Remodelling complement coercion interpretation

Frederick Gietz
Department of Linguistics
University of Toronto
Toronto, ON
frederick.gietz@utoronto.ca

Barend Beekhuizen
Department of Linguistics
University of Toronto
Toronto, ON
barend.beekhuizen@utoronto.ca

Abstract

Existing (experimental and computational) linguistic work uses participant paraphrases as a stand-in for event interpretation in complement coercion sentences (e.g. *she finished the coffee* → *she finished drinking the coffee*). We present crowdsourcing data and modelling that supports broadening this conception. In particular, our results suggest that sentences where many participants do not give a paraphrase, or where many different paraphrases are given are informative about to how complement coercion is interpreted in naturalistic contexts.

1 Interpreting word meanings in context

A central aspect of pragmatic reasoning is to construe utterance meaning which is not overly given in the sentence (Grice, 1975). This paper uses crowdsourcing and computational modeling to explore the range of possible interpretations in a particular grammatical construction in which implicit meaning is (frequently) to be inferred, namely *complement coercion* sentences. These are sentences like *they finished the coffee* or *she began a book*, where the entity-type direct object is 'coerced' into an event-type interpretation applying to that direct object (e.g., 'they finished **drinking** the coffee' or 'she began **writing** the book').

The traditional treatment is that these sentences involve a case of type-shifting, where the direct object whose extension is a physical entity is instead interpreted as an event involving that direct object (Pustejovsky and Bouillon, 1995). On this account, readers leverage the lexical-semantic information of the direct object itself to arrive at a specific event (e.g., **drink** for 'they finished the coffee'). In contrast to the type-shifting account, the pragmatic account of Piñango and Deo (2016) suggests that readers instead pragmatically retrieve a relevant scale to enrich the interpretation of aspectual verbs that have entity-type direct objects. This scale can be temporal in the case of an eventive interpretation (e.g. *I sat down and began the book*) but also spatial (e.g. *The marker begins the trail*). Crucially for our purposes, pragmatic enrichment is not a lexical process resolving a specific verb. This enrichment can take place to a greater or lesser extent, in principle even allowing for a lack of enrichment, although that option is not presented explicitly in their paper.

Complement coercion has drawn attention from different communities of scholars. Psycholinguists found that (simply put) complement coercion incurs a processing cost (McElree et al., 2001; Traxler et al., 2002), while computational linguists have shown an interest in complement coercion as a challenging case of automatically retrieving implicit aspects of sentence interpretation (Lapata and Lascarides, 2003; Roberts and Harabagiu, 2011; Zarcone et al., 2012; Chersoni et al., 2017). Interestingly, both lines of research inherit the assumption from the type-shifting account that complement coercion sentences have a verb paraphrase that represents the interpreted event, and largely design test items based on this assumption. In our paper, we focus on the computational task of modeling the interpretation of sentences containing complement coercion and the light it can shed on the two theoretical accounts, but we briefly touch on the implications for experimental work in §7.

For computational formulations of the task of complement coercion interpretation, inheriting the 'obligatory (semantic) resolution' property from the type-shifting account means that coercion interpretation is conceptualized as a case of multi-label classification in which models predict a single event label (verb) which is then evaluated against annotator consensus about the correct event label (Lapata and Lascarides, 2003; Zarcone and Padó, 2010; Zarcone et al., 2012). In §2, we demonstrate that this conception obscures many relevant and

Proceedings of the Society for Computation in Linguistics (SCiL) 2022, pages 158-170.
Held on-line February 7-9, 2022

interesting cases of complement coercion where verb paraphrase is not sufficient to represent interpretation. Then, in §3 and §4, we introduce models for complement coercion interpretation designed for the simple verb prediction task. In §5 and §6, we highlight two types of cases which break from the typical examples shown in the theoretical literature – cases where participants prefer not to give any verb paraphrase ('blanks') and cases where participants are divided on which verb to use ('low-consensus'). By building improvements to our models to handle these two cases, we suggest that complement coercion is best modeled as a form of (optional, or at least gradient) pragmatic enrichment rather than as obligatory semantic completion.

2 Elicitation study

2.1 Crowdsourcing with Blank responses

Existing experimental and computational work relies on (crowdsourced) norming data to determine how complement coercion sentences are interpreted (Zarcone and Padó, 2010; McElree et al., 2001; Traxler et al., 2002; Frisson and McElree, 2008). Comparing the sentences in experimental and computational studies with cases of complement coercion from a corpus of naturally occurring text (the Corpus of Contemporary American English, or COCA: Davies, 2009), we observed that the interpretation of naturally occurring cases often differs from the hand-crafted examples used in experimental and computational work, in particular in that hand-crafted examples typically allow for a clear verb paraphrase, often a single one, whereas naturally-occurring sentences often seem to lack this property.

This exploratory observation led us to design a new elicitation experiment in which we used naturally-occurring cases of complement coercion. First, candidate complement coercion sentences, containing an aspectual verb (*begin, end, start, finish, complete*) and a likely coerced entity object were extracted from COCA using heuristics discussed in Appendix A. 300 of the $4,583$ likely instances were sampled for an elicitation experiment, in which we asked participants to fill a blank between the verb and the direct object (e.g., *She finished ____ her book*), similar to the papers cited above. In contrast to these approaches, and in line with our expectation that not all cases readily elicit verb paraphrases, participants were instructed that blanks should be left empty if no verb was felt to fit it. Appendix B presents an example of the elicitation prompt and several participants responses. [1] Using Testable (Testable), we gathered on average 19 (range: 15–20) responses per item.

In line with our initial intuition, our data displayed a large amount of 'blank' responses. In 138/300 sentences (46%), the most common response was a blank one. The remaining 162 sentences displayed substantial variation in the degree to which participants agreed with each other. Defining the *consensus* of an item as the proportion of participants who gave the dominant response, our data displays a median consensus of 55% (IQR: 40%–74%). To illustrate: an example such as (1-a) has a similar direct object as (1-b), yet received a majority of blank responses (12/19, vs. 1/15 for the latter). Similarly, example (1-c) received 4/19 *paint* responses, versus 9/15 for example (1-b), both again with similar direct objects. In contrast, constructed cases often have a high consensus compared to naturalistic examples (e.g., example (1-d) had 58% of participants in the norming study of (Frisson and McElree, 2008) respond with *paint*).

(1)
a. Lordier began ___ the painting with a very light sketch of the major shapes...
b. In 1951 he began ___ a second mural, a portrayal of St. Joseph as the master craftsman...
c. You will see the final texture effect when you click OK. You have just completed ___ your textured picture.
d. The artist began ___ the portrait in his studio in the city.

2.2 Interpretive strategies vary across cases

We believe the prevalence of blank responses and low-consensus responses is not an effect of poor annotator training or different annotator conceptions of the target event, as Elazar et al. (2020) suggest, but instead an effect of the varying demands on the pragmatic resolution of the event that different examples bring about. In cases like (1-a), the implicit event is not critical to understanding the sentence; rather, the manner of the event (*with a very light sketch . . .*) is more salient to interpretation. Partici-

[1] This final 300 includes 16 (5.3%) with an inanimate subject, e.g. *A pretty bow completes the picture.* We kept these sentences to compare participant responses as they fit the complement coercion pattern by definition, recognizing they potentially form a subcategory or separate aspectual verb sense.

pants may consider the specific nature of the event to be backgrounded, and for that reason elect to leave the response blank.

Similarly, in (1-c), the act of completion seems to be the primary message of the sentence rather than the specific nature of what is completed. Unlike in (1-a), blank responses do not dominate, but participants display a lower degree of consensus about which verb to fill out than for (1-b): for (1-c), 4/15 respond with *paint*, but we also find highly similar responses like *edit* (3/15), *make, design, print* and *render* (all 1/15), that all convey a sense of creation.[2] Elazar et al. (2020) argue that such low-consensus cases potentially reflect respondents' different construals of the same situation. We propose instead that the fact that 11/15 responses reflect a general sense of creation is indicative of speakers agreeing on the broad sense of the coerced event (here: 'creation'), but disagreeing when forced to come up with a specific verb to fit that broad event.

Other cases are found in the data where no verb is dominant, but where participants still give some verb responses sharing the same broad sense. For example, both sentences (2-a) and (2-b) receive majority blank responses (12/20 and 11/19 responses, respectively), but other responses include verbs about creation like *make* and *build*. Similarly, the sentences in (3) all had the most popular response *write*, but none had it as a majority (4/20, 6/19, and 7/19 responses, respectively). Less popular responses included *publish*, *make*, and *compile*.

(2) a. Lau next inserts a set of wire filaments into the chamber... He completes ___ the setup by fitting a quartz cover on the top of the reactor.
b. Complete ___ the rig by threading a double-length of wire leader through the tube and egg sinkers.

(3) a. Together with Sky Telescope's Roger W. Sinnott, Tirion has just finished ___ a new edition of his classic Sky Atlas 2000.
b. McGruder began ___ his politically charged hip-hop comic strip for his college newspaper.
c. He did finish ___ Harvard Man - a story, he says, about sex, drugs, madness, orgasm, philosophy, and college basketball fixing.

[2]For completeness: two further responses were blank, and *click* and *prepare* were both given once.

An anonymous reviewer points out that some sentences may receive blank responses due to factors besides the event interpretation. Specifically, the fill-in-the-blank style of crowdsourcing may discourage responses which are valid verb interpretations but which cause grammaticality issues or redundancy when given overtly. For example, in sentence (4-a), 15/19 participants give a blank response, compared to 3/19 who give the verb '*install*' and 1/19 with the verb '*make*.' It is likely that some participants leave this sentence blank to avoid an ungrammatical double-gerund construction. However, a grammatically similar sentence, like (4-b), receives a 90% consensus response in *eat*. Similarly, the presence of a direct object ending in *-ing* may keep participants from presenting verbs ending in *-ing*. While we are certain that these low-level factors impact consensus rates to some extent, there are many counter-examples to such explanation, among cases like (2-a) and (3), where an explanation in terms of grammaticality or redundancy avoidance cannot be given.

(4) a. FINISHING ___ THE ROLL-OFF ROOF RAILS
b. Pauline and Juliet are finishing their grapes as they watch Hilda and Walter on the tennis court .

Overall, we take inconsistent and blank responses to be information (rather than noise) about how participants actually resolve these sentences when reading. For sentences which receive many blank responses or low-consensus among responses, we suggest that participants only resolve the interpretation to a specific verb because the task formulation forces or nudges them to do so.

This leads us to suggest a novel account for the two new types of cases introduced in this paper. For both types, the presence of complement coercion does not obligatorily lead to a particular event interpretation, as implicated by the account of Pustejovsky and Bouillon (1995). Rather, they fit better the account of Piñango and Deo (2016), where speakers are said to interpret an aspectual verb as related to some pragmatically determined scale but not necessarily to resolve that interpretation to the level of a specific event. Note that it is only this property of 'obligatory resolution' on

which we compare the two accounts – this paper does not make claims about any further differing properties of the two accounts.

Overall, we take sentences where the top response is a blank and sentences with a low consensus rate as indicative that not all cases of complement coercion need to be 'resolved' to a specific event, as labeled by a specific verb. Communication can succeed even when the interpretated event is left vague or underspecified (Frisson, 2009), and models of complement coercion interpretation should capture the proposed variation in the interpretation process, as evidenced by participants' diverse types of responses. In other words, we seek to build models which make more informative predictions than a single verb paraphrase, and we argue that re-conceptualizing the task allows us to better understand the linguistic (pragmatic) properties of complement coercion in return.

3 Modeling complement coercion

3.1 Redefining the modelling task

To recapitulate: previous work defines the modeling task of complement coercion interpretation as the prediction of a single, high-consensus verb paraphrase for a given sentence in line with the theoretical conception of complement coercion as involving obligatory resolution to the level of a particular event. Our annotation data show that only a minority of all sentences display a consensus of over 50%, and for almost half the items a blank response is dominant – two effects we argued not to be due to poor annotation or improperly trained annotators, but instead to the varying pragmatic demands on the resolution of apparent cases of complement coercion. In the following sections, we evaluate complement coercion interpretation models on our new dataset.

Crucially, the gold labels derived from our dataset differ in two ways from from those as formulated by similar tasks (e.g., Zarcone et al. (2012); Chersoni et al. (2017)): (1) the correct label for items is taken to be 'blank' if the dominant response was 'blank', and (2) all the low-consensus cases are included, using the dominant response as their label. We consider these changes to see how models that follow the 'predict-the-verb' task formulation fare on these two groups of cases in §4, after which we look at two extensions that explicitly take the varying pragmatic demands on interpretation into account in §5 and §6.

3.2 Models of complement coercion

Several models have been defined to model complement coercion detection (whether a case of an aspectual verb plus direct object is coercive or not) and interpretation (which event is to be inferred to 'fill the blank'). Existing models build on the intuition that the direct object of a sentence is uniquely informative for constructing a coercion interpretation (Lapata and Lascarides, 2003; Roberts and Harabagiu, 2011; Zarcone et al., 2012). First, we use the **Example Based Learning** model, or EBL (broadened from McGregor et al., 2017's coercion identification model). EBL is our only supervised model. For a given test sentence *S*, EBL predicts the interpretation to be the most common interpretation of training sentences that have the same direct object as *S*. For example, if 6/9 of the training sentences containing the direct object *book* have the top response *write*, 2/9 *read*, and 1/9 'blank', EBL will predict the answer *write* when presented with a test sentence containing the direct object *book*. If a direct object of a test sentence does not occur in the training set, EBL predicts a blank.

A second model, the **Co-occurrence counts** model (COOC) operates on the same intuition but leverages raw unlabeled corpus data instead of labeled training data. This model is a simple application of similar models from other verb-prediction tasks (Lenci, 2011; Zarcone et al., 2012). It assumes that the verb a particular direct object occurs with in a corpus (as a direct object) will also be the most likely coercion interpretation. More specifically, the COOC model predicts the top corpus verb for a specific direct object.

Finally, we define the **Prototype vector** model (Chersoni et al., 2017) (PROTO). For a given direct object, instead of predicting the top verb by co-occurrence in a corpus, average word vectors for the top k verbs that the direct object co-occurs with into a prototype vector P, and predict the closest verb in that vector space to P. Here we use pretrained word2vec (Mikolov et al., 2013) vectors from the *gensim* implementation in Python.

All three models rest on the assumption that a specific direct object (type) will predict a single recurrent interpretation. This approach has the disadvantage that it is unable to predict different interpretations for different tokens of the same direct object (such as 'make' vs 'drink' for *finish the coffee*). As a point of comparison to the models above, we furthermore used a large language model (BERT;

Devlin et al., 2019) to predict based on the context of the entire sentence rather than the direct object alone, thus allowing for different interpretation for different tokens. We adapt BERT as a model for coercion interpretation by treating the fill-in-the-blank position as a masked token to be predicted. BERT yields a distribution of relative confidences for each item in its vocabulary when used for this task. This means for each sentence, we define our **BERT** model to predict the top verb from the top *k* items in this distribution.

3.3 Experimental set-up

For this study, we split the 300 sentences in our dataset randomly into 150 training and 150 testing sentences. Within the 150 test items, we evaluate the accuracy of the models in predicting the response (either a single verb or a blank). To assess whether model performance is the same for the different cases as discussed in §2, we also report the accuracy scores for three salient groups of test items: items with a 'blank' top response ($n = 69$), cases with a non-blank top response at or above the median of 55% consensus (High Consensus, or HC, $n = 41$), and cases with a non-blank top response below the median consensus (Low Consensus, or LC, $n = 40$). In line with Roberts and Harabagiu (2011), we skipped predictions of semantically general verbs like *have* and *say*.

4 Modelling dominant verb responses

4.1 Results & Discussion

Accuracy for each model is reported in Table 1 as the **-T** ([T]op verb predicting) models. We expect models to be unable to predict **Blanks**, as they are defined to find the top-ranked verb. Interestingly, we observe that EBL performs well on this subset of the data at .870 accuracy, compared to near-zero scores for the other models. EBL's high performance, however, seems to be an artefact of data scarcity: many direct objects in the test set do not exist in the (small) training set and therefore cannot be predicted for, so EBL performs well by accident rather than by design. The non-zero scores for the other models can similarly be attributed to a few cases in which no verb among the co-occurrence data or top model predictions could be found.

	Overall	Blanks	HC	LC
EBL-T	.620	.870	.512	.300
COOC-T	.233	.058	.439	.317
COOC-B	.480	.710	.293	.275
PROTO-T	.200	.000	.488	.250
PROTO-B	.333	.623	.073	.100
BERT-T	.327	.029	.731	.425
BERT-B	.493	.435	.682	.400

Table 1: Accuracy for 4 models and variants for the entire dataset as well as the three subsets.

Turning to the **degree of consensus** next, we see, that all models perform worse on LC items than on HC items. In line with our analysis of LC items in §2, we take this difference to be indicative of the difficulty of predicting a single specific verb when LC items may have underspecified (or possibly ambiguous) meanings. Given that other datasets use deliberately high-consensus items, we believe this furthermore illustrates the challenge of modelling when using a more naturalistic sample of complement coercion sentences. For the LC sentence in example (5) each model makes a different prediction (COOC: *plant*, PROTO: *produce*, BERT: *grow*) all of which are incorrect for the dominant answer *sow* (given by 8/20 participants). The comparable closeness of some answers, however, suggests that rethinking the modeling task might be insightful, which we will do in §6.

(5) ...for a fall crop. Start ___ seeds in pots in early to midsummer, setting out six- to eight-week-old transplants in late summer or early fall in full sun and enriched soil.

Finally, focusing on a between-model comparison, we note that BERT outscores the other models on both HC and LC items. We believe this is because (1) BERT doesn't suffer from data scarcity as much as the other models by being trained on larger amounts of data, and (2) it is a token level model and can thereby make different predictions for sentences with the same direct object. This latter property leads BERT to correctly distinguish cases like *Nikolai finished a piece of stewed rabbit* (*eat*) from *Annie finishes the piece, lowering the bow* (*play*), where EBL predicts *see* in both cases, and the other models predict *take*. The fact that the use of contextualized, token-level representations leads to an increased performance on non-Blank responses suggests that information beyond direct-object noun is relevant in establishing the inferred event, where there is one (i.e., for the HC and LC cases).

Sentence	Responses	Predict top verb (§4; **-T**)	Allow for blanks (§5; **-B**)	Predict broad sense (§6; **-S**)
Lordier began the painting with a very light sketch...	BLANK (12), draw (2), redecorate, sketch, create, brush, paint	draw (11%)	BLANK (63%)	BLANK (63%)
Those so inclined can start the meal with vodka and tonic...	eat (7), BLANK (6), have (5), pair	eat (37%)	eat (37%)	CONSUME (63%)
She finished his back, then rearranged the sheet to do his legs. The top half of him was loose as a fish...	massage (11), BLANK (4), stretch (2), do (2), cover	massage (55%)	massage (55%)	OTHER (80%)
Although she was in residence for only about ten months she probably completed as many as ninety vases.	make (5), sculpt (3), BLANK (3), build, craft, create, shape	make (33%)	make (33%)	CREATE (73%)

Table 2: Example sentences, responses, and gold-standard answers for each dataset

5 Modelling blank responses

In a way, the approach taken in §4 set the models up to fail, as they have no mechanism to recognize that in cases such as (1-a) and (1-c), the interpretation does not 'need' to be resolved. In the remainder of this paper, we present two simple steps in the direction of broadening models' ability to interpret these cases in a way that is in line with their analysis as presented in §2. First, in this section, we update the models to explicitly predict the label to be 'blank' for items whose dominant response was blank. Then, in §6, we consider low-consensus items as cases where a broad sense is available for the event interpretation, but no specific verb resolution is necessary. These two strategies differ significantly from other datasets that approach this issue. Our changes in the nature of the correct label are illustrated in Table 2.

5.1 Updating models to predict 'blanks'

If we accept 'blanks' as valid modal participant responses to complement coercion interpretation cases, we need to allow models of complement coercion to have decision mechanisms to predict that response. For all unsupervised models, it is relatively straightforward to extend the unsupervised models by building in a threshold of confidence below which the models predict a blank response, reflecting the intuition that there is no single good verb the model can predict in response to the item. We call these models as a group **-B** ([B]lank predicting) models. As each model uses a different type of metric, applying a confidence threshold looks different for each one. We tuned specific thresholds by maximizing overall accuracy on our training set.

For the COOC model, for a given direct object and all uses of a verb with that direct object in the corpus, calculate the percentage p of all uses comprised by the top verb. If p is above a set threshold k, predict the top verb. Otherwise, predict a blank. We report results for an optimal k=12%. For the PROTO model, we build in a threshold based on the cosine similarity of the prototype vector to its nearest verb neighbor. If the similarity exceeds k, predict the verb corresponding to that nearest neighbor. Otherwise, predict a blank. (Optimal k=0.79). Finally, for BERT, we currently predict the verb with the highest confidence from BERT's masked prediction. To build in a threshold, we manually limit the number of predictions for the blank to check – if no verb is found in the top k words, predict a blank instead (Optimal k=5).

5.2 Results & Discussion

Table 1 presents the results for the blank-predicting models on the rows marked as **-B**. The addition of a tuned threshold mechanism to predict blanks improves accuracy on the Blanks subset (and thereby on the Overall accuracy) for all three models. For example, on the sentence *Some people with entrepreneurial spirit are still starting ___ farms*, all three blank-predicting models correctly predict a blank where their original versions incorrectly attempted verb responses.

However, this improvement comes at a cost for

the COOC model (a drop from .439 to .293 for HC and .317 to .275 for LC items) and the PROTO model (a drop from .488 to .073 for HC and .250 to .100 for LC items). Looking at the model predictions, the tuned threshold leaves these two models with a very good recall for predicting Blank responses, but a low precision: both models mark many cases that have a dominant verb response as 'blanks'. For instance, in *The others are already finishing their granola* the dominant response *eat* is correctly predicted by the original COOC model, but the updated model wrongly predicts a blank. The decrease in performance on HC and LC cases is much smaller for BERT, with accuracies dropping only from .731 to .682 (HC) and from .425 to .400 (LC). However, BERT only predicts 43.5% of the Blank items correctly, compared to 71% (COOC) and 62.3% (PROTO) of cases.

What we take this to mean is that sentences with a dominant 'blank' response have a particular contextual profile: they may have different kinds of direct objects, or they may contain more adjunct phrases of the kind of *with a very light sketch* in example (1-a). Such properties could make models recognize comparably reliably that the interpretation should be a blank. (A further investigation of these contextual properties is left for future research.) Simple unsupervised models overgeneralize that blank-prediction, but for BERT the explicit prediction of blanks comes at a comparably low cost, suggesting that there is contextual signal that correlates with participant responses being dominantly 'blank'. We take this to be converging evidence for the coherence of a group of 'blank-dominant responses' as a distinct type of complement coercion responses.

6 Modelling low-consensus responses

We next consider expanding the interpretation of our models to better handle low-consensus cases. Just as we modelled 'blank' interpretations by expanding the possible predictions of models, we adapt our task to low-consensus cases by changing the possible predicted classes.

One practical problem with these cases is that the task of predicting a single verb might penalize a model which guesses a different but semantically very similar token from the correct top response. For example, in sentence (6), our BERT-B model incorrectly guessed *construct* instead of the top response *build* given by 7/19 participants. Although intuitively these answers both involve creation of an object through handiwork, our dataset judges one correct and one incorrect.

(6) Amtrak has recently announced that it will begin ___ a high-speed rail system connecting New York, Boston, and Washington, D.C., in 1998.

In the case of these low consensus sentences, predicting a single verb might be too restrictive. Instead, we consider a second change to the evaluation and models. Namely, we replace individual verb answers with broad senses of meaning that cover a shared property of multiple verb responses across items.

A few approaches have previously modelled the concept of broader senses in complement coercion interpretations. For example, Shutova and Teufel (2009) clustered many possible interpretations to short verb+object phrases. For the pair *finish video*, this includes *film, shoot, take, produce, make...* as one cluster, *watch, view, see, examine...* as another, and *edit, cut, redact, screen...* as a third. Models were then evaluated on how closely their unsupervised clustering of the same items matched annotator clusters. Our modelling differs from this unsupervised clustering in two key ways. First, we pre-define senses that apply across many coercion phrases, rather than creating clusters for specific verb+object combinations. Second, we test predictions on individual sentences rather than predicting one cluster over all possible sentences for ambiguous phrases.

6.1 Updating models to capture broad senses

Modeling responses with broad meaning senses requires two updates: (1) reworking our dataset where correct answers consist of broad senses, and (2) updating our models to be able to predict these new classes. Among the responses to our elicitation task, we found two coherent groups of broad senses of verbs: verbs that involve some form of creation (CREATE; e.g., *make, build, write*) and verbs involving some form of consumption (CONSUME; e.g., *read, eat, drink, watch*). All unique responses were manually assigned to one of these two groups, or tagged as OTHER (e.g., *destroy, massage, hold*) if they didn't belong to either group. For each sentence, we then define the "correct" broad sense as the most popular broad sense category among all participant responses to the sentence.

In order to update our models so that they predicted broad senses rather than individual verbs, we needed a procedure to map a model's single verb prediction into a broad sense category. We used the hand-labels of senses for the gold standard answers to automate the prediction of a broad sense based on a model's originally predicted verb. For each category, we combined word vectors for all tagged examples into a single averaged vector representing that sense. When a model would predict a single verb V, it instead predicts the category whose average vector is closest to the vector for V. In this way, we update the verb predicting models to instead predict broad sense labels.

6.2 Results & Discussion

We report performance of each model in Table 3. We used the same training/test split as in the original dataset, and kept all sentences where the dominant response was a 'blank.' As in §4 and §5, we report the accuracy overall and on the three subsets of the data. This means the gold-standard for our updated dataset includes only 4 possible classes to predict: the 3 broad senses CREATE, CONSUME, OTHER, as well as BLANK.

Given the different inventory of categories, a direct comparison with the results in §5 is not possible; instead we compare relative increases of performance across models and subsets of the data. Because the broad sense predictions are derived from the verb prediction of the same models, we do note that sentences where a -T or -B model predicted incorrectly but a -S model predicted correctly are cases where the model predicted the correct sense but failed to match the exact verb. Sentences where all models were incorrect are sentences where the prediction belongs to an entirely different event sense (e.g., predicting *cook*/CREATE when the gold standard is *eat*/CONSUME).

High-consensus vs low-consensus: For BERT and EBL, the two best performing models on the broad sense dataset, we remark that the HC and LC cases show similar accuracies. In contrast to other modelling in §4 and §5, where accuracy was low for low-consensus items in particular, the broad sense prediction task shows less difference between the categories.

This improvement for low-consensus cases can be attributed to low-consensus sentences which received incorrect single verb predictions on the previous task but now received the correct broad interpretation. For example, in sentence (7), BERT-B predicts *draw* which was incorrect for the verb prediction task (correct: *write*, 6/19 participants). However, *draw* is categorized under the sense CREATE for the broad sense task, which is correct with 11/19 responses.

	Overall	Blanks	HC	LC
EBL-S	.673	.870	.561	.450
COOC-S	.487	.710	.317	.275
PROTO-S	.393	.623	.219	.225
BERT-S	.547	.435	.634	.650

Table 3: Accuracy for 4 models on the modified broad-senses dataset as well as its three subsets.

(7) McGruder began ___ his politically charged hip-hop comic strip for his college newspaper

The close performance for HC and LC cases on the broad senses task supports our intuition that many coercion sentences involve broad sense interpretation. We suggest that individual verb paraphrases for interpretations may be an artefact of tasks that prompt specific verb responses. That is, participants may be able to provide specific verb interpretations when prompted, but outside of the context of the elicitation task only resolve the broad sense and leave the specific nature of the event unspecified.

7 Discussion

By reframing the possible classes predicted in a coercion interpretation task, we break from the typical paradigm of considering complement coercion as analogous to verb paraphrase, that is: as an obligatory (semantic) resolution of a particular event. In redefining the computational task we also reconsider how complement coercion has traditionally been represented – as a type-shift from object to event, or a pragmatic process of interpreting a relevant scale. We acknowledge that neither of these accounts is formulated for modeling interpretation as a predictive task, and as such our work does not constitute a full comparison of all aspects of these accounts. Nonetheless, the fit with these accounts differs for the property at issue in our work, namely whether event resolution is obligatory.

The type-shifting account of Pustejovsky and

Bouillon (1995) forwards that each direct object contains within its lexical-semantic qualia the possible event interpretations for a coercion sentence. This translates readily to a single-verb prediction task. Under a generous reading, we could broaden this account to explain cases like (6) where multiple semantically similar verbs (e.g., *construct, build*) make valid interpretations. That is, we can rethink the type-shift accounts to consider broad event senses rather than specific paraphrases, just as we did for low-consensus cases in §6.

Still, the prevalence of sentences where a blank response was dominant suggests there are many coercion sentences where a verb paraphrase is difficult to access or absent altogether. Building models that predict blanks is difficult to link to a type-shifting account, which builds on the assumption of complement coercion as a process of event interpretation based on the direct object noun. In contrast these examples are well covered by the scalar interpretation account of Piñango and Deo (2016) which does not necessitate an event interpretation.

Although our work is not intended as a model of the psycholinguistic result that complement coercion incurs processing cost, we remark that most experimental work has used stimuli which are deliberately high-consensus constructed examples. Our present work illustrates the breadth of complement coercion sentences and outlines three general patterns – high-consensus, low-consensus, and majority blank responses – only the first of which is represented in experimental stimuli.

Noteably, Frisson and McElree (2008) investigate the effect of response consensus on processing cost, finding no difference in cost for reading sentences with high vs. low consensus. This finding is used to show that ambiguity between interpretations does not modulate the processing cost. Our introduction of low-consensus cases which share a broad sense complicates this picture by suggesting that not all low-consensus coercion sentences involve ambiguity between broad senses. As well, even norming work from Frisson and McElree (2008) forced participants to choose a verb, leaving potential blank cases unexplored. While our findings do not make predictions for the processing cost observed in experimental work, they suggest potential new classes of experimental stimuli for future work in the form of low-consensus items with a single broad sense and blank-dominant cases.

Overall, the broad spectrum of coercion examples covering multiple sub-classes illustrates that the process of interpretation goes beyond selecting a single appropriate verb paraphrase. Indeed, the presence of many blank responses suggests that it may go beyond event interpretation as well. As such, our work suggests that using naturalistic data and analyzing the semantic-pragmatic properties of observed cases is critical to developing a more complete insight into a phenomenon like complement coercion.

References

Emmanuele Chersoni, Alessandro Lenci, and Philippe Blache. 2017. Logical metonymy in a distributional model of sentence comprehension. In *Sixth Joint Conference on Lexical and Computational Semantics (* SEM 2017)*, pages 168–177.

Mark Davies. 2009. The 385+ million word corpus of contemporary american english (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, 14(2):159–190.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Yanai Elazar, Victoria Basmov, Shauli Ravfogel, Yoav Goldberg, and Reut Tsarfaty. 2020. The extraordinary failure of complement coercion crowdsourcing. In *Proceedings of the First Workshop on Insights from Negative Results in NLP*, pages 106–116.

Steven Frisson. 2009. Semantic underspecification in language processing. *Language and Linguistics Compass*, 3(1):111–127.

Steven Frisson and Brian McElree. 2008. Complement coercion is not modulated by competition: Evidence from eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(1):1.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Maria Lapata and Alex Lascarides. 2003. A probabilistic account of logical metonymy. *Computational Linguistics*, 29(2):261–315.

Alessandro Lenci. 2011. Composing and updating verb argument expectations: A distributional semantic model. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 58–66.

Brian McElree, Matthew J Traxler, Martin J Pickering, Rachel E Seely, and Ray Jackendoff. 2001. Reading time evidence for enriched composition. *Cognition*, 78(1):B17–B25.

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*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

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

Maria Mercedes Piñango and Ashwini Deo. 2016. Reanalyzing the complement coercion effect through a generalized lexical semantics for aspectual verbs. *Journal of Semantics*, 33(2):359–408.

James Pustejovsky and Pierrette Bouillon. 1995. Aspectual coercion and logical polysemy. *Journal of Semantics*, 12(2):133–162.

Kirk Roberts and Sanda Harabagiu. 2011. Unsupervised learning of selectional restrictions and detection of argument coercions. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 980–990.

Ekaterina Shutova and Simone Teufel. 2009. Logical metonymy: Discovering classes of meanings. In *Proceedings of the CogSci Workshop on Semantic Space Models*, pages 29–34. Citeseer.

Testable. testable.org: One-stop solution for behavioral experiments, surveys, and data collection.

Matthew J Traxler, Martin J Pickering, and Brian McElree. 2002. Coercion in sentence processing: Evidence from eye-movements and self-paced reading. *Journal of Memory and Language*, 47(4):530–547.

Cornelia Maria Verspoor. 1997. Conventionality-governed logical metonymy. In *Proceedings of the Second International Workshop on Computational Semantics*, pages 300–312. Citeseer.

Alessandra Zarcone and Sebastian Padó. 2010. "i like work: I can sit and look at it for hours" type clash vs. plausibility in covert event recovery. *Proceedings of Verb 2010*, page 209.

Alessandra Zarcone, Jason Utt, and Sebastian Padó. 2012. Modeling covert event retrieval in logical metonymy: probabilistic and distributional accounts. In *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2012)*, pages 70–79.

A Appendix - Corpus Extraction Heuristics

For extracting likely complement coercion candidates in §2, we used a series of heuristics to narrow from corpus sentences to coercion candidates.

First, we extracted all sentences which used one of five aspectual verbs (*begin, complete, end, finish, start*) if the sentence also included an overt direct object. We then eliminated uses which included an overt complement verb (*I had just finished washing the dishes*). This left 44,810 aspectual verb sentences from the corpus. We further removed sentences at the beginning or end of a passage in the corpus, i.e., sentences where we could not present at least one other sentence of context on either side. This left 41,372 aspectual verb sentences.

Next, we used information about the direct object in the ontological database WordNet to remove non-coercion uses of aspectual verbs (Miller, 1995). While we suspect there is common ground between many aspectual verb uses regardless of direct object type, other work leaves out specific direct objects as separate senses (Verspoor, 1997; Elazar et al., 2020). We choose to narrow the field here for maximum analogy to past work. Specifically, we removed all sentences where the direct object had no extension which was a physical entity. This included unclear cases where a physical meaning was possible but not certain – for example, "work," "school," or "company" and any direct object ending in *-ion*, those being event nouns. We also removed sentences where the aspectual verb *start* took a direct object with a "motor vehicle" sense in WordNet, (e.g., *...start the car/engine*). This left 5,088 candidate coercion sentences.

Finally, we removed sentences with a particle, e.g., (*finish* up *the tea*). While these sentences do resemble complement coercion in most respects, we expected they would introduce grammaticality issues with the fill-in-the-blank paraphrase task, potentially discouraging paraphrases when a clear interpretation was available.

This process left 4,583 sentences where an aspectual verb takes a clear entity object, resembling complement coercion by all definitions in the literature. Of these, we randomly selected 300 to use for crowdsourcing.

B Appendix - Dataset Materials

B.1 Materials

For our crowdsourcing experiment, we recruited online via Testable, under approval from **Anonymous Institution**. Participants were paid at a rate of $15CAD per hour, for annotating 50 items taking approximately 20 minutes.

Participants were initially given the following instructions asking them either provide a verb paraphrase or leave a blank (boldface as presented to participants):

> In a sentence like "The thirsty athlete finished a bottle of water," we know that the athlete drank the bottle of water, even though the verb "drinking" is not present. We are interested in sentences where such "silent verbs" are and aren't present.
>
> In this survey, you will be shown sentences which may or may not have this kind of silent verb meaning. We have added a blank line to the sentence where a verb might go if available. You will have the option to "**fill in the blank**" to make the meaning explicit or characterize the event occurring. For example, given sentence (A)...
>
> (A) The thirsty athlete finished ___ a bottle of water.
>
> ...you might choose to fill in the blank with "drinking" Given sentence (B)...
>
> (B) The construction company completed ___ a new condo.
>
> ...you might choose to fill in the blank with "building" **If you choose to fill in the blank, please fill it with a single verb with an "-ing" ending.** Some sentences might not have any reasonable input. In these cases, **you may leave the input blank.** For example, in (C)...
>
> (C) I began ___ the day by stretching.
>
> ...the sentence is fine as is, and doesn't necessarily imply a specific verb. You might choose to leave the sentence blank if you cannot think of any reasonable verb. Don't spend too much time on any one item – your gut feeling is most important. If you don't think of any verb after a few seconds, leave it blank and move on to the next question.

Items were presented in groups of 5. Participants were reminded of the instructions after every block of 5 items. An example of 5 items as displayed in a browser is shown in Figure 1.

B.2 Example items and responses

In this section we include 3 item examples from each of the 3 categories discussed in §2: high-consensus (top response above median consensus), low-consensus (top response below median consensus), or blank (top response was a blank).

B.2.1 High-consensus examples

(8) All this attention The Third Policeman is getting would've stunned its author. He finished ___ the book in 1940 at the dawn of World War II. Bad timing for a comic novel.
TOP RESPONSE: *write*, 100% (15/15)
ALL RESPONSES: writing (15)

(9) Yet a closer look reveals subtle touches of Sikes' brush. He finished ___ the walls in an aged plaster texture in warm shades of light gold and gray. He marbleized a pair of columns in similar neutral tones yet made them pop with metallic gold accents.
TOP RESPONSE: *paint*, 93.3% (14/15)
ALL RESPONSES: painting (14), building (1)

(10) Have I done something before 9/11 or after? When did I start ___ guitar? That was after.
TOP RESPONSE: *play*, 73.7% (14/19)
ALL RESPONSES: playing (14), learning (4), practicing (1)

B.3 Majority blank examples

(11) The peace has proven lasting, much to WIPNET's surprise. Gradually Liberia has started ___ the long road back from war. The country was absolutely devastated by 13 years of war, says Bushkofsky.
TOP RESPONSE: *BLANK*, 70.0% (14/20)
ALL RESPONSES: BLANK (14), taking (2), recovering (2), walking (1), building (1)

(12) Ants far exceed human beings in nastiness, Wilson has written. If ants had nuclear

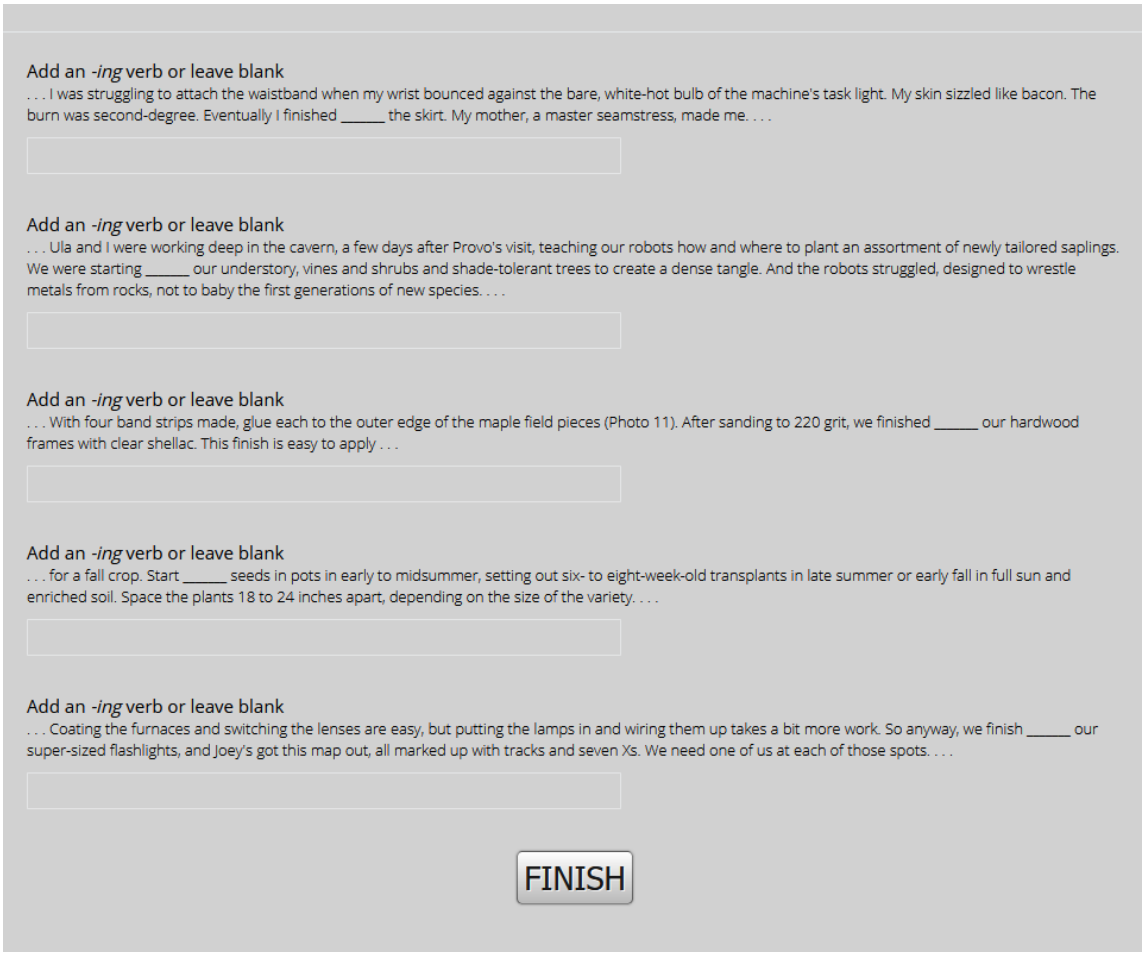

Figure 1: Screenshot of 5 items as displayed to a participant during annotation.

weapons, they would probably end ___ the world in a week. He told me there were only 20 people in the world who knew enough to identify and classify ants...
TOP RESPONSE: *BLANK*, 65.0% (13/20)
ALL RESPONSES: BLANK (13), destroying(4), annihilating(1), fighting(1), living(1)

(13) ...is the similar-size Keystone of Hercules. We complete our ___ circuit around the rim of the sky by looking southwest. Here dramatic Scorpius is well past its prime height, but it's still not too late for good looks at twinkling Antares and other illustrious Scorpius treasures.
TOP RESPONSE: *BLANK*, 65.0% (14/20)
ALL RESPONSES: BLANK (14), building (2), developing (1), doing (1), making (1), walking (1)

B.4 Low-consensus examples

(14) Erica became unconscious immediately. The technicians completed ___ the X-ray. Despite the Portlock's concerns, the technicians told them it was okay to take Erica home, even though she was still unconscious.
TOP RESPONSE: *take*, 30.0% (6/20)
ALL RESPONSES: taking (6), BLANK (5), analyzing (3), scanning (2), examining (1), making (1), performing (1), running (1)

(15) "He must like you a lot." Tyla finished ___ Julienne's hair, went to pick up her half-boots, and knelt to put them on her. Julienne was smiling, a dreamy, private softening of her lips.
TOP RESPONSE: *braid*, 21.1% (4/19)
ALL RESPONSES: braiding (4), combing (4), tying (3), brushing (2), BLANK (1), cutting (1), doing (1), making (1), styling (1), weaving (1)

(16) Painting outdoors allows me to capture light and color with much greater accuracy, he notes. He may complete ___ a piece during his first outing or return to the location the next day, but he often finishes paintings in his studio, as he did with Carmel Mission Bell Tower. Durborow especially enjoys the friendly competition and camaraderie of paint-outs, where artists work together on location, and he attends at least four such events a year.
TOP RESPONSE: *paint*, 33.3% (5/15)
ALL RESPONSES: painting (5), BLANK (3), drawing (3), assembling (1), building (1), making (1), writing (1)