The Extraordinary Failure of Complement Coercion Crowdsourcing

Yanai Elazar   Victoria Basmov   Shauli Ravfogel   Yoav Goldberg   Reut Tsarfaty
Computer Science Department, Bar Ilan University
Allen Institute for Artificial Intelligence
{yanaiela,vikasaeta,shauli.ravfogel}@gmail.com
{yoav.goldberg,reut.tsarfaty}@gmail.com

Abstract

Crowdsourcing has eased and scaled up the collection of linguistic annotation in recent years. In this work, we follow known methodologies of collecting labeled data for the *complement coercion* phenomenon. These are constructions with an *implied* action — e.g., "I started a new book I bought last week", where the implied action is *reading*. We aim to collect annotated data for this phenomenon by reducing it to either of two known tasks: Explicit Completion and Natural Language Inference. However, in both cases, crowdsourcing resulted in low agreement scores, even though we followed the same methodologies as in previous work. Why does the same process fail to yield high agreement scores? We specify our modeling schemes, highlight the differences with previous work and provide some insights about the task and possible explanations for the failure. We conclude that specific phenomena require tailored solutions, not only in specialized algorithms, but also in data collection methods.

1 Introduction

Crowdsourcing has become extremely popular in recent years for annotating datasets. Many works use frameworks like Amazon Mechanical Turk (AMT) by converting complex linguistic tasks into easy-to-grasp presentations which make it possible to crowdsource linguistically-annotated data at scale (Bowman et al., 2015; FitzGerald et al., 2018; Dasigi et al., 2019; Wolfson et al., 2020).

In this work, we attempt to use existing methodologies for crowdsourcing linguistic annotations in order to collect annotations for *complement coercion* (Pustejovsky, 1991, 1995), a phenomenon involving an implied action triggered by an event-selecting verb. Specifically, certain **verb classes** require an *event-denoting* complement, as in: "I **started** *reading* a book", "I **finished** *eating* the

Task	Annotations
Explicit	
After a heartfelt vow, she agrees and the two begin kissing as the preacher tries to **continue** __ the ceremony.	{officiating}, ϕ
Entailment	
Hunter waited for max to **finish** his burger before asking him again. $\rightsquigarrow$ Hunter waited for max to **finish swallowing** his burger before asking him again.	ENT NEU CON

Table 1: Examples for the two modeling and annotation schemes used in this work. Both examples are labeled with different (disagreeing) answers. In the Explicit modeling, each label is a set, which can be empty (ϕ) (meaning that no event is implied), or not (and thus the context suggests an implied event). The second modeling follows the NLI scheme, a standard approach for evaluating language understanding. The ENT, NEU and CON labels refer to the entail, neutral and contradict labels accordingly.

cake", etc. However, such *event-denoting* complements might remain implicit, not appearing in the surface form. Consider for instance, the sentence "I **started** __ a new book." Here the event that was started remains implicit. Our task is then, first, to detect that the verb '**started**' in this context implies some unmentioned event, and that probable events in this context are *reading* or *writing*. Furthermore, we wish to predict that for "I **started** __ the book I bought yesterday", the more probable event is *reading*, rather than *writing*.

This phenomenon (described in detail in Section 2) seems intuitive at first, and easy-to-grasp by non-experts. However, we find that collecting annotated data for this task via crowdsourcing is very challenging, achieving low agreement scores between annotators (§3), despite using two common collection methods in frequently used setups. The two framings we use for data collection along with examples for them are presented in Table 1.

These low agreement scores come as a surprise, given the large body of previous work on crowd-sourcing linguistic annotations. Why do such issues arise when collecting data for *complement coercion*, while for similar phenomena the same approaches yield successful results? Although it is difficult to answer this question, we aim to highlight the similarities and the differences with other tasks, and provide some insights into this question.

2   Background

Complement Coercion   We are interested in the linguistic phenomenon of *complement coercion*.[1] In complement coercion, there is a clash between an expectation for a verb argument denoting an event, and the appearance of a noun argument denoting an entity. Uncovering the covert event requires the comprehender to infer the implied event by invoking the comprehender's lexical semantics and/or world knowledge (Zarcone et al., 2017).

Consider Examples 1 and 2 below, with an implicit event of *reading* or *writing* missing in the surface form. Inferring the implicit event (marked __) is necessary in order to construe the full semantics of this sentence.

1. I **started** __ a new book.

2. I **started** __ a new book I bought last week.

The reconstruction of the covert event requires an interplay between semantics[2] and world knowledge. In example 1 above, the prefix "I started __ " with the event-selecting verb *started* triggers expectations for some event-denoting object (*reading*, *writing*, *eating*, *watching*, etc). The object that follows, "a new book", narrows down the expectations — based on world knowledge. As McGregor et al. (2017) puts it, "Different nouns grant privileged access to different activities, particularly those which are most frequently performed with the entities they denote". Although the entity narrows down the set of possible events, the implied event might remain ambiguous (in Example 1, both *reading* and *writing* are plausible, but *eating* is not). As can be seen in Example 2, additional context, as in "I bought last week", provides further world-knowledge cues, towards accessing a more specific event (in this case

reading is more likely than *writing*), thus resolving the remaining ambiguity.

Complement coercion is particularly frequent with certain verb classes, including *aspectual verbs* — verbs that "describe the initiation, termination, or continuation of an activity" (Levin, 1993) — such as: 'start', 'begin', 'continue' and 'finish' (McGregor et al., 2017). This set of verbs is the focus of our work. Note however, that such verbs may appear in similar constructions that do *not* imply any covert action or event. For instance, in the following sentence:

3. I **started** a new company.

Here, the verb 'start' is used as an entity-selecting (and not event-selecting) verb, a synonym of 'found' or 'establish'. See more examples of similar non-coercive constructions in Appendix B.

Annotated data for complement coercion (Pustejovsky et al., 2010) was collected in the past, based on a tailor-made annotation methodology (Pustejovsky et al., 2009), consisting of a multi-step process that includes word-sense disambiguation by experts. The annotation focused on coercion detection (as well as labeling the arguments type) and did not involve identifying the implied action. Here, we aim to collect complement coercion data via non-expert annotation, at scale, to test whether models can recover the implicit events and resolve the emerging ambiguities.

Crowdsourcing NLI   NLI, originally framed as Recognizing Textual Entailment (RTE), has become a standard framework for testing reasoning capabilities of models. It originated from the work by Dagan et al. (2005), where a small dataset was curated by experts using precise guidelines with a specific focus on lexical and syntactic variability rather than delicate logical issues, while dismissing cases of disagreements or ambiguity. Bowman et al. (2015); Williams et al. (2018) then scaled up the task and crowdsourced large-scale NLI datasets. In contrast to Dagan et al. (2005), the task definitions were short and loose, relying on the annotators' common sense understanding. Many works since have been using the NLI framework and the crowd-sourcing procedure associated with it to test models for different language phenomena (Marelli et al., 2014; Lai et al., 2017; Naik et al., 2018; Ross and Pavlick, 2019; Yanaka et al., 2020).

---

[1]Complement coercion has been studied in linguistics from many theoretical viewpoints. See Appendix A for background.

[2]E.g., understanding the difference between entity-denoting and event-denoting elements.

3 Copmlement Coercion Crowdsourcing

3.1 Explicit Completion Attempt

We begin by directly modeling the phenomenon. For a set of sentences containing possibly-coercive verbs, we wish to determine for each verb if it entails an implicit event, and if so, to figure out what the event is. This direct task-definition approach is reminiscent of studies that collected annotated data for other missing elements phenomena, such as Verb-Phrase Ellipsis (Bos and Spenader, 2011), Numeric Fused-Heads (Elazar and Goldberg, 2019), Bridging (Roesiger, 2018; Hou et al., 2018) and Sluicing (Hansen and Søgaard, 2020). However, when attempting to crowdsource and label complement coercion instances, we reach very low agreement scores in the first step: determining whether there is an implied event or not. We discuss this experiment in greater detail in Appendix C.

3.2 NLI for Complement Coercion

In light of the low agreements on explicit modeling of the task of complement coercion, we turn to a different crowdsourcing approach which was proven successful for many linguistic phenomena – using NLI as discussed above (§2). NLI was used to collect data for a wide range of linguistic phenomena: Paraphrase Inference, Anaphora Resolution, Numerical Reasoning, Implicatures and more (White et al., 2017; Poliak et al., 2018; Jeretic et al., 2020; Yanaka et al., 2020; Naik et al., 2018) (see Poliak (2020)). Therefore, we take a similar approach, with similar methodologies, and make use of NLI as an evaluation setup for the complement coercion phenomenon.

Here we do not directly model the identification and recovery of event verbs, but rather, we reduce it to an NLI task. Intuitively, if in Example 2 the semantically plausible implied event is *reading*, we expect the sentence "I **started** a book I bought last week" to *entail* a sentence that contains the event explicitly: "I **started** *reading* a book I bought last week" (Table 2).[3] In contrast, we expect "I **started** a book" to be *neutral* with respect to "I **started** *reading* a book", since both *reading* and *writing* are plausible in that context, and there is no reason to prefer one of these complements over the other. Examples of this format, along with the different labels we employ, are shown in Table 2.

Example	Label
I started a book I bought last week. ↝	
I started reading a book I bought last week.	ENT
I started a book. ↝	
I started reading a book.	NEU
I started eating a book.	CON

Table 2: Examples for NLI pairs with a complement coercion structure. The ENT, NEU and CON labels refers to entail, neutral and contradict accordingly.

Corpus Candidates In order to keep the task simple, we avoid complexities of lexical, semantic and grammatical differences. Each example is composed of a minimal-pair (Kaushik et al., 2019; Warstadt et al., 2020; Gardner et al., 2020) consisting of two sentences; one as the premise and the other as the hypothesis. We construct minimal pairs as follows: First, we extract dependency-parsed sentences from the Book Corpus (Zhu et al., 2015) containing the lemma of one of the verbs: 'start', 'begin', 'continue' and 'finish'.[4] Then, we keep sentences where the anchor verb is attached to another verb with an 'xcomp' dependency[5] (e.g. 'started' in "started reading"). These sentences are used as the hypotheses. To construct the premises, we remove the dependent verb (e.g. 'read'), as well as all the words between the anchor and the dependent verb (e.g. 'to' in the infinitive form: "to read"). Additional examples are provided in Appendix D.

Note that this procedure sometimes generates ungrammatical or implausible sentences, which are flagged by the annotators.

Crowdsourcing Procedure We follow the standard procedure of collecting NLI data with crowdsourcing and collect annotations from Amazon Mechanical Turk (AMT). Specifically, we follow the instruction from Glockner et al. (2018), which involves three questions:

1. Do the sentences describe the same event?

2. Does the new sentence add new information to the original sentence?

3. Is the new sentence incorrect/ungrammatical?

We discard any example which at least one worker marked as incorrect/ungrammatical. If the answer

---

[3]We follow Bowman et al. (2015), who modeled entailment based on event coreference.

[4]These are frequent verbs that often appear in complement coercion constructions (McGregor et al., 2017).

[5]We use spaCy's parser (Honnibal and Johnson, 2015; Honnibal and Montani, 2017).

to the first question was negative, we considered the label as contradict. Otherwise, we considered the label as entail if the answer to the second question was negative, and neutral if it was positive. A screenshot of the interface is displayed in Figure 2 in the Appendix.

We require an approved rate of at least 99%, at least 5000 completed HITs, and filter workers to be from English-speaking countries. We also condition the turkers to pass a validation test with a perfect score. We pay 8 cents per HIT.

Results   We collect 76[6] pairs (after filtering ungrammatical sentences), each labeled by three different annotators. The Fleiss Kappa (Fleiss, 1971) agreement is $k = 0.24$. This score is remarkably low, compared to previous work that similarly collected NLI labels and achieved scores between 0.61 and 0.7. Why does this happen? Consider the following examples, along with their labels:

4. "We **finished** Letterman and I got up from the couch and said, I'm going to bed." $\rightsquigarrow$
   "We **finished** *watching* Letterman and I got up from the couch and said, I'm going to bed."
   ENT ENT ENT

5. "Flo set the sack of sausage and egg biscuits on the counter right as the young man **finished** his case." $\rightsquigarrow$
   "Flo set the sack of sausage and egg biscuits on the counter right as the young man **finished** *pleading* his case."
   ENT NEU CON

6. "We **start** the interviews later today." $\rightsquigarrow$
   "We **start** *shooting* the interviews later today."
   NEU CON CON

Example 4 was labeled by all three annotators as *entail*. However, annotators were in disagreement on examples 5, 6. Example 5 was annotated with all three possible labels (entail, contradict and neutral). Indeed, different readings of this phrase are possible — more formally, different readers *construe* the meaning of the utterance differently; "*[Construal] is a dynamic process of meaning construction, in which speakers and hearers encode and decode, respectively*" (Trott et al., 2020). An annotator who understands the word 'case' as a legal case, will choose *entail*, while an annotator

who interprets 'case' as a bag and imagines a different background story (for example, a young man packing a brief-case), will choose *contradict*. Finally, an annotator who thinks of both scenarios will choose *neutral*, which can be argued to be the correct answer. However, we find that for a human hearer, holding both scenarios in mind at the same time is hard, which we attribute to the *construal* of meanings. When a human construes an interpretation, they construes it in a single fashion until primed otherwise. So, it is not natural to conceive competing meaning scenarios when one is already "locked in" on a specific construal.

Although the sentence pairs were carefully built to exclude lexical and syntactic variances, ambiguous sentences such as the above recur throughout the dataset. We believe that these disagreements are inherent to this type of problem, and are not due to other factors such as poor annotations. As evidence, the authors of this work also annotated a subset of these examples and reached a similar (low) agreement.

4   Discussion

Inherent Disagreements in Human Textual Inferences   Recently, Pavlick and Kwiatkowski (2019) discussed a similar trend of disagreements in five popular NLI datasets (RTE (Dagan et al., 2005), SNLI (Bowman et al., 2015), MNLI (Williams et al., 2018), JOCI (Zhang et al., 2017) and DNC (Poliak et al., 2018)). In their study, annotators had to select the degree to which a premise entails a hypothesis, on a scale (Chen et al., 2020) (instead of discrete labels). Pavlick and Kwiatkowski (2019) show that even though these datasets are reported to have high agreement scores, specific examples suffer from inherent disagreements. For instance, in about 20% of the inspected examples, "there is a nontrivial second component" (e.g. entailment and neutral). Our findings are related to theirs, although not identical: while the disagreements they report are due to the individuals' interpretations of a situation, in our case, disagreements are due to the difficulty in imagining a different scenario. While some works propose to collect annotator disagreements and use them as inputs (Plank et al., 2014; Palomaki et al., 2018) (see Pavlick and Kwiatkowski (2019) for an elaborated overview), this will not hold in our case, because only one of the labels is typically correct.

However, the bottom-line is the same: these dis-

---

[6]We stopped at 76 examples since we did not see fit to annotate more data with the low agreements we obtained.

agreements cannot be dismissed as 'noise', they are more profound. We hypothesize that when tackling specific phenomena like the one we address in this work, which involve sources of disagreements that are often 'ignored' (not intentionally) during the collection of large datasets,[7] these sources of disagreements are highlighted and manifest themselves more clearly. This results in low agreement scores as we see in our study.

Scale Annotations Recent works have proposed to collect labels for NLI pairs on a scale (Pavlick and Kwiatkowski, 2019; Chen et al., 2020; Nie et al., 2020). Although we agree that this technique may produce a more fine-grained understanding of human judgments, Pavlick and Kwiatkowski (2019); Nie et al. (2020) observed that scale annotations may result in a multi-modality of the distribution. The different distributions can be viewed as different construals, where each individual interprets the example differently.

Task Definition Another issue might arise from the task definition itself. As opposed to annotation efforts for linguistic tasks such as parsing (Marcus et al., 1993) and semantic role labeling (Carreras and Màrquez, 2005) that are carried out by expert annotators and often have annotation guidelines of dozens of pages, the transition to crowdsourcing has reduced the guidelines to a few phrases, and expert annotators have been replaced by laymen. This transition required to simplify the guidelines and to avoid complex definition and corner-cases. Even though crowdsourcing enabled an easier annotation process and collection of huge amounts of data, it also came with a cost: lack of refined definitions and relying on people's "common sense" and "intuition". However, as we see in this work, such intuitions are not consistent across individuals and are not sufficient for some tasks. We believe that, similar to the issues mentioned above, the lack of proper definitions tends to amplify disagreements when dealing with specific phenomena, which was often the reason behind the elaborated and long guidelines in classic datasets (Kalouli et al., 2019).

Possible Solution As we approach "solving" current NLP dataset, which were once perceived as complicated, we also reach an understanding that the datasets at hand do not reflect the full capacity of language, and specific linguistic phenomena, which may posses specific challenges, are lost in

the crowds. Some phenomena turn out to be more complex, and require specific solutions. In this work we show that, like we do with algorithmic solutions we need to reconsider the data collection process. We hold that data collection for these phenomena also require training of the annotators (Roit et al., 2020; Pyatkin et al., 2020), whether experts or crowdsourcing workers, and may also require coming up with novel annotation protocols.

Another potential solution is to use deliberation between the workers as a mean to improve agreement (Schaekermann et al., 2018). With respect to the disagreements we observed, a deliberation between workers would allow them to share the construals each individual had imagined, thus reaching a consensus on the labels. It would also serve as a training for recovering more construals, allowing them to better identify the *neutral* cases.

5 Conclusions

In this work, we attempt to crowdsource annotations for complement coercion constructions. We use two modeling methods, which were successful in similar settings, but resulted in low agreement scores in our setup. We highlight some of the issues we believe are causing the disagreements. The main one being different construals (Trott et al., 2020) of the utterances by different people — as well as the difficulty to consider a different one, once fixating on a specific construal — that led to different answers. We connect our findings to previous work that observed some inherent disagreement in human judgments in popular datasets, such as SNLI and MNLI (Pavlick and Kwiatkowski, 2019). Although this issue is less prominent in these datasets (which is manifested as higher agreement scores), we notice that when tackling a *specific* phenomenon, e.g. involving implicit elements, these issues may arise.

We also argue that the *lack* of detailed definitions in the commonly used NLI tasks may lead to poor performance on small buckets of language-specific phenomena. This drop might be lost in large-scale datasets, but may have critical effects when modeling and studying specific phenomena. As a community, we claim, we should seek to identify those buckets and further investigate them, using more profound approaches for data collection, with clear and grounded definitions. We hope that our attempted trial in data collection will allow others to learn from our failure.

---

[7]Due to large scale annotations, 'marginal' phenomena might be ignored to keep the instructions clear and concise.

References

Roberto G. de Almeida and Roberto G. Veena D. Dwivedi. 2008. Coercion without lexical decomposition: Type-shifting effects revisited. *The Canadian Journal of Linguistics / La revue canadienne de linguistique*, 53:301 – 326.

Johan Bos and Jennifer Spenader. 2011. An annotated corpus for the analysis of vp ellipsis. *Language Resources and Evaluation*, 45(4):463–494.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Xavier Carreras and Lluís Màrquez. 2005. Introduction to the conll-2005 shared task: Semantic role labeling. In *Proceedings of the ninth conference on computational natural language learning (CoNLL-2005)*, pages 152–164.

Tongfei Chen, Zhengping Jiang, Adam Poliak, Keisuke Sakaguchi, and Benjamin Van Durme. 2020. Uncertain natural language inference. In *Proceedings of The 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pages 177–190. Springer.

Pradeep Dasigi, Nelson F Liu, Ana Marasovic, Noah A Smith, and Matt Gardner. 2019. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5927–5934.

Yanai Elazar and Yoav Goldberg. 2019. Where's my head? definition, data set, and models for numeric fused-head identification and resolution. *Transactions of the Association for Computational Linguistics*, 7:519–535.

Nicholas FitzGerald, Julian Michael, Luheng He, and Luke Zettlemoyer. 2018. Large-scale qa-srl parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2051–2060.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hanna Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating nlp models via contrast sets. *arXiv preprint*.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.

Daniele Godard and Jacques Jayez. 1993. Towards a proper treatment of coercion phenomena. In *Sixth Conference of the European Chapter of the Association for Computational Linguistics*, Utrecht, The Netherlands. Association for Computational Linguistics.

A. E. Goldberg. 1995. *Constructions: A construction grammar approach to argument structure.* Chicago: University of Chicago Press.

Victor Petrén Bach Hansen and Anders Søgaard. 2020. What do you mean'why?': Resolving sluices in conversations. In *AAAI*, pages 7887–7894.

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1373–1378.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1).

Yufang Hou, Katja Markert, and Michael Strube. 2018. Unrestricted bridging resolution. *Computational Linguistics*, 44(2):237–284.

E. Matthew Husband, Lisa A. Kelly, and David C. Zhu. 2011. Using complement coercion to understand the neural basis of semantic composition: Evidence from an fmri study. *Journal of Cognitive Neuroscience*, 23:3254–3266.

Ray Jackendoff. 1996. The architecture of the language faculty. MIT Press.

Ray Jackendoff. 2002. Foundations of language: Brain, meaning, grammar, evolution.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705, Online. Association for Computational Linguistics.

Aikaterini-Lida Kalouli, Annebeth Buis, Livy Real, Martha Palmer, and Valeria dePaiva. 2019. Explaining simple natural language inference. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 132–143.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2019. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.

Zoltán Kövecses and Günter Radden. 1998. Metonymy: Developing a cognitive linguistic view. *Cognitive linguistics*, 9(1):37–77.

Gina R. Kuperberg, Arim Choi, Neil Cohn, Martin Paczynski, and Ray Jackendoff. 2010. Electrophysiological correlates of complement coercion. *Journal of Cognitive Neuroscience*, 22:2685–2701.

Alice Lai, Yonatan Bisk, and Julia Hockenmaier. 2017. Natural language inference from multiple premises. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 100–109.

Beth Levin. 1993. *English Verb Classes and Alternations*. The University of Chicago Press.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, Roberto Zamparelli, et al. 2014. A sick cure for the evaluation of compositional distributional semantic models. In *LREC*, pages 216–223.

Brian McElree, Liina Pylkkänen, Martin J. Pickering, and Matthew J. Traxler. 2006. A time course analysis of enriched composition. *Psychonomic Bulletin & Review*, 13:53–59.

Stephen McGregor, Elisabetta Jezek, Matthew Purver, and Geraint Wiggins. 2017. A geometric method for detecting semantic coercion. In *IWCS 2017 - 12th International Conference on Computational Semantics - Long papers*.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data?

Jennimaria Palomaki, Olivia Rhinehart, and Michael Tseng. 2018. A case for a range of acceptable annotations. In *SAD/CrowdBias@ HCOMP*, pages 19–31.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Maria Mercedes Piñango and Ashwini Deo. 2016. Reanalyzing the complement coercion effect through a generalized lexical semantics for aspectual verbs. *J. Semantics*, 33:359–408.

Maria Piñango and Ashwini Deo. 2014. Reanalyzing the complement coercion effect through a generalized lexical semantics for aspectual verbs. *Journal of Semantics*, 33.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 742–751.

Adam Poliak. 2020. A survey on recognizing textual entailment as an nlp evaluation.

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. 2018. Collecting diverse natural language inference problems for sentence representation evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 67–81.

James Pustejovsky. 1991. The generative lexicon. *Comput. Linguistics*, 17:409–441.

James Pustejovsky. 1995. *The Generative Lexicon*. MIT Press, Cambridge, MA.

James Pustejovsky and Pierrette Bouillon. 1994. On the proper role of coercion in semantic typing. In *COLING 1994 Volume 2: The 15th International Conference on Computational Linguistics*.

James Pustejovsky, Jessica Moszkowicz, Olga Batiukova, and Anna Rumshisky. 2009. Glml: Annotating argument selection and coercion. In *Proceedings of the Eight International Conference on Computational Semantics*, pages 169–180.

James Pustejovsky, Anna Rumshisky, Alex Plotnick, Elisabetta Jezek, Olga Batiukova, and Valeria Quochi. 2010. Semeval-2010 task 7: Argument selection and coercion. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 27–32.

Valentina Pyatkin, Ayal Klein, Reut Tsarfaty, and Ido Dagan. 2020. Qadiscourse – discourse relations as qa pairs: Representation, crowdsourcing and baselines.

Ina Roesiger. 2018. BASHI: A Corpus of Wall Street Journal Articles Annotated with Bridging Links. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Paul Roit, Ayal Klein, Daniela Stepanov, Jonathan Mamou, Julian Michael, Gabriel Stanovsky, Luke Zettlemoyer, and Ido Dagan. 2020. Controlled crowdsourcing for high-quality qa-srl annotation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7008–7013.

Alexis Ross and Ellie Pavlick. 2019. How well do nli models capture verb veridicality? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2230–2240.

Mike Schaekermann, Joslin Goh, Kate Larson, and Edith Law. 2018. Resolvable vs. irresolvable disagreement: A study on worker deliberation in crowd work. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–19.

Sean Trott, Tiago Timponi Torrent, Nancy Chang, and Nathan Schneider. 2020. (re)construing meaning in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5170–5184, Online. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Aaron Steven White, Pushpendre Rastogi, Kevin Duh, and Benjamin Van Durme. 2017. Inference is everything: Recasting semantic resources into a unified evaluation framework. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 996–1005.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Tomer Wolfson, Mor Geva, Ankit Gupta, Matt Gardner, Yoav Goldberg, Daniel Deutch, and Jonathan Berant. 2020. Break it down: A question understanding benchmark. *Transactions of the Association for Computational Linguistics*.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, and Kentaro Inui. 2020. Do neural models learn systematicity of monotonicity inference in natural language? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL2020)*, pages 6105—-6117.

Lai Yao-Ying. 2017. *The complement coercion phenomenon: Implications for models of sentence processing*. Ph.D. thesis, Yale University.

Soyeon Yoon. 2012. Constructions, semantic compatibility, and coercion: An empirical usage-based approach.

Alessandra Zarcone, Ken McRae, Alessandro Lenci, and Sebastian Padó. 2017. Complement coercion: The joint effects of type and typicality. *Frontiers in Psychology*, 8:1987.

Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2017. Ordinal common-sense inference. *Transactions of the Association for Computational Linguistics*, 5:379–395.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

A Linguistic Background

Complement coercion has been studied in linguistics from many theoretical viewpoints. Lexical semantic accounts (such as Pustejovsky 1991, 1995 and others) and Construction Grammar accounts (e.g. Goldberg 1995) "attempt to formalize what semantic features of a lexical item have been changed to conform to those of the construction" (Yoon, 2012). One of the main approaches is the Type-Shifting analysis (Pustejovsky, 1991, 1995; Jackendoff, 1996, 2002), "which asserts that complement coercion involves a type-shifting operation that coerces the entity-denoting complement to an event"(Yao-Ying, 2017). Another approach (de Almeida and Dwivedi 2008 and others) "claims that complement coercion involves a hidden VP structure with an empty verb head, which is saturated by pragmatical inference in context" (Yao-Ying, 2017). Cognitive linguistics accounts (such as Kövecses and Radden 1998) exploit metonymy as the mechanism behind coercion constructions (Yoon, 2012). Complement coercion has been also extensively investigated in the framework of neurolinguistic research (for example, Kuperberg et al. 2010) and psycholinguistic studies (e.g., McElree et al. 2006). The latter often show that "coercion sentences elicit increased processing times" (Husband et al., 2011) compared with non-coercion sentences. Such theories as the Type-Shifting Hypothesis mentioned above and the Structured-Individual Hypothesis (Piñango and Deo, 2016) suggest different explanations for this associated processing cost (Yao-Ying, 2017).

B Complement Coercion: Counter Examples

Here we provide some additional examples of constructions that are similar to the ones in Examples 1,2 (the verb 'start' is followed by a non-event-denoting complement) but do *not* function as complement coercion constructions. Consider the following sentences:

7. I **started** a new company.

8. His name **started** the list.

9. Her wedding dress **started** a new tradition among brides.

In example 7 the verb 'start' is used as an entity-selecting (and not event-selecting) verb, a synonym of 'found', 'establish', so that there is no type clash.

Please read the following sentence:

"Alan started a new book."

What verb(s) can be added after **started** to make the sentence more complete?

First interpretation:

reading

Second interpretation:

writing

☐ No adequate verb in this context

Figure 1: A screenshot of the explicit task presented to the annotators.

In example 8 the verb 'start' is used in its 'non-eventive' (Zarcone et al., 2017) or 'stative' (Piñango and Deo, 2014) sense ('constitute the initial part of something'). When used this way, the verb 'start' does not exclusively select for eventive complements, so, again, there is no type clash. Also, some authors (Godard and Jayez, 1993; Yao-Ying, 2017; Pustejovsky and Bouillon, 1994) argue that in coercion constructions the subject should be an "intentional controller of the event" (Godard and Jayez, 1993). In example 9 this condition does not hold, therefore there is no coercion.

C Explicit Modeling

In the *Explicit Completion* approach, the goal is to add the implicit argument of the coercion construction, if such completion exists. For instance, in the sentence "I started __ a new book", possible completions are 'reading' and 'writing', and in Example 7 no completion fits. Concretely, given a sentence with a complement coercion verb candidate, the task is to complete it with a set of possible verbs that describe the covert event. As not all candidates function as parts of complement coercion constructions, annotators can mark that no additional verb is adequate in the context. In cases where there is more than one semantically plausible answer (e.g. Ex. 1), we ask annotators to provide two completion sets, each consisting of a group of semantic equivalent verbs, which correspond to different possible understandings of the text. A screenshot of the task presented to the turkers is shown in Figure 1.

This approach to task definition is reminiscent of those used for other missing elements phenom-

ena, such as Verb Phrase Ellipsis (Bos and Spenader, 2011), Numeric Fused-Heads (Elazar and Goldberg, 2019), Bridging (Roesiger, 2018; Hou et al., 2018) and Sluicing (Hansen and Søgaard, 2020). However, in contrast to these tasks, where the answers can usually be found in the context,[8] the answers in our case are more open-ended (although still bounded by some restrictions (Godard and Jayez, 1993; Pustejovsky and Bouillon, 1994)). This makes this task more challenging for annotation.

Corpus Candidates   In the *explicit completion* setting, we look for natural sentences that contain one of the following anchor verbs: 'start', 'begin', 'continue' and 'finish', - immediately followed by a direct object without any dependent verb in between.

Annotation Procedure   We use the same restrictions from the previous procedure and create a new validation test, tailored for the new task. We pay 4 cents per Hit.

Results   We collect annotations for 200 sentences, with two annotations per sentence. We compute the Fleiss Kappa (Fleiss, 1971) after a relaxation of the annotations into two labels: added a complement or not. Similarly to the previous modeling, the agreement score is $k = 0.18$, which is considered to be low. Consider the following examples:

9. "In 2011, Old Navy **began** ‗‗ a second rebranding to emphasize a family-oriented environment, known as Project ONE.", — $\{advertising, promoting, endorsing\}, \phi$

10. "After he had **finished** ‗‗ his studies Sadra began to explore unorthodox doctrines and as a result was both condemned and excommunicated by some Shi'i 'ulamā'.", — $\{pursuing, doing\}, \phi$

According to the definition of complement coercion, these examples do not require a complement. However, as can be seen from these examples, the proposed complements do contribute to an easier understanding of the sentence. We note that this concept of 'missing' is hard to explain and can be also subjective. Another obstacle is that strict adherence to the linguistic definition does not always

---

Sentence 1: "They jumped onto the pier and started the boat."

Sentence 2: "They jumped onto the pier and started untying the boat."

Do the sentences describe the same event?

○ Yes

◉ No

Does the new sentence add new information to the original sentence?

◉ Yes

○ No

Is the first sentence incorrect/ungrammatical?

○ Yes

◉ No

Figure 2: Screenshot of the interface shown to the turkers for collecting labels. This setup follows the instructions used for labeling NLI data in Glockner et al. (2018).

contribute to potential usefulness of the task for downstream applications. For this phenomenon, we did not follow the strict linguistic definition and used a more relaxed one. Additional examples along with their annotations are provided in Table 4.

D NLI Framing: Additional Material

We provide a screenshot of the NLI interface shown to the turkers in Figure 2.

NLI Data   We provide additional examples for the original and the modified sentences (hypotheses and premises accordingly) used in the NLI framing (§3.2), along with the three obtained labels, in Table 3.

Premise	Hypothesis	Annotations
that gives us something to work with if he starts trouble.	that gives us something to work with if he starts making trouble.	ENT ENT ENT
I do hope you will continue mrs. cox's incredible hospitality.	I do hope you will continue to enjoy mrs. cox's incredible hospitality.	CON CON CON
he asked me as he continued a tune.	he asked me as he continued to strum a tune.	NEU NEU CON
how would she continue questions like this?	how would she continue to answer questions like this?	CON CON CON
he finished a sip of coffee and replied, not surprised.	he finished taking a sip of coffee and replied, not surprised.	ENT ENT NEU
it was pike's idea to start these games.	it was pike's idea to start playing these games.	ENT NEU CON
I started deep breaths and tried to cleanse my mind.	I started taking deep breaths and tried to cleanse my mind.	ENT ENT NEU
I would like to finish this movie sometime in this year!	I would like to finish watching this movie sometime in this year!	ENT ENT CON

Table 3: Examples for NLI pairs with a complement coercion structure. The ENT, NEU and CON labels refers to the entail, neutral and contradict accordingly.

Text	Annotations
... it will likely travel in a parabola, *continuing* its stabilizing spin, ...	ϕ, ϕ
Afterwards, they decide to *continue* the pub crawl to avoid attracting suspicion.	$\{doing\}, \{doing\}$
I was surprised he did not *continue* his openness at the RFPERM.	$\{embue\}, \{showing, displaying, ...\}$
In 1994, he joined Motilal Oswal to *start* their institutional desk before moving to UBS in 1996.	$\{employ\}_1, \{work\}_2, \{working\}$
In 1943 she *started* a career as an actress with the stage name Sheila Scott a name ...	$\phi, \{pursuing\}$
..., giving him the opportunity to *continue* the work left by his predecessors as well as ...	$\phi, \{researching, studying\}$
In the Middle Ages it was a battle cry , which was used to *start* a Feud or a Combat reenactment.	$\phi, \{fighting\}$
In addition, deductions are taken if the man *finishes* the element on two feet ...	$\phi, \{competing\}$

Table 4: Examples for the Explicit modeling. ϕ denotes the empty set, meaning no event is implied. When a subscript is present it denotes the different interpretation of the sentence, by the same annotator.