The BECauSE Corpus 2.0: Annotating Causality and Overlapping Relations

Jesse Dunietz
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA 15213, USA
jdunietz@cs.cmu.edu

Lori Levin and **Jaime Carbonell**
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{lsl,jgc}@cs.cmu.edu

Abstract

Language of cause and effect captures an essential component of the semantics of a text. However, causal language is also intertwined with other semantic relations, such as temporal precedence and correlation. This makes it difficult to determine when causation is the primary intended meaning. This paper presents BECauSE 2.0, a new version of the BECauSE corpus with exhaustively annotated expressions of causal language, but also seven semantic relations that are frequently co-present with causation. The new corpus shows high inter-annotator agreement, and yields insights both about the linguistic expressions of causation and about the process of annotating co-present semantic relations.

1 Introduction

We understand our world in terms of causal networks – phenomena causing, enabling, or preventing others. Accordingly, the language we use is full of references to cause and effect. In the Penn Discourse Treebank (PDTB; Prasad et al., 2008), for example, over 12% of explicit discourse connectives are marked as causal, as are nearly 26% of implicit discourse relationships. Recognizing causal assertions is thus invaluable for semantics-oriented applications, particularly in domains such as finance and biology where interpreting these assertions can help drive decision-making.

In addition to being ubiquitous, causation is often co-present with related meanings such as temporal order (cause precedes effect) and hypotheticals (the *if* causes the *then*). This paper presents the Bank of Effects and Causes Stated Explicitly (BECauSE) 2.0, which offers insight into these overlaps. As in BECauSE 1.0 (Dunietz et al., 2015, in press), the corpus contains annotations for causal language. It also includes annotations for seven commonly co-present meanings when they are expressed using constructions shared with causality.

To deal with the wide variation in linguistic expressions of causation (see Neeleman and Van de Koot, 2012; Dunietz et al., 2015), BECauSE draws on the principles of Construction Grammar (CxG; Fillmore et al., 1988; Goldberg, 1995). CxG posits that the fundamental units of language are *constructions* – pairings of meanings with arbitrarily simple or complex linguistic forms, from morphemes to structured lexico-syntactic patterns.

Accordingly, BECauSE admits arbitrary constructions as the bearers of causal relationships. As long as there is at least one fixed word, any conventionalized expression of causation can be annotated. By focusing on causal *language* – conventionalized expressions of causation – rather than real-world causation, BECauSE largely sidesteps the philosophical question of what is truly causal. It is not concerned, for instance, with whether there is a real-world causal relationship within *flu virus* (virus causes flu) or *delicious bacon pizza* (bacon causes deliciousness); neither is annotated.

Nonetheless, some of the same overlaps and ambiguities that make real-world causation so hard to circumscribe seep into the linguistic domain, as well. Consider the following examples (with **causal constructions** in bold, CAUSES in small caps, and *effects* in italics):

(1) **After** I DRANK SOME WATER, *I felt much better*.
(2) **As** VOTERS GET TO KNOW MR. ROMNEY BETTER, *his poll numbers will rise*.
(3) **THE MORE** HE COMPLAINED, ***the less*** *his captors fed him*.
(4) THE RUN ON BEAR STERNS **created** *a crisis*.
(5) THE IRAQI GOVERNMENT will **let** *Represen-*

Proceedings of the 11th Linguistic Annotation Workshop, pages 95–104,
Valencia, Spain, April 3, 2017.

tative Hall visit next week.

Each sentence conveys a causal relation, but piggybacks it on a related relation type. (1) uses a temporal relationship to suggest causality. (3) employs a correlative construction, and (2) contains elements of both time and correlation in addition to causation. (4), meanwhile, is framed as bringing something into existence, and (5) suggests both permission and enablement.

Most semantic annotation schemes have required that each token be assigned just one meaning. BECauSE 1.0 followed this policy, as well, but this resulted in inconsistent handling of cases like those above. For example, the meaning of *let* varies from "allow to happen" (clearly causal) to "verbalize permission" (not causal) to shades of both. These overlaps made it difficult for annotators to decide when to annotate such cases as causal.

The contributions of this paper are threefold. First, we present a new version of the BECauSE corpus, which offers several improvements over the original. Most importantly, the updated corpus includes annotations for seven different relation types that overlap with causality: temporal, correlation, hypothetical, obligation/permission, creation/termination, extremity/sufficiency, and context. Overlapping relations are tagged for any construction that can also be used to express a causal relationship. The improved scheme yields high inter-annotator agreement. Second, using the new corpus, we derive intriguing evidence about how meanings compete for linguistic machinery. Finally, we discuss the issues that the annotation approach does and does not solve. Our observations suggest lessons for future annotation projects in semantic domains with fuzzy boundaries between categories.

2 Related Work

Several annotation schemes have addressed elements of causal language. Verb resources such as VerbNet (Schuler, 2005) and PropBank (Palmer et al., 2005) include verbs of causation. Likewise, preposition schemes (e.g., Schneider et al., 2015, 2016) include some purpose- and explanation-related senses. None of these, however, unifies all linguistic realizations of causation into one framework; they are concerned with specific classes of words, rather than the semantics of causality.

FrameNet (Ruppenhofer et al., 2016) is closer in spirit to BECauSE, in that it starts from meanings and catalogs/annotates a wide variety of lexical items that can express those meanings. Our work differs in several ways. First, FrameNet represents causal relationships through a variety of unrelated frames (e.g., CAUSATION and THWARTING) and frame roles (e.g., PURPOSE and EXPLANATION). As with other schemes, this makes it difficult to treat causality in a uniform way. (The ASFALDA French FrameNet project recently proposed a reorganized frame hierarchy for causality, along with more complete coverage of French causal lexical units [Vieu et al., 2016]. Merging their framework into mainline FrameNet would mitigate this issue.) Second, FrameNet does not allow a lexical unit to evoke more than one frame at a time (although SALSA [Burchardt et al., 2006], the German FrameNet, does allow this).

The Penn Discourse Treebank includes causality under its hierarchy of contingency relations. Notably, PDTB does allow annotators to mark discourse relations as both causal and something else. However, it is restricted to discourse relations; it excludes other realizations of causal relationships (e.g., verbs and many prepositions), as well as PURPOSE relations, which are not expressed as discourse connectives. BECauSE 2.0 can be thought of as an adaptation of PDTB's multiple-annotation approach. Instead of focusing on a particular type of construction (discourse relations) and annotating all the meanings it can convey, we start from a particular meaning (causality), find all constructions that express it, and annotate each instance in the text with all the meanings it expresses.

Other projects have attempted to address causality more narrowly. For example, a small corpus of event pairs conjoined with *and* has been tagged as causal or not causal (Bethard et al., 2008). The CaTeRS annotation scheme (Mostafazadeh et al., 2016), based on TimeML, also includes causal relations, but from a commonsense reasoning standpoint rather than a linguistic one. Similarly, Richer Event Description (O'Gorman et al., 2016) integrates real-world temporal and causal relations between events into a unified framework. A broader-coverage linguistic approach was taken by Mirza and Tonelli (2014), who enriched TimeML to include causal links and their lexical triggers. Their work differs from ours in that it requires arguments to be TimeML events; it requires causal connectives to be contiguous; and its guidelines define causality less precisely, relying on intuitive notions

of causing, preventing, and enabling.

2.1 BECauSE 1.0

Our work is of course most closely based on BECauSE 1.0. Its underlying philosophy is to annotate any form of *causal language* – conventionalized linguistic mechanisms used to appeal to cause and effect. Thus, the scheme is not concerned with what real-world causal relationships hold, but rather with what relationships are presented in the text. It defines causal language as "any construction which presents one event, state, action, or entity as promoting or hindering another, and which includes at least one lexical trigger." Each annotation consists of a **cause** span; an **effect** span; and a **causal connective**, the possibly discontinuous lexical items that express the causal relationship (e.g., *because of* or *opens the way for*).

3 Extensions and Refactored Guidelines in BECauSE 2.0

This update to BECauSE improves on the original in several ways. Most importantly, as mentioned above, the original scheme precluded multiple copresent relations. Tagging a connective as causal was taken to mean that it was primarily expressing causation, and not temporal sequence or permission. (In fact, temporal expressions that were intended to suggest causation were explicitly excluded.) Based on the new annotations, there were 210 instances in the original corpus where multiple relations were present and annotators had to make an arbitrary decision.[1] The new scheme extends the previous one to include these overlapping relations.

Second, although the first version neatly handled many different kinds of connectives, adjectives and nouns were treated in a less general way. Verbs, adverbs, conjunctions, prepositions, and complex constructions typically have two natural slots in the construction. For example, the *because* construction can be schematized as ⟨*effect*⟩ *because* ⟨*cause*⟩, and the causative construction present in *so loud it hurt* as ⟨*so cause*⟩ *(that)* ⟨*effect*⟩.

Adjective and noun connectives, however, do not offer such natural positions for ⟨cause⟩ and ⟨effect⟩. In the following example, BECauSE 1.0 would annotate the connective as marked in bold: *the **cause of** her illness **was** dehydration*. But this is an unparsimonious account of the causal construction: the copula and preposition do not contribute to the causal meaning, and other language could be used to tie the connective to the arguments. For instance, it would be equally valid to say *her illness' cause was dehydration*, or even *the chart listed her illness' cause as dehydration*. The new corpus addresses this by annotating just the noun or adjective as the connective (e.g., *cause*), and letting the remaining argument realization language vary. A number of connectives were similarly refactored to make them simpler and more consistent.

Finally, version 1.0 struggled with the distinction between the causing event and the causing agent. For example, in *I caused a commotion by shattering a glass*, either the agent (*I*) or the agent's action (*shattering a glass*) could plausibly be annotated as the cause. The guidelines for version 1.0 suggested that the true cause is the action, so the agent should be annotated as the cause only when no action is described. (In such cases, the agent would be considered metonymic for the action.) However, given the scheme's focus on constructions, it seems odd to say that the arguments to the construction change when a *by* clause is added.

The new scheme solves this by labeling the agent as the cause in both cases, but adding a MEANS argument for cases where both an agent and their action are specified.[2]

4 BECauSE 2.0 Annotation Scheme

4.1 Basic Features of Annotations

The second version of the BECauSE corpus retains the philosophy and most of the provisions of the first, with the aforementioned changes.

To circumscribe the scope of the annotations, we follow BECauSE 1.0 in excluding causal relationships with no lexical trigger (e.g., *He left. He wasn't feeling well.*); connectives that lexicalize the means or result of the causation (e.g., *kill* or *convince*); and connectives that underspecify the nature of the causal relationship (e.g., *linked to*).

[1] This is the total number, in the new corpus, of instances that are annotated with both causal and overlapping relations and which would have been ambiguous under the 1.0 guidelines – i.e., the guidelines did not either explictly exclude them or deem them always causal.

[2] Another possibility would have been to divvy up causes into CAUSE and AGENT arguments. Although FrameNet follows this route in some of its frames, we found that this distinction was difficult to make in practice. For example, a non-agentive cause might still be presented with a separate means clause, as in *inflammation triggers depression by altering immune responses*. In contrast, MEANS are relatively easy to identify when present, and tend to exhibit more consistent behavior with respect to what constructions introduce them.

As in BECauSE 1.0, the centerpiece of each instance of causal language is the **causal connective**. The connective is not synonymous with the causal construction; rather, it is a lexical proxy indicating the presence of the construction. It consists of all words present in every use of the construction. For example, the bolded words in ***enough*** *money* ***for*** *us* ***to*** *get by* would be marked as the connective. Annotators' choices of what to include as connectives were guided by a *constructicon*, a catalog of constructions specified to a human-interpretable level of precision (but not precise enough to be machine-interpretable). The constructicon was updated as needed throughout the annotation process.

In addition to the connective, each instance includes **cause** and **effect** spans. (Either the cause or the effect may be absent, as in a passive or infinitive.) BECauSE 2.0 also introduces the **means** argument, as mentioned above. Means arguments are annotated when an agent is given as the cause, but the action taken by that agent is also explicitly described, or would be but for a passive or infinitive. They are marked only when expressed as a *by* or *via* clause, a dependent clause (e.g., *Singing loudly, she caused winces all down the street*), or a handful of other conventional devices. If any of an instance's arguments consists of a bare pronoun, including a relativizing pronoun such as *that*, a coreference link is added back to its antecedent (assuming there is one in the same sentence).

The new scheme distinguishes three types of causation, each of which has slightly different semantics: CONSEQUENCE, in which the cause naturally leads to the effect; MOTIVATION, in which some agent perceives the cause, and therefore consciously thinks, feels, or chooses something; and PURPOSE, in which an agent chooses the effect because they desire to make the cause true. Unlike BECauSE 1.0, the new scheme does not include evidentiary uses of causal language, such as *She met him previously, because she recognized him yesterday*. These were formerly tagged as INFERENCE. We eliminated them because unlike other categories of causation, they are not strictly causal, and unlike other overlapping relations, they never also express true causation; they constitute a different sense of *because*.

The scheme also distinguishes positive causation (FACILITATE) from inhibitory causation (INHIBIT); see Dunietz et al. (2015) for full details.

Examples demonstrating all of these categories are shown in Table 1.

4.2 Annotating Overlapping Relations

The constructions used to express causation overlap with many other semantic domains. For example, the *if/then* language of hypotheticals and the *so* ⟨*adjective*⟩ construction of extremity have become conventionalized ways of expressing causation, usually in addition to their other meanings. In this corpus, we annotate the presence of these overlapping relations, as well.

A connective is annotated an instance of either causal language or a non-causal overlapping relation whenever it is used in a sense and construction that *can* carry causal meaning. The operational test for this is whether the word sense and linguistic structure allow it to be coerced into a causal interpretation, and the meaning is either causal or one of the relation types below.

Consider, for example, the connective *without*. It is annotated in cases like *without your support, the campaign will fail*. However, annotators ignored uses like *we left without saying goodbye*, because in this linguistic context, *without* cannot be coerced into a causal meaning. Likewise, we include *if* as a HYPOTHETICAL connective, but not *suppose that*, because the latter cannot indicate causality.

All overlapping relations are understood to hold between an ARGC and an ARGE. When annotating a causal instance, ARGC and ARGE refer to the cause and effect, respectively. When annotating a non-causal instance, ARGC and ARGE refer to the arguments that would be cause and effect if the instance were causal. For example, in a TEMPORAL relation, ARGC would be the earlier argument and ARGE would be the later one.

The following overlapping relation types are annotated:

- TEMPORAL: when the causal construction explicitly foregrounds a temporal order between two arguments (e.g., *once*, *after*) or simultaneity (e.g., *as*, *during*).
- CORRELATION: when the core meaning of the causal construction is that ARGC and ARGE vary together (e.g., *as*, *the more… the more…*).
- HYPOTHETICAL: when the causal construction explicitly imagines that a questionable premise is true, then establishes what would hold in the world where it is (e.g., *if… then…*).
- OBLIGATION/PERMISSION: when ARGE (effect) is an agent's action, and ARGC (cause)

	FACILITATE	INHIBIT
CONSEQUENCE	*We are in serious economic trouble* **because of** INADEQUATE REGULATION.	THE NEW REGULATIONS should **prevent** *future crises*.
MOTIVATION	WE DON'T HAVE MUCH TIME, **so** *let's move quickly*.	THE COLD **kept** *me* **from** *going outside*.
PURPOSE	*Coach them in handling complaints* **so that** THEY CAN RESOLVE PROBLEMS IMMEDIATELY.	(Not possible)

Table 1: Examples of every allowed combination of the three types of causal language and the two degrees of causation (with **connectives** in bold, CAUSES in small caps, and *effects* in italics).

is presented as some norm, rule, or entity with power that is requiring, permitting, or forbidding ARGE to be performed (e.g., *require* in the legal sense, *permit*).

- CREATION/TERMINATION: when the construction frames the relationship as an entity or circumstance being brought into existence or terminated (e.g., *generate*, *eliminate*).
- EXTREMITY/SUFFICIENCY: when the causal construction also expresses an extreme or sufficient/insufficient position of some value on a scale (e.g., *so... that...*, *sufficient... to...*).
- CONTEXT: when the construction clarifies the conditions under which the effect occurs (e.g,. *with*, *without*, *when* in non-temporal uses). For instance, *With supplies running low, we didn't even make a fire that night.*

All relation types present in the instance are marked. For example, *so offensive that I left* would be annotated as both causal (MOTIVATION) and EXTREMITY/SUFFICIENCY. When causality is not present in a use of a sometimes-causal construction, the instance is annotated as NON-CAUSAL, and the overlapping relations present are marked.

It can be difficult to determine when language that expresses one of these relationships was also intended to convey a causal relationship. Annotators used a variety of questions to assess an ambiguous instance, largely based on Grivaz (2010):

- **The "why" test:** After reading the sentence, could a reader reasonably be expected to answer a "why" question about the potential effect argument? If not, it is not causal.
- **The temporal order test:** Is the cause asserted to precede the effect? If not, it is not causal.
- **The counterfactuality test:** Would the effect have been just as probable to occur or not occur had the cause not happened? If so, it is not causal.
- **The ontological asymmetry test:** Could you just as easily claim the cause and effect are reversed? If so, it is not causal.
- **The linguistic test:** Can the sentence be rephrased as "It is because (of) X that Y" or "X causes Y?" If so, it is likely to be causal.

Figure 1 showcases several fully-annotated sentences that highlight the key features of the new BECauSE scheme, including examples of overlapping relations.

5 BECauSE 2.0 Corpus

5.1 Data

The BECauSE 2.0 corpus[3] is an expanded version of the dataset from BECauSE 1.0. It consists of:

- 59 randomly selected articles from the year 2007 in the Washington section of the New York Times corpus (Sandhaus, 2008)
- 47 documents randomly selected[4] from sections 2-23 of the Penn Treebank (Marcus et al., 1994)
- 679 sentences[5] transcribed from Congress' Dodd-Frank hearings, taken from the NLP Unshared Task in PoliInformatics 2014 (Smith et al., 2014)
- 10 newspaper documents (Wall Street Journal and New York Times articles, totalling 547 sentences) and 2 journal documents (82 sentences) from the Manually Annotated Sub-Corpus (MASC; Ide et al., 2010)

The first three sets of documents are the same dataset that was annotated for BECauSE 1.0.

5.2 Inter-Annotator Agreement

Inter-annotator agreement was calculated between the two primary annotators on a sample of 260

[3]Publicly available, along with the constructicon, at `https://github.com/duncanka/BECauSE`.

[4]We excluded WSJ documents that were either earnings reports or corporate leadership/structure announcements, as both tended to be merely short lists of names/numbers.

[5]The remainder of the document was not annotated due to constraints on available annotation effort.

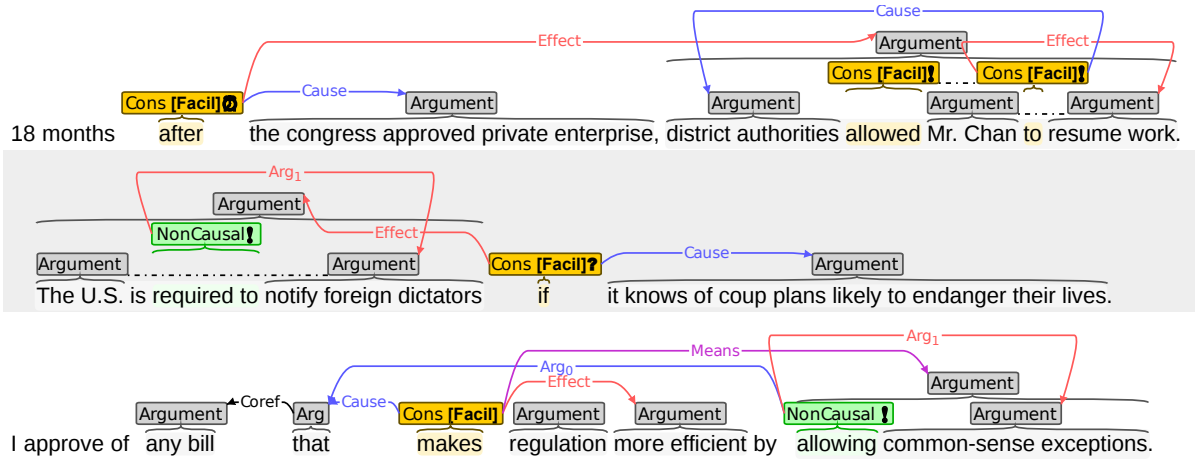

Figure 1: Several example sentences annotated in BRAT (Stenetorp et al., 2012). The question mark indicates a hypothetical, the clock symbol indicates a temporal relation, and the thick exclamation point indicates obligation/permission.

	Causal	Overlapping
Connective spans (F_1)	0.77	0.89
Relation types (κ)	0.70	0.91
Degrees (κ)	0.92	(n/a)
Cause/ARGC spans (%)	0.89	0.96
Cause/ARGC spans (J)	0.92	0.97
Cause/ARGC heads (%)	0.92	0.96
Effect/ARGE spans (%)	0.86	0.84
Effect/ARGE spans (J)	0.93	0.92
Effect/ARGE heads (%)	0.95	0.89

Table 2: Inter-annotator agreement for the new version of BECauSE. κ indicates Cohen's kappa; J indicates the average Jaccard index, a measure of span overlap; and % indicates percent agreement of exact matches. Each κ and argument score was calculated only for instances with matching connectives.

An argument's head was determined automatically by parsing the sentence with version 3.5.2 of the Stanford Parser (Klein and Manning, 2003) and taking the highest dependency node in the argument span.

Means arguments were not included in this evaluation, as they are quite rare – there were only two in the IAA dataset, one of which was missed by one annotator and the other of which was missed by both. Both annotators agreed with these two means arguments once they were pointed out.

sentences, containing 98 causal instances and 82 instances of overlapping relations (per the first author's annotations). Statistics appear in Table 2.

Overall, the results show substantially improved connective agreement. F_1 for causal connectives is up to 0.77, compared to 0.70 in BECauSE 1.0. (The documents were drawn from similar sources and containing connectives of largely similar complexity as the previous IAA set.) The improvement suggests that the clearer guidelines and the overlapping relations made decisions less ambiguous, although some of the difference may be due to chance differences in the IAA datasets. Agreement on causal relation types is somewhat lower than in version 1.0 – 0.7 instead of 0.8 (possibly because more instances are annotated in the new scheme, which tends to reduce κ) – but it is still high. Unsurprisingly, most of the disagreements are between CONSEQUENCE and MOTIVATION. Degrees are close to full agreement; the only disagreement appears to have been a careless error. Agreement on argument spans is likewise quite good.

For overlapping relations, only agreement on ARGEs is lower than for causal relations; all other metrics are significantly higher. The connective F_1 score of 0.89 is especially promising, given the apparent difficulty of deciding which uses of connectives like *with* or *when* could plausibly be coerced to a causal meaning.

5.3 Corpus Statistics and Analysis

The corpus contains a total of 5380 sentences, among which are 1803 labeled instances of causal language. 1634 of these, or 90.7%, include both cause and effect arguments. 587 – about a third – involve overlapping relations. The corpus also includes 583 non-causal overlapping relation annotations. The frequency of both causal and overlapping relation types is shown in Table 3.

A few caveats about these statistics: first, PURPOSE annotations do not overlap with any of the categories we analyzed. However, this should not be interpreted as evidence that they have no overlaps. Rather, they seem to inhabit a different part of the semantic space. PURPOSE does share some language with origin/destination relationships (e.g., ***toward*** *the goal of*, ***in order to*** *achieve my goals*), both diachronically and synchronically; see §7.

	CONSEQUENCE	MOTIVATION	PURPOSE	All causal	NON-CAUSAL	**Total**
None	625	319	272	1216	-	1216
TEMPORAL	120	135	-	255	463	718
CORRELATION	9	3	-	12	5	17
HYPOTHETICAL	73	48	-	121	24	145
OBLIGATION/PERMISSION	67	5	-	72	27	99
CREATION/TERMINATION	37	4	-	41	43	84
EXTREMITY/SUFFICIENCY	53	9	-	62	-	62
CONTEXT	17	15	-	32	25	57
Total	994	537	272	1803	583	2386

Table 3: Statistics of various combinations of relation types. Note that there are 9 instances of TEMPORAL+CORRELATION and 3 instances of TEMPORAL+CONTEXT. This makes the bottom totals less than the sum of the rows.

Second, the numbers do not reflect all constructions that express, e.g., temporal or correlative relationships – only those that can be used to express causality. Thus, it would be improper to conclude that over a third of temporals are causal; many kinds of temporal language simply were not included. Similarly, the fact that all annotated EXTREMITY/SUFFICIENCY instances are causal is an artifact of only annotating uses with a complement clause, such as *so loud I felt it*; *so loud* on its own could never be coerced to a causal interpretation.

Several conclusions and hypotheses do emerge from the relation statistics. Most notably, causality has thoroughly seeped into the temporal and hypothetical domains. Over 14% of causal expressions are piggybacked on temporal relations, and nearly 7% are expressed as hypotheticals. This is consistent with the close semantic ties between these domains: temporal order is a precondition for a causal relationship, and often hypotheticals are interesting specifically because of the consequences of the hypothesized condition. The extent of these overlaps speaks to the importance of capturing overlapping relations for causality and other domains with blurry boundaries.

Another takeaway is that most hypotheticals that are expressed as conditionals are causal. Not all hypotheticals are included in BECauSE (e.g., *suppose that* is not), but all conditional hypotheticals are[6]: any conditional could express a causal relationship in addition to a hypothetical one. In principle, non-causal hypotheticals could be more common, such as *if he comes, he'll bring his wife* or *if we must cry, let them be tears of laughter*. It appears, however, that the majority of conditional hypotheticals (84%) in fact carry causal meaning.

Finally, the data exhibit a surprisingly strong preference for framing causal relations in terms of agents' motivations: nearly 45% of causal instances are expressed as MOTIVATION or PURPOSE. Of course, the data could be biased towards events involving human agents; many of the documents are about politics and economics. Still, it is intriguing that many of the explicit causal relationships are not just about, say, politicians' economic decisions having consequences, but about why the agents made the choices they did. It is worth investigating further to determine whether there really is a preference for appeals to motivation even when they are not strictly necessary.

6 Lessons Learned

Our experience suggests several lessons about annotating multiple overlapping relations. First, it clearly indicates that a secondary meaning can be evoked without losing any of the original meaning. In terms of the model of prototypes and radial categories (Lewandowska-Tomaszczyk, 2007), the conventional model for blurriness between categories, an instance can simultaneously be prototypical for one type of relation and radial for another. For instance, *the ruling **allows** the police **to** enter your home* is a prototypical example a permission relationship. However, it is also a radial example of enablement (a form of causation): prototypical enablement involves a physical barrier being removed, whereas *allow* indicates the removal of a normative barrier.

A second lesson: even when including overlapping semantic domains in an annotation project, it may still be necessary to declare some overlapping domains out of scope. In particular, some adjacent domains will have their own overlaps with meanings that are far afield from the target domain. It

[6]We did not annotate *even if* as a hypothetical, since it seems to be a specialized concessive form of the construction. However, this choice does not substantially change the conclusion: even including instances of *even if*, 77% of conditional hypotheticals would still be causal.

would be impractical to simply pull all of these second-order domains into the annotation scheme; the project would quickly grow to encompass the entire language. If possible, the best solution is to dissect the overlapping domain into a more detailed typology, and only include the parts that directly overlap with the target domain. If this is not doable, the domain may need to be excluded altogether.

For example, we attempted to introduce a TOPIC relation type to cover cases like *The President is fuming **over** recent media reports* or *They're angry **about** the equipment we broke* (both clearly causal). Unfortunately, opening up the entire domain of topic relations turned out to be too broad and confusing. For example, it is hard to tell which of the following are even describing the same kind of topic relationship, never mind which ones can also be causal: *fought **over** his bad behavior* (behavior caused fighting); *fought **over** a teddy bear* (fought for physical control); *worried **about** being late*; *worried **that** I might be late*; *I'm skeptical **regarding** the code's robustness*. We ultimately determined that teasing apart this domain would have to be out of scope for this work.

7 Contributions and Lingering Difficulties

Our approach leaves open several questions about how to annotate causal relations and other semantically blurry relations.

First, it does not eliminate the need for binary choices about whether a given relation is present; our annotators must still mark each instance as either indicating causation or not. Likewise for each of the overlapping relations. Yet some cases suggest overtones of causality or correlation, but are not prototypically causal or correlative. These cases still necessitate making a semi-abitrary call.

The ideal solution would somehow acknowledge the continuous nature of meaning – that an expression can indicate a relationship that is not causal, entirely causal, slightly causal, or anywhere in between. But it is hard to imagine how such a continuous representation would be annotated in practice.

Second, some edge cases remain a challenge for our new scheme. Most notably, we did not examine every semantic domain sharing some overlap with causality. Relations we did not address include:

- Origin/destination (as mentioned in §5.3; e.g., *the sparks **from** the fire*, ***toward** that goal*)
- Topic (see §6)
- Componential relationships (e.g., ***As part of** the building's liquidation, other major tenants will also vacate the premises*)
- Evidentiary basis (e.g., *We went to war **based on** bad intelligence*)
- Having a role (e.g., ***As** an American citizen, I do not want to see the President fail*)
- Placing in a position (e.g., *This move **puts** the American people **at** risk*)

These relations were omitted due to the time and effort it would have taken to determine whether and when to classify them as causal. We leave untangling their complexities for future work.

Other cases proved difficult because they seem to imply a causal relationship in each direction. The class of constructions indicating necessary preconditions was particularly troublesome. These constructions are typified by the sentence *(**For** us) **to** succeed, we all **have to** cooperate*. (Other variants use different language to express the modality of obligation, such as *require* or *necessary*.) On the one hand, the sentence indicates that cooperation enables success. On the other hand, it also suggests that the desire for success necessitates the cooperation.[7] We generally take the enablement relationship to be the primary meaning, but this is not an entirely satisfying account of the semantics.

Despite the need for further investigation of these issues, our attempt at extending causal language annotations to adjacent semantic domains was largely a success. We have demonstrated that it is practical and sometimes helpful to annotate all linguistic expressions of a semantic relationship, even when they overlap with other semantic relations. We were able to achieve high inter-annotator agreement and to extract insights about how different meanings compete for constructions. We hope that the new corpus, our annotation methodology and the lessons it provides, and the observations about linguistic competition will all prove useful to the research community.

[7] Necessary precondition constructions are thus similar to constructions of PURPOSE, such as *in order to*. As spelled out in Dunietz et al. (2015), a PURPOSE connective contains a similar duality of causations in opposing directions: it indicates that a desire for an outcome causes an agent to act, and hints that the action may in fact produce the desired outcome. However, in PURPOSE instances, it is clearer which relationship is primary: the desired outcome may not obtain, whereas the agent is centainly acting on their motivation. In precondition constructions, however, both the precondition and the result are imagined, making it harder to tell which of the two causal relationships is primary.

References

Steven Bethard, William J Corvey, Sara Klingenstein, and James H. Martin. 2008. Building a corpus of temporal-causal structure. In *Proceedings of LREC 2008*, pages 908–915.

Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Padó, and Manfred 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)*. Association for Computational Linguistics.

Jesse Dunietz, Lori Levin, and Jaime Carbonell. 2015. Annotating causal language using corpus lexicography of constructions. In *The 9th Linguistic Annotation Workshop held in conjunction with NAACL 2015*, pages 188–196.

Jesse Dunietz, Lori Levin, and Jaime Carbonell. in press. Automatically tagging constructions of causation and their slot-fillers. *Transactions of the Association for Computational Linguistics*.

Charles J. Fillmore, Paul Kay, and Mary Catherine O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language*, 64(3):501–538.

Adele Goldberg. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago University Press.

Cécile Grivaz. 2010. Human judgements on causation in French texts. In *Proceedings of LREC 2010*. European Languages Resources Association (ELRA).

Nancy Ide, Christiane Fellbaum, Collin Baker, and Rebecca Passonneau. 2010. The manually annotated sub-corpus: A community resource for and by the people. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 68–73. Association for Computational Linguistics.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics.

Barbara Lewandowska-Tomaszczyk. 2007. Polysemy, prototypes, and radial categories. *The Oxford handbook of cognitive linguistics*, pages 139–169.

Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The Penn Treebank: Annotating predicate argument structure. In *Proceedings of the Workshop on Human Language Technology*, HLT '94, pages 114–119. Association for Computational Linguistics, Stroudsburg, PA, USA.

Paramita Mirza and Sara Tonelli. 2014. An analysis of causality between events and its relation to temporal information. In *Proceedings of COLING*, pages 2097–2106.

Nasrin Mostafazadeh, Alyson Grealish, Nathanael Chambers, James Allen, and Lucy Vanderwende. 2016. CaTeRS: Causal and temporal relation scheme for semantic annotation of event structures. In *Proceedings of the 4th Workshop on Events: Definition, Detection, Coreference, and Representation*, pages 51–61. Association for Computational Linguistics.

Ad Neeleman and Hans Van de Koot. 2012. *The Theta System: Argument Structure at the Interface*, chapter The Linguistic Expression of Causation, pages 20–51. Oxford University Press.

Tim O'Gorman, Kristin Wright-Bettner, and Martha Palmer. 2016. Richer Event Description: Integrating event coreference with temporal, causal and bridging annotation. *Computing News Storylines*, page 47.

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

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of LREC 2008*. European Language Resources Association (ELRA), Marrakech, Morocco.

Josef Ruppenhofer, Michael Ellsworth, Miriam RL Petruck, Christopher R. Johnson, Jan Scheffczyk, and Collin F. Baker. 2016. FrameNet II: Extended theory and practice.

Evan Sandhaus. 2008. The New York Times annotated corpus. *Linguistic Data Consortium, Philadelphia*.

Nathan Schneider, Jena D. Hwang, Vivek Srikumar, Meredith Green, Abhijit Suresh, Kathryn Conger, Tim O'Gorman, and Martha Palmer. 2016. A corpus of preposition supersenses. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 99–109. Association for Computational Linguistics.

Nathan Schneider, Vivek Srikumar, Jena D. Hwang, and Martha Palmer. 2015. A hierarchy with, of, and for preposition supersenses. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 112–123. Association for Computational Linguistics, Denver, Colorado, USA.

Karin K. Schuler. 2005. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA. AAI3179808.

Noah A. Smith, Claire Cardie, Anne L. Washington, and John Wilkerson. 2014. Overview of the 2014 NLP unshared task in poliinformatics. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107. Association for Computational Linguistics.

Laure Vieu, Philippe Muller, Marie Candito, and Marianne Djemaa. 2016. A general framework for the annotation of causality based on FrameNet. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of LREC 2016*. European Language Resources Association (ELRA), Paris, France.