Cococorpus: a corpus of copredication

Long Chen **Deniz Ekin Yavaş** **Laura Kallmeyer** **Rainer Osswald**
Heinrich Heine University Düsseldorf
{chen.long,deniz.yavas,laura.kallmeyer,rainer.osswald}@hhu.de

Abstract

While copredication has been widely investigated as a linguistic phenomenon, there is a notable lack of systematically annotated data to support empirical and quantitative research. This paper gives an overview of the ongoing construction of Cococorpus, a corpus of copredication, describes the annotation methodology and guidelines, and presents preliminary findings from the annotated data. Currently, the corpus contains 1500 gold-standard manual annotations including about 200 sentences with copredications. The annotated data not only supports the empirical validation for existing theories of copredication, but also reveals regularities that may inform theoretical development.

1 Introduction

Inherently polysemous nouns have multiple meaning facets. For example, the noun 'breakfast' has an object facet referring to its food reading, and an event facet referring to its dining reading. It is common to analyze its semantic type as a "dot-type" *food • event* (Pustejovsky, 1995; Cruse, 1995). Different meaning facets of a dot-type noun can be predicated by multiple predicates at the same time. This phenomenon is referred to as *copredication* (Pustejovsky, 1995; Asher, 2011). In (1), the verb 'bring' targets the object facet while 'late' targets the event facet.[1]

(1) Go *bring* your new sister some *late* **breakfast**.

Coercion, by contrast, refers to the phenomenon where the predicate targets a facet which is not available in the noun. (2) is an example of coercion. The verb 'resist' targets an event facet, but 'novel' is an *object • info* noun without event facets.

(2) I *resisted* a second **novel** for 14 years until Jack became a way out of a trap I got myself into with a multi-book contract.

Copredication has been studied extensively in linguistics. For example, a number of studies focus on the restrictions and the orders of copredication. Asher (2011) observed an asymmetry in the copredication of the polysemous noun 'city'. Retoré (2014) noticed that *football team* reading of the polysemous noun 'Liverpool' cannot copredicate with other readings. Chatzikyriakidis and Luo (2015) concluded that the copredication of 'newspaper' related to the *organization* reading is relatively restricted compared to its other facets. Ortega-Andrés and Vicente (2019) proposed the concept of 'activation package' to explain the possibility of the copredication over 'school'. Sutton (2022) discovered that the *physical entity* and *eventuality* reading of 'statement' cannot cooccur in a copredication construction, but either reading can copredicate with the *informational content* reading. Murphy (2024) claimed that complexity and coherence are the decisive factors for the predicate order within a copredication. Michel and Löhr (2024) further suggested that context is a more fundamental factor and explained the order of copredication with the notion of "expectation". Chen et al. (2025) proposed a distinction between 'primary facets' and 'secondary facets' to explain the asymmetry in the copredication of *food • event* nouns. However, most of these previous work is based on a small number of introspectively constructed sentences or cherry-picked typical cases, and there is little quantitative research on copredication to prove or disprove the proposed theories. Also, previous work mainly focused on prototypical copredication

[1] The example sentences in the paper are all taken from our annotated data, which come from BookCorpus (Zhu et al., 2015), accessed via https://huggingface.co/datasets/bookcorpus/.

Proceedings of the 21st Joint ACL - ISO Workshop on Interoperable Semantic Annotation (ISA-21), pages 31–40

instances, while a lot of borderline cases are not well-represented.

These limitations show the need of an annotated corpus of sentences with copredication. Currently there are few corpora related to copredication. Hanks and Pustejovsky (2005) created a lexicon with corpus usage patterns of words with semantic types information. However, the focus is on verbs instead of polysemous nouns and copredication constructions are not specifically addressed. Alonso et al. (2013) annotated in total 4500 sentences from three languages containing regular polysemous nouns, but most of the sentences contain single predications instead of copredications, and copredication is only treated as a kind of underspecification. Another valuable resource focusing on semantic types targeted by predicates is T-PAS (Typed Predicate Argument Structures; Jezek et al. 2014). T-PAS offers argument structure patterns for Italian verbs, annotated with semantic types. It also provides corpus instances for each verb pattern. However, although T-PAS includes a range of semantic types and is not limited to simple type nouns, instances involving dot-types are relatively infrequent, and the focus is on single-type predications. To our knowledge, no corpus to date has a specific focus on copredication.

In this paper we describe the construction of Cococorpus, an ongoing project that aims at a corpus with copredication and coercion annotations. The current version is restricted to English. Up to now, we have mainly been targeting the annotation of copredication, covering three kinds of dot-types, and we have manually annotated about 1500 gold-standard sentences from BookCorpus (Zhu et al., 2015) where a polysemous noun is predicated by multiple predicates, including around 200 copredication sentences. Cococorpus also contains around 18000 silver-standard sentences acquired through automatic annotation.

2 Annotation overview

For our copredication annotation, we focused on three common dot-types in language: *food • event*, *info • event*, and *object • info*, which are related to the three major ontological categories *phys(ical)-obj(ect)*, *information*, and *event*.[2] These dot-types are selected because they are relatively better-studied and their facets are rather easy to distinguish from each other.

[2] *food* is a subtype of *phys-obj*

For each dot-type, five nouns with relatively high frequency and little ambiguity are selected. The selected nouns are listed in Table 1. At the moment, we focus on the copredication construction V+Adj+N for annotation. The source of our data is BookCorpus (Zhu et al., 2015), not only because it is free and easily accessible with a considerable number of sentences, but also because of the diversity of the genres of the texts and the contemporary, naturalistic language of the texts.

The annotation pipeline includes the following stages:

1. Automatic extraction of target constructions containing the selected nouns
2. Preliminary annotation using the classifier from Yavas et al. (2023)
3. Manual annotation by two trained annotators
4. Disagreement resolution through discussion

More specifically, the annotation focuses on the relationship between each selected noun and the adjective or verb in the sentence that predicates the noun. For example, in (3), the relation between 'ate' and 'breakfast' would be annotated as ARTIFACT,[3] and the relation between 'quick' and 'breakfast' would be annotated as EVENT.

(3) They packed their bags, *ate* a *quick* **breakfast** of dry cereal and headed south.

Our labeling scheme primarily follows the original labels of the classifier. The basic facet selection labels consists of ARTIFACT for the predication over the object facet, INFORMATION for the predication over the info(rmation) facet, and EVENT for the predication over the event facet. To account for cases where a predicate simultaneously targets multiple facets,[4] we added four composite labels: ARTIFACT_INFO, ARTIFACT_EVENT, EVENT_INFO, and ARTIFACT_EVENT_INFO. Furthermore, coercion is distinguished from facet selection during human annotation, so four additional labels are employed: COERCION_ARTIFACT, COERCION_EVENT, COERCION_INFO and COERCION_OTHER, which indicate

[3] This label stands for the object facet. It comes from the T-PAS taxonomy, from which our classifier has been trained on. In this paper the annotation labels are presented in small capitals, and semantic types are presented in italics.

[4] e.g. in 'read a book', 'read' targets both the object facet and the info facet of 'book'.

Dot types	Selected nouns
food • event	breakfast, buffet, dinner, feast, meal
info • event	conversation, lecture, response, speech, submission
object • info	brochure, diary, novel, summary, textbook

Table 1: The selected nouns of each dot-type

coercion to the object facet, event facet, info facet, and other facets, respectively. Although theoretically, facet selection and coercion can both happen in one single predication, and coercion can involve multiple facets, our current annotated corpus does not contain such instances, and therefore the corresponding combined labels are not implemented yet. Additionally, it is usually unclear which facets light verbs target in copredication constructions, so we specifically annotate light verbs with the label LVB and exclude them in our current research, such as the relation between 'have' and 'dinner' in the light verb construction 'have an early dinner'. And for technical reasons, deleting sentences directly during annotation is not possible, so we also employ a label called DELETE to mark the exclusion of a sentence.

3 The construction of the corpus

3.1 Automatic extraction and preliminary classification

We extract sentences containing the candidate nouns from the BookCorpus and parse them using spaCy's transformer-based pipeline for English.[5] Our goal is to identify sentences where candidate nouns are the direct object of a verb (`dobj`) and modified by an adjective (`amod`) simultaneously. Sentences meeting both criteria are classified in the next step using the classifiers developed by Yavas et al. (2023) for copredication detection.

In their study, Yavas et al. (2023) develop multilingual classifiers to identify semantic argument types in both verbal and adjectival predications. They train separate binary support vector machine classifiers for several semantic types. These classifiers employ contextualized word embeddings generated by pre-trained language models, specifically the multilingual RoBERTa model (Conneau et al., 2020). Given their contextualized word embeddings in the sentences as input, the classifiers classify the relation between the predicate and its argument based on the targeted semantic type. Fig. 1 illustrates the classification process. Yavas et al. (2023) demonstrate that these classifiers effectively detect verb-adjective copredications across multiple languages, making them well-suited for our study. Specifically, we employ six classifiers corresponding to three semantic types for both verbal and adjectival predications: *information*, *event*, and *artifact*. An example of copredication detection using the classifiers is illustrated in Fig. 2.

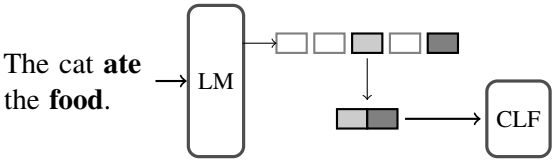

Figure 1: The figure illustrates the working principles of the classifiers developed by Yavas et al. (2023). The classifiers are trained to classify the relation between a predicate and its argument in a sentence using their contextualized word embeddings from a pre-trained language model as input.

3.2 Manual annotation

3.2.1 Annotation platform and format

The manual annotation and adjudication were conducted using the INCEpTION annotation platform (Klie et al., 2018).[6] As illustrated in Fig. 3, the annotation interface displays the automatic annotation, with each relation between a predicate and a noun represented by a labeled directional arrow. Annotators could modify the relations by selecting the corresponding arrows and adjusting the assigned labels through a drop-down menu on the right side of the page.

The adjudication interface (Fig. 4) employs a color-coded system to indicate the annotation status of the sentences. On the left side, the green cells and the white cells mark sentences that require no further modification: Green indicates that the annotators did not change the automatic annotation, and white indicates that they changed it in the same way. Red cells indicate unresolved disagreements.

[5] spacy.io/models/en#en_core_web_trf

[6] inception-project.github.io

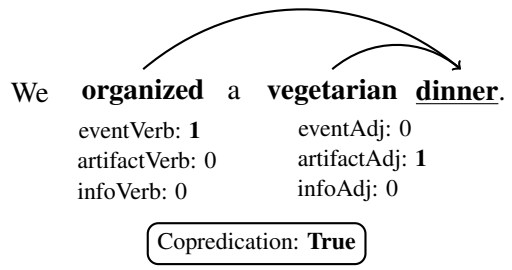

Figure 2: Classification of verbal and adjectival predications in a sentence. Copredication is detected when different types of classifiers assign positive labels to different predications.

The adjudication panel presents a comparative view of both annotators' decisions in the central display area. The adjudicator could either select one of the existing annotations or create a new annotation in the upper panel to establish the final decision.

3.2.2 Annotation guidelines

Our annotation distinguishes facet selection from argument coercion. For example, in (4-a), the predicate 'finish' targets the type event, which is not a meaning facet of the noun 'novel', so the relation between 'finish' and 'novel' is annotated as COERCION_EVENT;[7] in (4-b), although the noun 'novel' does have an info facet, coercion still occurs because according to the context, the content the person 'posted' is the metadata of the novel (e.g. description or advertisement) instead of the actual content, so the relation should be annotated as COERCION_INFORMATION instead of INFORMATION.

(4) a. I really have to *finish* this current **novel** before researching another one.
b. I have *posted* the new **novel** Rome's Evolution on Amazon, B & N, Kobo and Smashwords.

Regarding facet selection, the nouns in the current annotation task are preselected, so it is clear which facets these nouns have. The major annotation task thus reduces to identifying which facets are targeted by given predicates. Operationally, we distinguish selected facets by substitution. Taking the object facet as an example, replace the dot-type noun with physical objects such as 'stone'. The phrase 'throw the stone' is felicitous but '#memorize the stone' is semantically anomalous, which proves that 'throw' targets object facets but 'memorize' cannot.

The facet selection of certain predicates is context-dependent, requiring annotators to determine the most possible targeted facets based on the context from the sentence. As a representative case, the verb 'remember' can target only the object facet of 'breakfast' (as in (5-a)) or target both the event and the object facets of 'meal' simultaneously (as in (5-b)).

(5) a. Her pleased expression tells me she likes that I *remember* her favorite **breakfast**.
b. Once passed the initial security checkpoint it occurred to Jane that she could not *remember* her last **meal**, but there were lines outside all the food stalls.

Some predicates exhibit a high flexibility in meaning facet selection or their predication involve other mechanisms instead of standard facet selection. For example, in 'love the book', any part of the book can be targeted by 'love'; in 'his own book', the adjective 'own' is more focused on the possession relation instead of the facets of 'book'. In such cases, determining the targeted facets becomes both methodologically challenging and theoretically uninformative. Therefore, these predicates were systematically excluded from our annotation. These predicates include:

- 'like' verbs: like, love, hate, prefer
- 'equal' verbs: be, equal, mean
- quoting verbs: say, mention
- adjectives describing type/token: certain, particular, single, same, such
- adjectives describing number: many, much, more, enough, some, few, extra
- adjectives describing entirety/identity: own, whole, entire, complete, actual, real, other, another
- adjectives describing quality: good, nice, fine, best, great, fantastic, wonderful, amazing, bad, awful

[7] While events such as reading and writing are often analyzed as being part of the qualia structure of 'novel', they do not count as formal quale but as telic quale (cf. Pustejovsky 1995, Sect. 6.2); that is, they do not qualify as event facets of the noun.

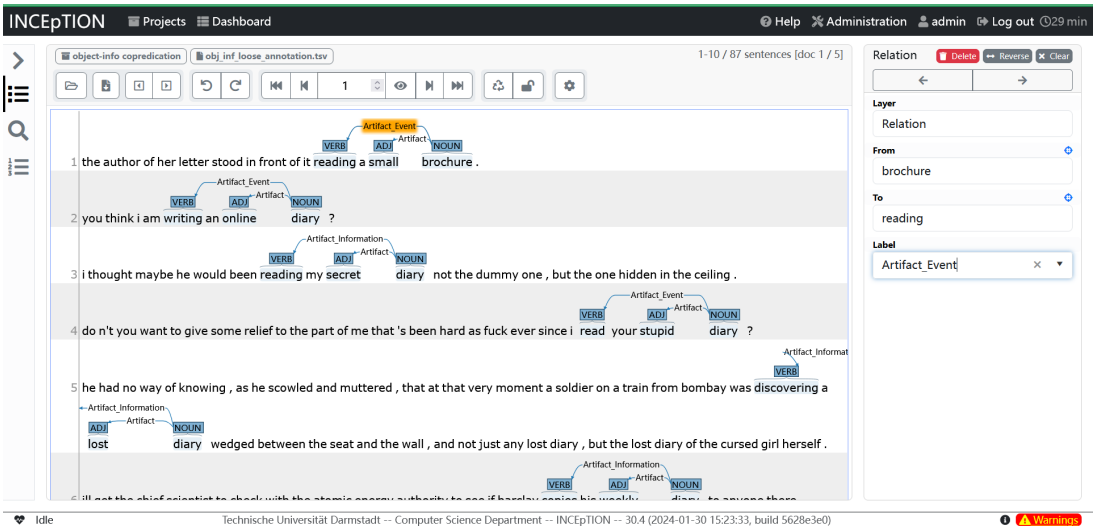

Figure 3: Annotation interface

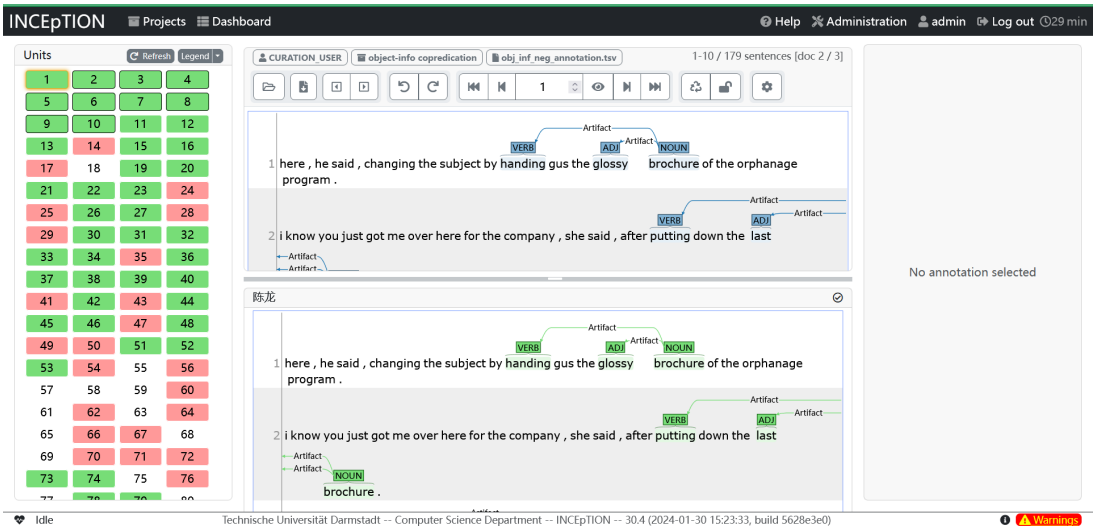

Figure 4: Adjudication interface

4 The analysis of the annotation

4.1 Statistics of the annotated data

The extraction process yielded varying numbers of candidate sentences across different dot-types. For the selected *food • event* nouns, we extracted 6838 sentences with multiple predication from BookCorpus. From this subcorpus, we randomly selected 300 'positive candidates' where the classifier detected copredication (the verb-noun relation and the adjective-noun relation are different) and 300 'negative candidates' (both predicates target the same facet). For *info • event* nouns, we got 9129 sentences with multiple predication, from which we similarly selected 300 positive and 300 negative cases for annotation. *object • info* nouns provided only 832 sentences in total due to the relatively low frequency of the five chosen nouns. Consequently, we selected only 150 'positive' and 150 'negative' candidates for annotation.

	V_{obj}	V_{ev}	V_{both}	$V_{coercion}$	Σ
A_{obj}	303	16	12	1	332
A_{ev}	108	42	17	2	169
A_{both}	2	4			6
Σ	413	62	29	3	507

Table 2: Statistics of the annotation of *food • event* nouns

4.1.1 *food • event*

Among the 600 chosen sentences of the five *food • event* nouns, 507 valid annotations were obtained. Around 100 sentences were excluded,

mainly because they involve light verb constructions or parsing errors. Table 2 reveals the frequency of the predicate tokens that target different facets of the nouns.[8] The majority of verbs (413 out of 507) target the object facet, and approximately two-thirds of adjectives (332 out of 507) target the object facet. Additionally, around 30 sentences contain predicates that simultaneously target both facets. Furthermore, three sentences were identified as involving a coercion.

(6) a. "I *promised* you guys a hot **meal**", said Ellen, sighing.
b. Her expression changes to that of a lioness *stalking* her next **meal**.

(6) presents two annotated instances of coercion. In (6-a), the verb 'promise' typically targets an event facet and the sentence means 'I promised to get you guys a hot meal'. Although the noun 'meal' has an event facet, the type of the event facet is *eating*, which is not the giving event 'promise' targets. Thus, the relation between 'promise' and 'meal' is annotated as COERCION_EVENT. In (6-b), the object of the verb 'stalk' usually needs to be an animate object rather than food, so the predication is annotated as COERCION_OTHER.

While our annotation can provide empirical evidence for some theoretical analyses on copredication, we are not yet able, due to the limited number of dot types and copredication instances, to fully validate or falsify theories related to dot objects with multiple facets (as in Asher, 2011, Ortega-Andrés and Vicente, 2019, Sutton, 2022, etc.) or specific to word items and context (as in Michel and Löhr, 2024). Murphy (2024) proposed a principle called Incremental Semantic Complexity (ISC) and concluded that in copredication, the concrete readings would come earlier than abstract readings (in linear order). It is assumed that physical objects are more concrete than information, and information is more concrete than events. This assumption is supported to a large extent by our annotation statistics on copredication over *food • event* nouns. As shown in Table 2, in most copredication instances, the verb, which precedes the adjective, is targeting the object facet. However, there are still 16 cases where the verb targets the event facet, indicating that the ISC is rather a tendency than a strict principle.

As observed in Chen et al. (2025), for *food • event* nouns in the copredication pattern V+Adj+N, the verb targeting the event facet and the adjective targeting the object facet is a preferred order. When the adjective instead targets the event facet, copredication is only possible when the adjective is *facet-addressing*. Facet-addressing adjectives, contrary to *facet-picking* adjectives, are adjectives that do not affect the availability of a facet of a dot-type noun. For example, 'quick' is a facet-addressing adjective, because 'quick lunch' is still a dot-type and can be copredicated by object-targeting verbs like 'cook' and 'order'. By contrast, 'slow' is a facet-picking adjective since the object facet is not available anymore in 'slow lunch', and '#cook a slow lunch' or '#order a slow lunch' are not acceptable. However, according to Table 2, only 16 cases align with the assumption that object facet is targeted first, whereas there are 108 cases where the copredication works in the less preferred direction. This discrepancy can be partially explained by the dominance of object-targeting verbs in the corpus.

Further analysis of the 108 copredication cases with the event-targeting adjectives reveals that there are only 23 adjective types in the 108 cases. These adjectives fall into the following four semantic categories:

- order related: *first, last, next, fourth, new*
- time related: *quick, slow, early, late, occasional*
- specialness related: *special, customary, regular, unexpected, obligatory, worthy, easy*
- other: *romantic, solitary, civilized, corporate*

Notably, all of them except 'slow' are facet-addressing adjectives. In all the copredication instances involving 'slow', the verb was always 'eat' in our dataset. We discovered in previous studies that 'eat' can also take *food • event* nouns modified by any adjectives as objects, probably due to its relatively light meaning in the context of meals. Moreover, the phrase 'eat a slow meal' could be understood as eating a meal slowly, which makes the phrase felicitous.

[8] V_{obj} stands for the cases where the verb targets the object facet; V_{both} means that the verb targets both facets; $V_{coercion}$ means the verb is annotated as having coercion. Similar interpretations apply to the other symbols of the table.

	V_{inf}	V_{ev}	V_{both}	$V_{coercion}$	Σ
A_{inf}	53	28	1	1	83
A_{ev}	14	189	2		205
Σ	67	117	3	1	288

Table 3: Statistics of the annotation of *info • event* nouns

4.1.2 *info • event*

From the initial set of 600 chosen sentences containing *info • event* nouns, only 288 yielded valid annotations. The low proportion of valid annotations is due to two factors: (1) the selected nouns are deverbal nouns and frequently occur in light-verb constructions, and (2) 'response', 'submission' and 'speech' also have other meanings unrelated to info facets. The distribution of facet selection for these *info • event* nouns is presented in Table 3.

In the 288 instances of multiple predication, there are only 42 instances (14.6%) of copredication. In 14 of the instances, the verb targets the info facet and the adjective targets the event facet; in the other 28 instances copredication works in the other order. This implies either the ISC from Murphy (2024) might be too strict or *information* and *event* are actually close to each other in terms of complexity.

The proportion of copredication over *info • event* nouns is significantly lower than that observed with *food • event* nouns. This is consistent with the observation in Chen et al. (2025) that for *info • event* nouns, both facets tend to be secondary facets, that might be inaccessible if the other facets are targeted first, and copredication in the construction V+Adj+N can only happen when the adjective is facet-addressing.

The 42 copredication instances only include 10 different adjective types. These adjectives can be classified into the following four semantic categories and they are all facet-addressing predicates:

- order related: *last*
- time related: *rapid, earlier, lengthy, little*
- atmosphere related: *bickering, heated*
- speaker related: *private, unwilling, hasty*

4.1.3 *object • info*

The distribution of predications over *object • info* nouns is presented in Table 4, derived from 253 valid annotations out of 300 candidate sentences.

	V_{obj}	V_{inf}	V_{both}	$V_{coercion}$	Σ
A_{obj}	56	2	12		70
A_{inf}	30	31	96	23	180
$A_{coercion}$	1	1	1		3
Σ	87	34	109	23	253

Table 4: Statistics of the annotation of *object • info* nouns

	V_{obj}	V_{inf}/V_{both}	$V_{coercion}$	Σ
A_{obj}	56	14		70
A_{inf}	30	127	23	180
$A_{coercion}$	1	2		3
Σ	87	143	23	253

Table 5: Updated statistics of the annotation of *object • info* nouns

This dot-type exhibits a notably higher frequency of coercion, with the coercion to the event facet being particularly predominant. In 22 of the 23 coercion instances, the verb targets an event facet, suggesting a possible tendency of the direction of coercion.

Regarding copredication, the 32 instances revealed a significant asymmetry, which is consistent with the ISC suggested by Murphy (2024) but seems contradictory to the observation in Chen et al. (2025). According to Chen et al. (2025), there is little restriction on the copredication of *object • info* nouns and copredication can happen in both orders. However, in only two cases in our annotated data, the verb targets the event facet and the adjective targets the object facet. This may be attributed to two factors. First, the frequency of info-targeting adjectives is relatively high (180 out of 253 instances). Secondly, high-frequency verbs like 'read', 'write', and 'publish' are annotated as targeting both facets, resulting in a high number in the third column of the table. Interestingly, in 96 of the 109 instances, the adjective targets the info facet, suggesting that verbs like 'read' probably "mainly" targets the info facet. If we take this into account and combine the cases where the verb targets the info facet and the verb targets both facets, the updated statistics (as in Table 5) reveals a more balanced distribution of copredications.

4.2 Disagreement analysis

The inter-annotator agreement is listed in Table 6. The primary reasons of inter-annotator disagree-

Dot types	Agreement
food • event	0.64
info • event	0.43
obj • info	0.67

Table 6: The inter-annotator agreement in Cohen's Kappa

ment can be summarized as follows.

4.2.1 The exclusion of the sentences

As is in (7), the adjective 'complete' refers to part-whole relations and can target any facets; so one of the annotator followed the guideline and labeled DELETE while the other annotator decided by context, which suggests that 'complete' targets the info facet, and annotated INFORMATION.

(7) To buy the *complete* **novel**, Trail of Storms, click here.

4.2.2 Difficult contexts

In some sentences, the predicate can target both facets of the noun, but the context provided by the sentence is insufficient or complicated, which also results in the disagreement between annotators.

(8) a. Do you *remember* that first **dinner**?
b. He was cooking a *special* **dinner** for her and he had finally found the perfect ring to put on her finger, a heart shaped diamond surrounded by smaller stones and set in platinum.

The annotation of (8-a) presented a challenge due to lack of context. The annotators labeled ARTIFACT and EVENT respectively. The adjudication process selected EVENT as the final decision, because conceptually, it is more plausible to recall a dining event while forgetting specific culinary details than remembering only the food while forgetting the associating eating event. Thus, the event facet is established as the default selection for such contextually underspecified cases in terms of remembering a meal.

In (8-b), the context provides competing clues. The verb 'cook' implies a special food preparation, while the sentential context subsequently indicates the dining experience being 'special', resulting in the divergent annotation of the relation between 'special' and 'dinner'. The final decision is that 'special' targets the event facet, as the contextual evidence provided no substantive indication of unusual food characteristics that warrants an object facet selection. This annotation example reflects the difficulty in the identification of copredication regarding predicates with a wider choice of facets.

4.2.3 Borderline light verbs

There is little consensus regarding the definition of light verb constructions, which is reflected in our annotation of high-frequency *info • event* noun patterns including 'give a speech/lecture', 'make a response/speech/conversation' and 'deliver a speech'. One annotator label them as LVB and the other treat them as regular verb phrases, contributing substantially to the relatively low agreement in the annotation of *info • event* nouns.

To resolve the disagreement, we implemented the diagnostics proposed by Fleischhauer and Neisani (2020), such as replacing the verb with its synonyms and examining the acceptability and the meaning of the new phrase. Application of these diagnostics reveal that the verbs 'deliver', 'give' and 'make' cannot be easily substituted ('#send a speech/lecture', '#produce a conversation' are ungrammatical). Consequently, the verbs mentioned above are regarded as light verbs during the final adjudication.

4.2.4 Unclear distinctions between facets

Some of the disagreement arises from the unclear distinctions between facets, especially the object facet and the info facet of *object • info* nouns.

(9) a. He tried to talk a lot about theories and make funny stories at times to let students feel like they were not drones *downloading* the latest **textbook** that the publishing company decided could be revised for the twelfth time in a row for twelve years straight.
b. During this time, she published a *short* **novel**.

The ontological status of digital texts presents a significant challenge to our annotation, as illustrated in example (9-a), which involves the predication over electronic versions of textbooks rather than traditional physical books. It is arguable whether the PDF file and the strings in the computers are a kind of physical object or more about the information.

A similar puzzle also exists with traditional paper books, as demonstrated in example (9-b). The

adjective 'short' unambiguously targets the info facet of the 'novel', but at the same time, the length of the printed characters in the physical book is also short. The printed symbols are ontologically different from the object facet of the 'novel', which is usually made of paper and consists of covers, and also different from the info facet of the 'novel', which does not have a physical form. Frequent verbs including 'read' and 'write' also have the same problem. The entity a person 'read' in a novel is not the paper material but rather the printed matter on the paper, which is not exactly the object facet of 'novel'. The statistics shown in Fig. 4 also suggests that the facet these verbs target is probably closer to info facet than object facet. Currently, the label INFORMATION is decided for both controversial cases, but these disagreements highlight a need for a comprehensive revisit of the analysis of *object • info* nouns and an investigation of the possible existence of a third meaning facet.

4.2.5 Borderline coercions

The unclear distinction between coercion and facet selection is also a reason for the disagreement between annotators.

(10) a. Anticipating an *angry* **conversation** he would not want to overhear, Mark hurried to the shower.
b. It will handle all your daily chores, provide *intelligent* **conversation** and need absolutely no maintenance.

The examples (10-a) and (10-b) exemplify a systematic pattern of human-targeting adjectives modifying *info • event* nouns, relating to the behavioral manner of the participant during the event. On the one hand, the constructions display typical features of coercion. For example, the syntactic transformation of these phrases are restricted. Transformations such as '?The conversation is angry', '#It provides a conversation that is intelligent' are marginally acceptable or unacceptable. Furthermore, these adjectives cannot easily modify the event facet or info facet of the nouns of other types, e.g. '?an angry book' requires some context to be acceptable, and '#an intelligent meeting' is infelicitous.

On the other hand, the treatment of these cases as facet selection can also be justified. First, the interpretation of the phrase is specific, unlike the typical coercion examples such as 'finish the book', where the event that is 'finished' is implicit and needs to be specified by context. Second, the usage of human-targeting adjectives for *info • event* nouns is productive. Other adjectives of this kind including 'cheerful', 'friendly', 'polite', and 'honest' modifying *info • event* nouns also exist in our annotated data.

Currently, these instances are annotated as facet selection. A more comprehensive analysis and annotation on coercion will be left for future research.

5 Conclusion and future work

The construction of Cococorpus is an ongoing project. Currently, we achieved an annotation of more than 1000 sentences with multiple predications over inherently polysemous nouns, among which 198 sentences exhibit copredication. The annotated data can serve as an empirical evidence for some linguistic analyses on copredication, such as a tendency of ISC from Murphy (2024) and the distinction between primary and secondary facets of polysemous nouns from Chen et al. (2025). The annotated data and annotation guideline are published at the project Github page (https://github.com/CoCoCo-Project).

Many aspects of copredication remain to be addressed in future annotation efforts. The current annotation framework is limited to the construction V+Adj+N, while other typical copredication constructions, such as (reduced) relative clauses and multiple adjectives, are yet to be incorporated. Furthermore, the existing coverage of dot-types and nouns is also relatively limited. We plan to expand our annotation scope to include: (1) additional dot-types and lexical items, particularly nouns exhibiting multiple facets (e.g., *school*, *city*) and those with debatable facet classifications (e.g., *annotation*); (2) a broader range of predicate types; (3) cross-linguistic investigations to examine potential variations in copredication phenomena across different languages.

Acknowledgements

This work is part of the project "Coercion and Copredication as Flexible Frame Composition" funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), project number 440934416. We would like to thank the anonymous reviewers for their valuable comments.

References

Héctor Martínez Alonso, Bolette Sandford Pedersen, and Núria Bel. 2013. Annotation of regular polysemy and underspecification. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 725–730.

Nicholas Asher. 2011. *Lexical meaning in context: A web of words*. Cambridge University Press.

Stergios Chatzikyriakidis and Zhaohui Luo. 2015. Individuation criteria, dot-types and copredication: A view from modern type theories. In *Proceedings of the 14th Meeting on the Mathematics of Language (MoL 2015)*, pages 39–50.

Long Chen, Laura Kallmeyer, and Rainer Osswald. 2025. Primary vs. secondary meaning facets of polysemous nouns. *Empirical issues in syntax and semantics: Selected papers from CSSP 2023*, page 27.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

D. Alan Cruse. 1995. Polysemy and related phenomena from a cognitive linguistic viewpoint. In Patrick Saint-Dizier and Evelyne Viegas, editors, *Computational lexical semantics*, pages 33–49. Cambridge University Press.

Jens Fleischhauer and Mozhgan Neisani. 2020. Adverbial and attributive modification of persian separable light verb constructions. *Journal of Linguistics*, 56(1):45–85.

Patrick Hanks and James Pustejovsky. 2005. A pattern dictionary for natural language processing. *Revue Française de linguistique appliquée*, 10(2):63–82.

Elisabetta Jezek, Bernardo Magnini, Anna Feltracco, Alessia Bianchini, and Octavian Popescu. 2014. T-PAS; a resource of typed predicate argument structures for linguistic analysis and semantic processing. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 890–895, Reykjavik, Iceland. European Language Resources Association (ELRA).

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart De Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th international conference on computational linguistics: system demonstrations*, pages 5–9.

Christian Michel and Guido Löhr. 2024. A cognitive psychological model of linguistic intuitions: Polysemy and predicate order effects in copredication sentences. *Lingua*, 301:103694.

Elliot Murphy. 2024. Predicate order and coherence in copredication. *Inquiry*, 67(6):1744–1780.

Marina Ortega-Andrés and Agustín Vicente. 2019. Polysemy and co-predication. *Glossa: a journal of general linguistics*, 4(1).

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

Christian Retoré. 2014. The Montagovian Generative Lexicon ΛTy_n: a type theoretical framework for natural language semantics. In *TYPES: International Workshop on Types and Proofs for Programs, April 2013, Toulouse, France*, pages 202–229.

Peter Roger Sutton. 2022. Restrictions on copredication: a situation theoretic approach. In *Proceedings of the 32nd Semantics and Linguistic Theory Conference*, pages 335–355.

Deniz Ekin Yavas, Laura Kallmeyer, Rainer Osswald, Elisabetta Ježek, Marta Ricchiardi, and Long Chen. 2023. Identifying semantic argument types in predication and copredication contexts: A zero-shot cross-lingual approach. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 310–320.

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.