0.91	Fleiss' Kappa	n/a	n/a	0.51	Fleiss' Kappa	320 (stories) (Mostafazadeh et al., 2016a)
StoryLine Extraction	2	n/a	n/a	n/a	n/a	0.638	Dice Coefficient	258 (news)
BECauSE 2.1	5	0.86‡	F_1	0.70	F_1	0.80	Cohen's Kappa	>116 (news) (Sandhaus, 2008) (Marcus et al., 1994) (Ide et al., 2010) (Smith et al., 2014)

* Calculated for 3103 tokens. † Only for CONTINGENCY relations. ** Only 4 of 9 are causal. ‡ Spans only.

Table 3: Results from previous causal annotation studies.

different writing styles, ranging from the wordy, rant-like style of Reddit documents to more succinct news reporting.

	Reddit	News	Fables	Overall
Connective spans (F_1)	0.82	0.75	0.75	0.77
Connectives + overlap (F_1)	0.86	0.81	0.81	0.83
Types of causation (κ)	0.78	0.89	0.82	0.83
Argument spans (F_1)	0.76	0.72	0.68	0.71
Arguments + overlap (F_1)	0.91	0.83	0.85	0.86
Argument labels (κ)	0.93	0.86	0.91	0.90

Table 4: Annotation performance across different text types, with and without partial overlap for span identification. κ = Cohen's Kappa.

Since the main obstacle faced by the present study is to provide a means of establishing agreement on instances of vague CEP categories—and specifically distinguishing between CAUSE and ENABLE—we provide the percentage of how often annotators agreed on the CAUSE and ENABLE labels in Table 5. These scores demonstrate that annotators were able to reliably differentiate between these categories across different document types.

	CAUSE vs. ENABLE agreement
Reddit	78.57%
News	89.25%
Fables	80.95%
Overall	82.48%

Table 5: Percentage of agreement in cause type between CAUSE and ENABLE across the various genres.

Finally, we perform a one-way ANOVA comparing overall F_1 scores across genres for all documents, which yields a p-value of 0.29 showing no significant effect of data type on IAA. This demonstrates the robustness of our guidelines across genres, which included specific instructions for genre-specific idiosyncrasies such as the appearances of abbreviations and shorthands in Reddit posts.

6.2. Statistics

The analysis of our corpus provides numerous interesting insights. The corpus contains a total of 150 doubly-annotated documents, which featured 870 annotations of causal constructions between both annotators, with 22 of our 300 annotated documents containing no causal annotation at all. As shown in Figure 4, CAUSE-type instances dominated all instances of annotated causal language. This was to be expected since test 3.2 of the CRDT tests for CAUSE-type instances asks annotators whether the textual context presents the Cause as necessary and sufficient for the Effect. In the limited context of a 200-token document, many authors present the Cause as contextually necessary and sufficient in some way for the Effect to occur.

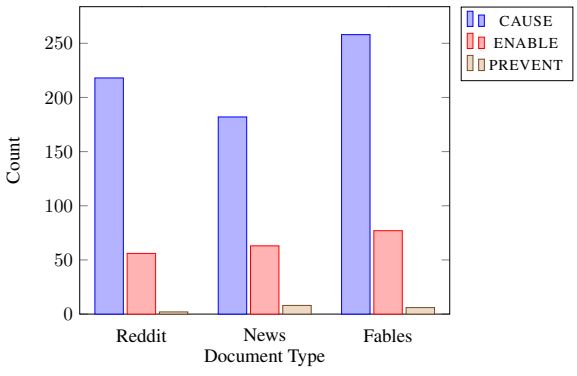

Figure 4: Counts of CEP across document types.

Table 6 is also of interest because it demonstrates that Fables had the most annotations of causal language, while News contained the least. We hypothesize that this is because of the narrative, event-driven structure of Fables, which have been popularly used for temporal

annotations for this reason (Bethard et al., 2012). The same reasoning may explain the less frequent use of causal relations in news data—news articles are more concerned with reporting states of affairs than making attributions of causality.

	Reddit		News		Fables		Total
Category	n	%	n	%	n	%	n
CAUSE	218	79	182	71.9	258	75.7	658
ENABLE	56	20.3	63	24.9	77	22.5	196
PREVENT	2	0.7	8	3.2	6	1.8	16
Total	276	100	253	100	341	100	870

Table 6: Counts of CEP across document types.

Table 7 reports the most popular connectives across the different document types. Firstly, note that the most frequent five connectives account for approximately half of all instances of annotated causal language. While our findings generally align with Dunietz's counts of connective patterns in the BECauSE corpus (our most frequent five appear in his top seven), it is interesting to note that their frequencies vary across document type. For example, the conditional only appears 8 times in the CNN news data, highlighting the factual nature of news reporting. Furthermore, while *'after'* appears as our fourth most popular connective pattern, these instance occur almost exclusively in the CNN data (with 41 counts, compared to only 4 in Reddit and 6 in Fables). Similarly, while *'because'* occurs in the top five most frequently appearing connectives, 77.8% of these appearances were in Reddit. This is most likely due to the stream-of-consciousness style of Reddit writing, where writers are not so concerned with diversifying their word choice. Finally, Table 8 lists the connectives that were used exclusively for either CAUSE or ENABLE throughout the entire corpus. While some pairings seem intuitive (e.g., *'let'* and *'allow'* denoting ENABLE relations), others are less so (e.g., *'with'* denoting CAUSE relations).

6.3. Summary of findings

In summary, this project reached IAA scores of $F_1 = 0.77$ for connective spans, $\kappa = 0.83$ for causation categorization of connectives, $F_1 = 0.71$ for argument spans, and $\kappa = 0.90$ for argument labels. Also observe that allowing for partial overlap only increases connective identification F_1 from 0.82 to 0.86, while argument identification improves from 0.71 to 0.86. This is to be expected, since connective spans are pre-delimited in the Constructicon for annotators, while argument spans are not. Furthermore, the most frequently annotated connectives in our corpus aligned with those in the BECauSE corpus. The sub-corpus of Fables contained the most occurrences of causal language, while News had the least. Finally, analysis of the connectives and their types across different sub-corpora reveal some interesting trends, such as connectives that appear frequently in one document type but not another, or connectives that only appear as CAUSE or ENABLE.

7. Discussion

A limitation of the surface construction labeling approach is its inability to represent long-distant, document-level causal relations. Consider the following text taken from one of the Reddit posts: *'I'm pretty much being called a liar and a cheat. Happened to anyone else? So, I literally cried when my TA told me.'* Intuitively, the accusation of plagiarism described in the first sentence could be construed as a Cause of the narrator *'literally crying'*. However, this causal relation is not annotatable according to our guidelines because (i) it is not demarcated by a lexical connective, and (ii) even with the connective *'so'* before *'I literally cried...'*, the span is not enough to fit into the construction of <*Cause*>*, so* <*Effect*> as the left argument of *'so'* is not the accusation of plagiarism.

A potential direction for future researchers may be to annotate a wider, more varied datasets when choosing text to annotate. While the straightforward and clean language used in news and short stories may enable higher IAA, using noisy data such as Reddit posts test the robustness of annotation schemes.

Finally, the IAA of our project demonstrates the feasibility of using CEP categorization in causal relation annotation. However, we did not include Dunietz's other causal dimensions of `Motivation`, `Purpose`, and `Consequence`. Thus, a natural next step in future research would be to integrate these aforementioned three categories and CEP into a single scheme. This expansion of dimensions annotated in the same layer would provide more insight into how causal relations are described in text.

8. Conclusion

In this paper, we introduced a decision based method for annotating causal categories across various genres of text. Our annotation scheme was designed to capture the categories of CAUSE, ENABLE, and PREVENT, and their many edge cases which are difficult for annotators to consistently identify in practice. We showed that, by using this method, annotators can achieve IAA which is comparable to previous studies. Furthermore, our method performs equally well across genres, highlighting the robustness of our annotation scheme. Finally, we observed a number of interesting differences in usage and frequency of causal language across different genres.

9. Acknowledgements

We gratefully acknowledge the support of the Amazon Alexa AI grant. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Alexa AI. We also thank Yingying Chen, Yuxin (Jessica) Ji, Claire Fenton, and Yifeng Wu for feedback on the annotation guidelines.

Causal	Reddit		News		Fables			
Connective	n	Frequency	n	Frequency	n	Frequency	Total	Overall %
to	48	17.39%	24	9.49%	46	13.49%	118	13.56%
for	29	10.51%	30	11.86%	42	12.32%	101	11.61%
if	30	10.87%	8	3.16%	47	13.78%	85	9.77%
after	4	1.45%	41	16.21%	2	0.59%	47	5.40%
because	35	12.68%	4	1.58%	6	1.76%	45	5.17%
Total	146	52.90%	107	42.30%	143	41.94%	396	45.52%

Table 7: Comparison of popular connectives across different document types.

Causal Connective	Type	Reddit	News	Fables	Total
make	CAUSE	6	8	15	29
with	CAUSE	4	4	10	18
cause	CAUSE	4	6	0	10
let	ENABLE	0	0	6	6
allow	ENABLE	2	3	0	5
have	CAUSE	0	2	3	5

Table 8: Count of connectives annotated exclusively as either CAUSE or ENABLE and $n \geq 5$.

10. Bibliographical References

Baglini, R. and Siegal, E. A. B.-A. (2020). Direct causation: A new approach to an old question. *University of Pennsylvania Working Papers in Linguistics*, 26:19–28.

Bar-Asher Siegal, E. and Boneh, N. (2019). Sufficient and necessary conditions for a non-unified analysis of causation. *Proceedings of the 36th West Coast Conference on Formal Linguistics*, pages 55–60.

Bethard, S., Kolomiyets, O., and Moens, M.-F. (2012). Annotating story timelines as temporal dependency structures. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2721–2726, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Bonial, C., Bonn, J., Conger, K., Hwang, J. D., and Palmer, M. (2014). PropBank: Semantics of new predicate types. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3013–3019. European Language Resources Association (ELRA).

Caselli, T. and Vossen, P. (2017). The event StoryLine corpus: A new benchmark for causal and temporal relation extraction. In *Proceedings of the Events and Stories in the News Workshop*, pages 77–86, Vancouver, Canada, August. Association for Computational Linguistics.

Dunietz, J., Levin, L., and Carbonell, J. G. (2015). Annotating causal language using corpus lexicography of constructions. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 188–196.

Dunietz, J., Levin, L., and Carbonell, J. (2017a). Automatically Tagging Constructions of Causation and Their Slot-Fillers. *Transactions of the Association for Computational Linguistics*, 5:117–133, 06.

Dunietz, J., Levin, L., and Carbonell, J. (2017b). The BECauSE corpus 2.0: Annotating causality and overlapping relations. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 95–104, Valencia, Spain, April. Association for Computational Linguistics.

Dunietz, J. (2018). *Annotating and Automatically Tagging Constructions of Causal Language*. Ph.D. thesis, Carnegie Mellon University.

Graff, D. (2002). *The AQUAINT Corpus of English News Text*. 09.

Gundle, K., Dingel, M., and Koenig, B. (2010). "to prove this is the industry's best hope": Big tobacco's support of research on the genetics of nicotine addiction. *Addiction*, 105(6):974–983.

Ide, N., Baker, C., Fellbaum, C., and Passonneau, R. (2010). The manually annotated sub-corpus: A community resource for and by the people. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 68–73, Uppsala, Sweden, July. Association for Computational Linguistics.

Kingsbury, P. R. and Palmer, M. (2003). Propbank: the next level of treebank.

Klie, J.-C., Bugert, M., Boullosa, B., Eckart de Castilho, R., and Gurevych, I. (2018). The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico, August. Association for Computational Linguistics.

Lauer, S. and Nadathur, P. (2020). Causal necessity, causal sufficiency, and the implications of causative verbs. *Glossa: a journal of general linguistics*, 5:49–105.

Marcus, M., Kim, G., Marcinkiewicz, M. A., MacIntyre, R., Bies, A., Ferguson, M., Katz, K., and Schasberger, B. (1994). The Penn Treebank: Annotating predicate argument structure. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Miltsakaki, E., Prasad, R., Joshi, A., and Webber, B. (2004). The Penn Discourse Treebank. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May. European Language Resources Association (ELRA).

Mirza, P. and Tonelli, S. (2016). CATENA: CAusal and TEmporal relation extraction from NAtural language texts. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 64–75, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Mirza, P., Sprugnoli, R., Tonelli, S., and Speranza, M. (2014). Annotating causality in the TempEval-3 corpus. In *Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL)*, pages 10–19, Gothenburg, Sweden, April. Association for Computational Linguistics.

Mostafazadeh, N., Chambers, N., He, X., Parikh, D., Batra, D., Vanderwende, L., Kohli, P., and Allen, J. (2016a). A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California, June. Association for Computational Linguistics.

Mostafazadeh, N., Grealish, A., Chambers, N., Allen, J., and Vanderwende, L. (2016b). CaTeRS: Causal and temporal relation scheme for semantic annotation of event structures. In *Proceedings of the Fourth Workshop on Events*, pages 51–61, San Diego, California, June. Association for Computational Linguistics.

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

Pearl, J. (2009). *Causality*. Cambridge University Press, Cambridge, UK, 2 edition.

Pitler, E., Raghupathy, M., Mehta, H., Nenkova, A., Lee, A., and Joshi, A. (2008). Easily identifiable discourse relations. In *Coling 2008: Companion volume: Posters*, pages 87–90, Manchester, UK, August. Coling 2008 Organizing Committee.

Prasad, R., Miltsakaki, E., Dinesh, N., Lee, A., Joshi, A. K., Robaldo, L., and Webber, B. L. (2006). The penn discourse treebank 2.0 annotation manual.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008). The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA).

Prasad, R., Webber, B., Lee, A., and Joshi, A. (2019). Penn Discourse Treebank Version 3.0.

Pustejovsky, J., Verhagen, M., Saurí, R., Moszkowicz, J., Gaizauskas, R., Katz, G., Mani, I., Knippen, R., and Setzer, A. (2006). *TimeBank 1.2*. 01.

Sandhaus, E. (2008). The New York Times Annotated Corpus.

Smith, N. A., Cardie, C., Washington, A., and Wilkerson, J. (2014). Overview of the 2014 NLP unshared task in PoliInformatics. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 5–7, Baltimore, MD, USA, June. Association for Computational Linguistics.

UzZaman, N., Llorens, H., Allen, J., Derczynski, L., Verhagen, M., and Pustejovsky, J. (2012). Tempeval-3: Evaluating events, time expressions, and temporal relations.

Whorf, B. L. (1956). *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf*. MIT Press.

Wolff, P. and Song, G. (2003). Models of causation and causal verbs. *Cognitive Psychology*, 47:276–332.

Wolff, P. and Thorstad, R. (2017). Force dynamics. *The Oxford handbook of causal reasoning*, pages 147–168.

Wolff, P. and Zettergren, M. (2002). A vector model of causal meaning. In *Proceedings of the twenty-fifth annual conference of the cognitive science society*. Erlbaum.

Wolff, P., Klettke, B., Ventura, T., and Song, G. (2005). Expressing causation in english and other languages.

Wolff, P. (2007). Representing causation. *Journal of experimental psychology. General*, 136:82–111, 03.