Semantically Constrained Multilayer Annotation: The Case of Coreference

Jakob Prange* **Nathan Schneider**
Georgetown University

Omri Abend
The Hebrew University of Jerusalem

Abstract

We propose a coreference annotation scheme as a layer on top of the Universal Conceptual Cognitive Annotation foundational layer, treating units in predicate-argument structure as a basis for entity and event mentions. We argue that this allows coreference annotators to sidestep some of the challenges faced in other schemes, which do not enforce consistency with predicate-argument structure and vary widely in what kinds of mentions they annotate and how. The proposed approach is examined with a pilot annotation study and compared with annotations from other schemes.

1 Introduction

Unlike some NLP tasks, coreference resolution lacks an agreed-upon standard for annotation and evaluation (Poesio et al., 2016). It has been approached using a multitude of different markup schemas, and the several evaluation metrics commonly used (Pradhan et al., 2014) are controversial (Moosavi and Strube, 2016). In particular, these schemas use divergent and often (language-specific) syntactic criteria for defining candidate mentions in text. This includes the questions of whether to annotate entity and/or event coreference, whether to include singletons, and how to identify the precise span of complex mentions. Recognition of this limitation in the field has recently prompted the Universal Coreference initiative,[1] which aims to settle on a single cross-linguistically applicable annotation standard.

We think that many issues stem from the common practice of creating mention annotations from scratch on the raw or tokenized text, and we suggest that they could be overcome by reusing structures from existing semantic annotation, thereby ensuring compatibility between the layers. We advocate for the design pattern of a **semantic foundational layer**, which defines a basic semantic structure that additional layers can refine or make reference to. Some form of predicate-argument structure involving entities and propositions should serve as a natural semantic foundation for a layer that groups coreferring entity and event mentions into clusters.

Here we argue that Universal Conceptual Cognitive Annotation (UCCA; Abend and Rappoport, 2013) is an ideal choice, as it defines a foundational layer of predicate-argument structure whose main design principles are cross-linguistic applicability and fast annotatability by non-experts. To that end, we develop and pilot a new layer for UCCA which adds coreference information.[2] This coreference layer is constrained by the spans already specified in the foundational predicate-argument layer. We compare these manual annotations to existing gold coreference annotations in multiple frameworks, finding a healthy level of overlap.

Our contributions are:

- A discussion of multilayer design principles informed by existing semantically annotated corpora (§2).
- A semantically-based framework for mention identification and coreference resolution as a layer of UCCA (§3). Reusing UCCA units as mentions facilitates efficient and consistent multilayer annotation. We call the framework **Universal Conceptual Cognitive Coreference** (UCoref).
- An in-depth comparison to three other coreference frameworks based on annotation guidelines (§4) and a pilot English dataset (§5).

*Contact: jakob@cs.georgetown.edu

[1] https://sites.google.com/view/crac2019/

[2] Our annotations are available under https://github.com/jakpra/UCoref.

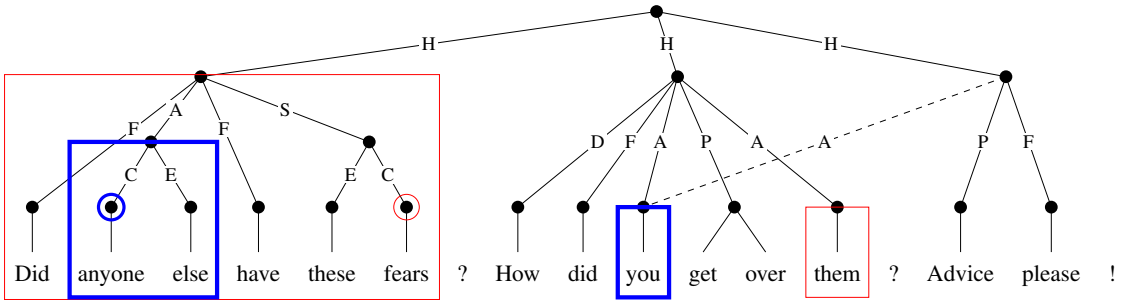


Figure 1: A foundational UCCA analysis of three consecutive sentences from the Richer Event Description corpus, with examples of coreferent units superimposed (boxes). The context is that the speaker is posting a message to a forum in which she shares her own fears and asks for advice; *you* is coreferent with *anyone else*, and *them* refers back to the whole first scene.[3] Circled nodes indicate semantic heads/minimal spans, as determined by following State (S) and Center (C) edges. In the third sentence, *Advice please!*, the addressee/adviser is a salient, but implicit Participant (A) which is expressed with a remote (dashed) edge to a prior mention. Remaining categories are abbreviated as: H – Parallel Scene, P – Process, E – Elaborator, D – Adverbial, F – Function.

2 Background and Motivation

We first consider the organization of semantic annotations in corpora, arguing that UCCA's representation of predicate-argument structure should serve as a foundation for coreference annotations.

2.1 Approaches to Semantic Multilayering

A major consideration in the design of coreference annotation schemes, as well as meaning representations generally, is what the relevant annotation targets are and whether they should be normalized across layers when the text is annotated for multiple aspects of linguistic structure. Should coreference be annotated completely independently of decisions about syntactic phrases and semantic predicate-argument structures? On the one hand, this decoupling of the annotations might absolve the coreference annotators from having to worry about other annotation conventions in the corpus. On the other hand, this is potentially a recipe for inconsistent annotations across layers, making it more difficult to integrate information across layers for complex reasoning in natural language understanding systems. Moreover, certain details of coreference annotation may be underdetermined such that relying on other layers would save coreference annotators and guidelines-developers from having to reinvent the wheel.

We can examine existing semantic annotation schemes with regard to two closely related criteria: a) **anchoring**, i.e. the previously determined underlying structure (characters, tokens, syntax, etc.) that defines the set of possible annotation targets in a new layer; and b) **modularity**, the extent to which multiple kinds of information are expressed as separate (possibly linked) structures/layers, which may be annotated in different phases.

Massively multilayer corpora. A few corpora comprise several layers of annotation, including semantics, with an emphasis on modularity of these layers. One example is OntoNotes (Hovy et al., 2006), annotated for syntax, named entities, word senses, PropBank (Palmer et al., 2005) predicate-argument structures, and coreference. Another example is GUM (Zeldes, 2017), with layers for syntactic, coreference, discourse, and document structure. Both of these resources cover multiple genres. Different layers in these resources are anchored differently, as noted below.

Token-anchored. Many semantic annotation layers are specified in terms of character or token offsets. This is the case for UCCA's Foundational Layer (§2.2), FrameNet (Fillmore and Baker, 2009), RED (O'Gorman et al., 2016), all of the layers in GUM, and the named entity and word sense

[3] Following UCCA's philosophy, we interpret both *fears* and *them* mainly as evoking the emotional state of having fears (i.e., "how did you get over them" ≈ "how did you get over being afraid"). This analysis abstracts away from the more direct reading as the specific objects of fear; but either way, the proper semantic head of the first sentence has to be *fears* (not *have*), and from our flexible minimum/maximum span policy it follows that any mention coreferring with *fears* automatically corefers with the whole scene.

Further, we interpret both *anyone else* and *you* as referring to the unknown-sized set of audience members sharing the speaker's fears. Whereas *you* introduces a presupposition that this set is non-empty, this is not the case for the negative polarity item *anyone else*. Although questionable in terms of cohesion (as the presupposition created by *you* fails if the answer to the first question is 'no'), this is a typical phenomenon in conversational data and can be explained by recognizing that the second question is implicitly conditional: "**If so,** how did you get over them?"

annotations in OntoNotes. Though the guidelines may mention syntactic criteria for deciding what units to semantically annotate, the annotated data does not explicitly tie these layers to syntactic units, and to the best of our knowledge the annotator is not constrained by the syntactic annotation.

Syntax-anchored. Semantic annotations explicitly defined in terms of syntactic units include: PropBank (such as in OntoNotes); and the coreference annotations in the Prague Dependency Treebank (PDT; Nedoluzhko et al., 2016). In addition, PDT's "deep syntactic" tectogrammatical layer, which is built on the syntactic analytic layer, can be considered quasi-semantic (Böhmová et al., 2003).

Transformed syntax. In other cases, semantic label annotations enrich skeletal semantic representations that have been deterministically converted from syntactic structures. One example is Universal Decompositional Semantics (White et al., 2016), whose annotations are anchored with PredPatt, a way of converting Universal Dependencies trees (Nivre et al., 2016) to approximate predicate-argument structures.

Sentence-anchored. The Abstract Meaning Representation (AMR; Banarescu et al., 2013) is an example of a highly integrative (anti-modular) approach to sentence-level meaning, without anchoring below the sentence level. AMR annotations take the form of a single graph per sentence, capturing a variety of kinds of information, including predicate-argument structure, sentence focus, modality, lexical semantic distinctions, coreference, named entity typing, and entity linking ("Wikification"). English AMR annotators provide the full graph at once (with the exception of entity linking, done as a separate pass), and do not mark how pieces of the graph are anchored in tokens, which has spawned a line of research on various forms of token-level alignment for parsing (e.g. Flanigan et al., 2014; Pourdamghani et al., 2014; Chen and Palmer, 2017; Szubert et al., 2018; Liu et al., 2018). Chinese AMR, by contrast, is annotated in a way that aligns nodes with tokens (Li et al., 2016).

Semantics-anchored. The approach we explore here is the use of a *semantic* layer as a foundation for a different type of semantic layer. Such approaches support modularity, while still allowing annotation reuse. A recent example for this approach is multi-sentence AMR (O'Gorman et al., 2018), which links together the previously annotated per-sentence AMR graphs to indicate coreference across sentences.

2.2 UCCA's Foundational Layer

UCCA is a coarse-grained, typologically-motivated scheme for analyzing abstract semantic structures in text. It is designed to expose commonalities in semantic structure across paraphrases and translations, with a focus on predicate-argument and other semantic head-modifier relations. Formally, each text passage is annotated with a directed acyclic graph (DAG) over semantic elements called **units**. Each unit, corresponding to (anchored by) one or more tokens, is labeled with one or more semantic **categories** in relation to a parent unit.

The foundational layer[4] specifies a DAG structure organized in terms of **scenes** (events/situations mentioned in the text). This can be seen for three sentences in figure 1, where each corresponds to a Parallel Scene (denoted by the category label H) as three events are presented in sequence. A scene unit is headed by a predicate, which is either a State (S), like *these fears*, or a Process (P), like *get over*. Most scenes have at least one Participant (A), typically an entity or location—in this case, the individuals experiencing fear. Semantic refinements of manner, aspect, modality, negation, causativity, etc. are marked with the category Adverbial (D). Time (T) is used for temporal modifiers. Within a non-scene unit, the semantic head is marked Center (C), while semantic modifiers are Elaborators (E). Function (F) applies to words considered to add no semantic content relevant to the scene structure.

Some additional structural properties are worthy of note. An **unanalyzable unit** indicates that a group of tokens form a multiword expression with no internal semantic structure, like *get over* 'surmount'. A **remote edge** (reentrancy, shown as a dashed line in figure 1) makes it possible for a unit to have multiple parent units such that the structure is not a tree. This is mainly used when a Participant is shared by multiple scenes. Texts are annotated in passages generally larger than sentences, and remote edges may cross sentence boundaries—for example, when a Participant mentioned in one sentence is implicit in the next, such as *you* as the implicit advice-giver in the sentence *Advice please!*. **Implicit units** are null elements used when there is a salient piece of the meaning that is implied but not expressed overtly *anywhere* in the passage. (If

[4] Annotation guidelines: https://github.com/UniversalConceptualCognitiveAnnotation/docs/blob/master/guidelines.pdf

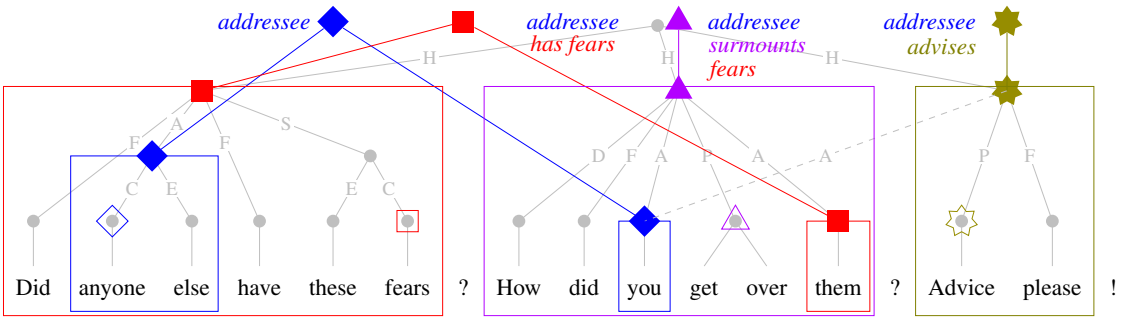


Figure 2: The reference layer UCoref on top of UCCA's foundational layer. A new "referent node" is introduced as a parent for each cluster of coreferring mentions. Colors and shapes indicate coreferring mentions. By virtue of the remote Participant edge (dashed line), the addressee referent implicitly participates in the third scene as well.

the third sentence from figure 1 was annotated in isolation, the advice-giver would be represented by an implicit unit.)

2.3 Insufficiency of the Foundational Layer

In addition to the benefits of a semantic foundational layer for coreference annotation (§2.1), we point out how adding such a layer to UCCA would rectify shortcomings of the foundational layer.

First and foremost, UCCA currently lacks any representation of "true" coreference, i.e., the phenomenon that two or more *explicit* units are mentions of the same entity. Second, though remote edges are helpful for indicating that a Participant is shared between multiple scenes, this is problematic if the referent is mentioned multiple times in the passage. Because the information that those mentions are coreferent is missing, the choice which mention to annotate with a remote edge is underdetermined. This leads to multiple conceptually equivalent choices that are formally distinct, opening the way for spurious disagreements. For example, the implicit advice-giver in figure 1 could be marked equally well with a remote edge to *anyone else* instead of *you*, resulting in a structurally diverging graph (taking the presented analysis as the reference).[5] And third, many other implicit relations relevant to coreference (e.g., implied common sense part/whole relations, via *bridging*) are not exposed in the foundational layer of UCCA. A layer that annotates identity coreference could be extended with such additional information in the future.

[5]While additional, more restrictive guidelines could to some extent curb such confusion (e.g., by specifying that the closest appropriate mention to the left should always be chosen as the remote target), this would require the foundational layer annotators to be confident in the notion of coreference to determine which mentions are "appropriate", eliminating the modularity and intuitiveness we desire.

3 The UCoref Layer

The underlying hypothesis of this work is that the spans of words that form referring expressions, i.e., evoke or point back to entities and events in the world, are also grouped as semantic units in the foundational layer of UCCA. This assumption is motivated by the fundamental principles of UCCA as a neo-Davidsonian theory: The basic elements of a discourse are descriptions of scenes (≈ events), and their basic elements are participants (≈ entities). We can thus automatically identify scene and participant units as referring. With this high-precision preprocessing and a small set of simple guidelines for identifying other UCCA units as referring, the process of *mention identification* in UCoref is very efficient. Figure 2 illustrates how UCoref interacts with the foundational layer. Four referents and six mentions (two singletons) are identified based on the criteria below.

Scene and Participant units. The vast majority of referent mentions can be identified by two simple rules: 1) All **scene** units are considered mentions as they constitute descriptions of actions, movements, or states as defined in the foundational layer guidelines. 2) Similarly, all **Participant** units are considered mentions as they describe entities that are contributing to or affected by a scene/event (including locations and other scenes/events).

Special attention should be paid to relational nouns like *teacher* or *friend* that both refer to an entity and evoke a process or state in which the entity generally or habitually participates.[6] According to the UCCA guidelines, these words are analyzed internally (as both **P**/**S** and **A** within a nested unit over the same span), in addition to the

[6]A teacher is a person who teaches and a friend is a person who stands in a friendship relation with another person. Cf. Newell and Cheung (2018); Meyers et al. (2004).

context-dependent incoming edge from their parent. However, the inherent scene (of teaching or friendship) is merely *evoked*, but not *referred to*, and it is usually invariant with respect to the explicit context it occurs in. Moreover, treating one span of words as two mentions would pose a significant complication. Thus, we consider these units only in their role as Participant (and not scene) mentions.

Non-scene-non-participant units. A certain subset of the remaining unit types are considered to be mention *candidates*. This subset is comprised of the categories, *Time*, *Elaborator*, *Relator*, *Quantity*, and *Adverbial*. We give detailed guidelines for these categories, as well as for coreference markup, in the supplementary material (appendix A).

Center units. For simplicity, a referring unit with a *single* Center usually does not require its Center to be marked separately, as a unit always corefers with its Center (see §4 and §5.1 about how this relates to the min/max span distinction).

Multi-Center units receive a different treatment: One use of multi-Center units is coordination, where each conjunct is a Center. Here we do want to mark up the conjuncts in addition to the whole coordination unit—provided the whole unit is referring by one of the other criteria—and assign them to separate coreference clusters. Another class of multi-Center units, which we call *relative partitive constructions*, is less straightforward to handle. Consider a phrase like *the top of the mountain*. The intuition given in the UCCA guidelines is that while the phrase is syntactically and, to some extent, semantically headed by *top*, it can only be fully understood in relation to *mountain*; thus, both words should be Centers. This construction is clearly less symmetric than coordination, but at this point we do not have a reliable way of formally distinguishing the two in preprocessing, purely based on the UCCA structure and categories. Thus, multi-Center units deserve a more nuanced manual UCoref analysis in future work; however, for the sake of consistency and simplicity, we treat all multi-Center units in the same way as we treat coordinations in our pilot annotation (§5).

Implicit units. Implicit units may be identified as mentions and linked to coreferring expressions just like any other unit, as long as they meet the criteria outlined above.

4 Comparison with Other Schemes

The task of coreference resolution is far from trivial and has been approached from many different angles. Below we give a detailed analysis of the theoretical differences between three particular frameworks: OntoNotes (Hovy et al., 2006), Richer Event Description (RED; O'Gorman et al., 2016), and the Georgetown University Multilayer corpus (GUM; Zeldes, 2017).

Singletons and events. RED and UCoref annotate all nominal entity, nominal event, and verbal event mentions, including singletons.[7] OntoNotes does not include singleton mentions in the coreference layer.[8] Further, only those verbal mentions that are coreferent with a nominal are included. GUM includes all nominal mentions, including singletons and nominal event mentions, and follows the OntoNotes guidelines for verbal mentions.

Syntactic vs. semantic criteria. GUM and OntoNotes, despite not being *anchored* in syntax, specify syntactic criteria for mention and coreference annotation. The criteria in RED and UCoref, on the other hand, are fundamentally semantic. Rough syntactic guidance is only given where appropriate and at no time is a decisive factor.

Minimum and maximum spans. The policy on mention spans is often one of two extremes: *minimum* spans (also called *triggers* or *nuggets*), which typically only consist of the head word or expression that sufficiently describes the type of entity or event; or *maximum* spans (also called *full mentions*), containing all arguments and modifiers. GUM and OntoNotes generally apply a maximum span policy for nominal mentions, with just a few exceptions.[9] For verbal mentions, OntoNotes chooses minimum spans, whereas GUM annotates full clauses or sentences. RED always uses minimum spans, except for time expressions, which follow the TIMEX3 standard (Pustejovsky et al., 2010). One of the main advantages of UCoref is that the preexisting predicate-argument and head-modifier structures of the foundational layer allow a flexible and reliable mapping between minimum and maximum span annotations. Addition-

[7] For event coreference specifically, see also EventCorefBank (ECB; Bejan and Harabagiu, 2010) and the TAC-KBP Event Track (Mitamura et al., 2015), which uses the ACE 2005 dataset (LDC2006T06; Doddington et al., 2004).

[8] A separate layer records all *named* entities, however, and non-coreferent ones can be considered singleton mentions.

[9] The GUM guidelines specify that clausal modifiers should not be included in a nominal mention.

ally, UCoref has 'null' spans, corresponding to implicit units in UCCA.[10]

Predication. OntoNotes does not assert a coreference relation between copular arguments.[11] RED distinguishes several relation types depending on the "predicativeness" of the expression and in particular asserts a set-membership (i.e., non-identity) relation when the second argument is indefinite. In GUM, relation types are assigned based on different criteria,[12] and, depending on the polarity and modality of the copula, its arguments may be marked as coreferring mentions, even if they are indefinite.[13] A slightly different distinction is made in UCoref, where, thanks to the foundational layer, evokers of set-membership and attributive relations are marked as stative scenes in which the modified entity participates. Definite identity is handled in the same way as in RED, as well as relational nouns except for the special case of generic mentions (appendix A.2).

Apposition. In RED and OntoNotes, punctuation is considered a strict criterion for marking appositives, while GUM relies solely on syntactic completeness. In OntoNotes and GUM, ages specified after a person's name are considered separate appositional mentions, coreferring with the name mention they modify. UCoref takes advantage of UCCA's semantic Center-Elaborator structure, abstracting away from superficial markers like punctuation which may not be available in all genres and languages (details in appendix A.2).

Prepositions. Whereas OntoNotes and GUM stick to the syntactic notion of NPs, UCoref includes prepositions and case markers within mentions. This does not have a major effect on coreference, but contributes to consistency between languages that vary in the grammaticalization of their case marking.

Coordination. Our treatment of coordinate entity mentions is adopted and expanded from the GUM guidelines, where the span containing the full coordination is only marked up if it is antecedent to a plural pronominal mention. OntoNotes does not specify how coordinations in particular should be handled; while the guidelines state that out of *head-sharing* (i.e., elliptic) mentions only the largest one should be picked, we assume that coordinations of multiple *explicitly headed* phrases are not targeted as mentions in addition to the conjuncts. The minimum span approach of RED precludes marking full coordinations in addition to conjuncts.

Summary. That OntoNotes does not annotate singleton mentions makes it the most restrictive out of the compared frameworks. Despite its emphasis on syntax, GUM is closer to our framework as it includes singletons and marks full spans for non-singleton events; the marking of bridging relations, directed coreference links, and information status present in GUM is beyond our scope here. RED is conceptually closest to UCoref in marking all entity, time, and event mentions, except for the difference in span boundaries. This can largely be resolved as we will show in §5.1.

5 Pilot Annotation

In order to evaluate the accessibility of the annotation guidelines given above and in appendix A, and facilitate empirical comparison with other schemes, we conducted a pilot annotation study. We annotated a small English dataset consisting of subsets of the OntoNotes (LDC2013T19), RED (LDC2016T23), and GUM[14] corpora with the UCCA foundational and coreference layers.[15]

The OntoNotes documents are taken from blog posts, the GUM documents are WikiHow instructional guides, and the RED documents are online forum discussions. Because all annotations were done by a single annotator each and not reviewed, our results are to be understood as a proof of concept; measuring interannotator agreement will be

[10] The coreference layer of the Prague Dependency Treebank (Nedoluzhko et al., 2016), quite similarly to the proposed framework, marks null-mentions arising from control verbs, reciprocals, and dual dependencies (in general, null-nodes arising from obligatory valency slot insertions into the tectogrammatical layer)—the syntactic equivalents of implicit units and remote edges in UCCA. Further, in case the mention is a root of a nontrivial subtree, it is underspecified whether the mention spans only the root, the whole subtree or some part of it.

[11] Neither do Poesio and Artstein (in the ARRAU corpus; 2008).

[12] In particular, the notion of *bridging* is interpreted differently between GUM and RED: GUM reserves it for entities that are expected (from world knowledge) to stand in some relationship (e.g., part/whole) with each other, which is reflected in a definite initial mention of the 'bridging target' (*My car is broken; it's* ***the motor***). RED uses it for copular predications involving relational/occupational nouns like ***John*** *is* ***a/the killer***, which are simple 'coref' (or 'ana'/'cata', if one mention is a pronoun) relations in GUM. We consider neither of these definitions in this work (see appendix A.2).

[13] See also Chinchor (1998).

[14] https://github.com/amir-zeldes/gum

[15] Since the RED documents are not tokenized (character spans are used for mention identification), we preprocessed them with the PTB tokenizer and the Punkt sentence splitter using Python NLTK.

	GUM	OntoNotes	RED
sentences	70	17	24
tokens	1180	303	302
↪ non-punct	1030	261	274
UCCA units	1436	336	379
↪ candidates	911	195	186

Table 1: Overview of our pilot dataset. *Candidates* refers to the UCCA units that are filtered by category for mention candidacy before manual annotation.

necessary in the future to gauge the difficulty of the task and quality of guidelines/data.

Table 1 shows the distribution of tokens and UCCA foundational units, and table 2 compares the distribution of UCoref units with the respective "native" annotation schema for each corpus. We can see that about one third of all UCCA units are identified as mentions, in all corpora. The automatic candidate filtering based on UCCA categories simplifies this process for the annotator by removing about one third to one half of units. There are similar amounts of scene and Participant units (both of which are always mentions), but it is important to note that Participant units can also refer to events. We can see this reflected by the majority of referent units being event referents. We can also see that most of the referents in GUM, RED, and UCoref are in fact singletons, and the number of non-singleton referents is quite similar between each scheme and UCoref. Most implicit units and targets of remote edges are part of a non-singleton coreference cluster, which confirms the issue of spurious ambiguity we pointed out in §2.3.

5.1 Recovering Existing Schemes

Next we examine the differences in gold annotations between our proposed schema and existing schemas and how we can (re)cover annotations in established schemas from our new schema. We can interpret this experiment as asking: If we had a perfect system for UCoref, could we use that to predict GUM/OntoNotes/RED-style coreference? And vice versa, if we had an oracle in one of those schemes, and possibly oracle UCoref mentions, how closely could we convert to UCoref?[16]

Exact mention matches. A naïve approach would be to look at the token spans covered by all mentions and reference clusters and count how often we can find an exact match between UCoref and one of the existing schemes.

[16]See also Zeldes and Zhang (2016), who base a full coreference resolution system on this idea.

In UCoref, we use maximum spans by default, but thanks to the nature of the UCCA foundational layer, minimum spans can easily be recovered from Centers and scene-evokers. For schemas with a *minimum span* approach, we can switch to a minimum span approach in UCoref by choosing the head unit of each maximum span unit as its representative mention. This works well between UCoref and RED as they have similar policies for determining semantic heads, which is crucial for, e.g., light verb constructions. This would be problematic, however, when comparing to a minimum span schema that uses syntactic heads. For schemas with a *non-minimum span* approach, we keep only the maximum span units from UCoref and discard any heads that have been marked up representatively for their parent (e.g., as remote targets).

Fuzzy mention matches. Because our theoretical comparison in §4 exposed systematically diverging definitions of what to include in a mention span, we also apply an evaluation that abstracts away from some of these differences. We greedily identify one-to-one alignments for maximally overlapping mentions, as measured by the Dice coefficient:

$$m_A^*, m_B^* \leftarrow \underset{m_A \in (A \setminus L_A), m_B \in (B \setminus L_B)}{\arg\max} \frac{|m_A \cap m_B|}{|m_A| + |m_B|}$$

where L_A (L_B) records the mentions from annotation A (B) aligned thus far, and stopping when this score falls below a threshold μ. μ is a hyperparameter controlling how much overlap is required: $\mu = 1$ corresponds to exact matches only, while $\mu = 0$ includes all overlapping mention pairs as candidates (we report fuzzy match results for $\mu = 0$). Once a mention is aligned it is removed from consideration for future alignments.

We align referents by the same procedure. Results are reported in table 3.

5.2 Findings

We can see in table 3 that UCoref generally covers between 60% and 80% of exact *mentions* in existing schemes ('R' columns), however, the amount of UCoref units that are present in other schemes varies greatly, between 21.3% (OntoNotes) and 79.5% (RED; 'P' columns). This is generally expected based on our theoretical analysis in §4. Fuzzy match has a great effect on the maximum span schemes in GUM and OntoNotes, resulting in up to 100% of mentions being aligned, and a lesser,

	WikiHow		Blog		Forum	
	GUM	UCR	ONT	UCR	RED	UCR
mentions	288	466	40	128	120	117
↪ event	158	208	–	47	70	54
↪ entity/A	127	215	–	66	47	58
↪ other	3	43	–	14	3	5
↪ NE	–	–	10	–	–	–
↪ IMP	–	26	–	6	–	4
↪ remote	–	10	–	3	–	1

	WikiHow		Blog		Forum	
	GUM	UCR	ONT	UCR	RED	UCR
referents	155	291	20	96	82	78
↪ event	108	180	–	43	58	47
↪ entity	47	108	–	46	21	27
↪ time	0	3	–	7	3	4
↪ non-singleton	46	36	10	13	9	18
↪ IMP	–	26	–	1	–	4
↪ remote	–	7	–	2	–	1

Table 2: Distribution of mentions and referents in the datasets. **Mentions**: Under *event*, we count UCoref (UCR) scenes, GUM mentions of the types 'event' or 'abstract', and RED EVENTs; under *entity*, we count UCR A's, GUM 'person', 'object', 'place', and 'substance' mentions, and RED ENTITYs. *NE* stands for OntoNotes (ONT) named entities and IMP and remote for implicit and remote UCR units. A coreference cluster (**referent**) is classified as an *event* referent if there is at least one event mention of that referent, as a *time* referent if there is at least one UCR T / GUM 'time' / RED TIMEX3 mention of that referent, and as an *entity* referent otherwise; we also report how many of the IMP and remote units are part of non-singleton referents.

	mentions										**referents**									
	GUM			OntoNotes			RED				GUM			OntoNotes			RED			
	P	R	F	P	R	F	P	R	F		P	R	F	P	R	F	P	R	F	
=	41.7	60.6	49.4	21.3	67.5	32.3	*79.5*	*77.5*	*78.5*	=	19.3	24.4	21.6	3.1	15.2	5.2	*52.6*	*50.0*	*51.3*	
										≈	31.6	40.0	35.3	9.4	45.0	15.5	*66.7*	*63.4*	*65.0*	
≈	67.0	97.2	79.3	31.5	100.0	47.9	*88.9*	*86.7*	*87.8*	=	43.9	55.6	49.0	5.2	25.0	8.6	*66.7*	*63.4*	*65.0*	
										≈	59.6	75.6	66.7	10.4	50.0	17.2	*80.8*	*76.8*	*78.8*	

Table 3: Exact (=) and fuzzy (≈) referent matches based on exact and aligned mentions between UCoref and GUM, OntoNotes, and RED. Precision (P) and recall (R) are measured treating gold UCoref annotation as the prediction and gold annotation in each respective existing framework as the reference. Italics indicate minimum UCoref spans are used. Implicit UCoref units are excluded from this evaluation, and children of remote edges are only counted once (for their primary edge).

but still positive effect on RED.[17] We observe a similar trend for *referent* matches, which follows partly from the mismatch in mention annotation, and partly from diverging policies in marking coreference relations, as discussed above. Whether or not singleton event and/or entity referents are annotated has a major impact here. Below we give examples for sources of non-exact mention matches that can be resolved using fuzzy alignment.

GUM and OntoNotes. A phenomenon that is trivially resolvable using fuzzy alignments is punctuation, which is excluded from all UCoref units, but included in GUM and OntoNotes. Another group of mentions recovered are prepositional phrases, where UCoref includes prepositions (*to them*, *since the end of 2005*), and GUM and OntoNotes do not (*them*, *the end of 2005*). As mentioned in §4, GUM deviates from its maximum span policy for clausal modifiers of noun phrases, which are stripped off from the mention. Noun phrases modified in this way can be fuzzily aligned with the maximum spans in UCoref, even if the modifier is very long: *people who are stuck on themselves intolerant of people different from them rude or downright arrogant* (UCoref) gets aligned with *people* (GUM).

RED. Almost 80% of both RED and UCoref mentions match exactly, but there are some cases of divergence: 1) One subset of these are time expressions like *this morning*, where, as pointed out above, RED marks maximum spans. However, in UCoref these are internally analyzable—thus their Center will be extracted for minimum spans (here, *morning*). On the other hand, idiomatic multiword expressions (MWEs) such as verb-particle constructions (e.g., *pass away* 'die') are treated as unanalyzable in UCCA, but only the syntactic head (*pass*) is included in RED. 2) Also interesting are predicative prepositions and adverbials in copular or expletive constructions: *there will be lots of good dr.s and nurses around*. Here, UCoref chooses *around* as the (stative) scene evoker (and would mark the prepositional object as a participant, if it is explicit), while RED chooses the copula *be*. 3) UCCA treats some verbs as modifiers rather than predicates themselves: e.g., *stopped* in

[17] Note, though, that this evaluation only shows us *if* we can find a fuzzy alignment, not whether the aligned spans are actually equivalent. As purely span-based alignment is prone to errors, a future extension to the algorithm should take information about (ideally semantic) heads into account.

i m [sic] *stopped feeling her move* and *it seemed* in *it seemed tom* [sic] *take forever*. The former, as an aspectual secondary verb, is labeled Adverbial (D); the latter, which injects the perspective of the speaker, is labeled Ground (G). Since we do not generally consider these categories referring, these are not annotated as mentions in UCoref, though they are in RED.

5.3 Discussion

For the non-minimum span schemas GUM and OntoNotes, we can use a fuzzy mention alignment based on token overlap to find many pairs which aim to capture the same mention, only under different annotation conventions. RED is most similar to UCoref in defining what counts as a mention, though our corpus analysis showed that the notion of *semantic heads* is interpreted differently for certain constructions, where UCCA is more liberal about treating verbs as modifiers rather than heads. While counting fuzzy matches allows us to recover partially overlapping spans (time expressions, verbal MWEs), other phenomena (adverbial copula constructions, secondary verbs) have inconsistent policies between the two schemes that require more elaborate methods to align. We can thus, to some extent, use UCoref to predict RED-style annotations, with the additional gain of flexible minimum/maximum spans and cross-sentence predicate-argument structure for a whole document. Furthermore, we see that UCoref subsumes all OntoNotes mentions and nearly all GUM mentions and is able to reconstruct coreference clusters in GUM with high recall.

6 Conclusion

We have defined and piloted a new, modular approach to coreference annotation based on the semantic foundational layer provided by UCCA. An oracle experiment shows high recall with respect to three existing schemes, as well as high precision with respect to the most similar of the three. We have released our annotations to fuel future investigations.

Acknowledgments

We would like to thank Amir Zeldes and two anonymous reviewers for their many helpful comments, corrections, and pointers to relevant literature. This research was supported in part by NSF award IIS-1812778 and grant 2016375 from the United States–Israel Binational Science Foundation (BSF), Jerusalem, Israel.

References

Omri Abend and Ari Rappoport. 2013. Universal Conceptual Cognitive Annotation (UCCA). In *Proc. of ACL*, pages 228–238, Sofia, Bulgaria.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proc. of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria.

Cosmin Bejan and Sanda Harabagiu. 2010. Unsupervised event coreference resolution with rich linguistic features. In *Proc. of ACL*, pages 1412–1422, Uppsala, Sweden.

Alena Böhmová, Jan Hajič, Eva Hajičová, and Barbora Hladká. 2003. The Prague Dependency Treebank: A three-level annotation scenario. In Anne Abeillé, editor, *Treebanks: Building and Using Parsed Corpora*, Text, Speech and Language Technology, pages 103–127. Springer Netherlands, Dordrecht.

Wei-Te Chen and Martha Palmer. 2017. Unsupervised AMR-dependency parse alignment. In *Proc. of EACL*, pages 558–567, Valencia, Spain.

Nancy A. Chinchor. 1998. Overview of MUC-7/MET-2. In *Proc. of the Seventh Message Understanding Conference (MUC-7)*, Fairfax, Virginia.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The Automatic Content Extraction (ACE) program - Tasks, data, and evaluation. In *Proc. of LREC*, pages 837–840, Lisbon, Portugal.

Charles J. Fillmore and Collin Baker. 2009. A frames approach to semantic analysis. In Bernd Heine and Heiko Narrog, editors, *The Oxford Handbook of Linguistic Analysis*, pages 791–816. Oxford University Press, Oxford, UK.

Jeffrey Flanigan, Sam Thomson, Jaime Carbonell, Chris Dyer, and Noah A. Smith. 2014. A discriminative graph-based parser for the Abstract Meaning Representation. In *Proc. of ACL*, pages 1426–1436, Baltimore, Maryland, USA.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: the 90% solution. In *Proc. of HLT-NAACL*, pages 57–60, New York City, USA.

Bin Li, Yuan Wen, Lijun Bu, Weiguang Qu, and Nianwen Xue. 2016. Annotating The Little Prince with Chinese AMRs. In *Proc. of LAW X – the 10th Linguistic Annotation Workshop*, pages 7–15, Berlin, Germany.

Yijia Liu, Wanxiang Che, Bo Zheng, Bing Qin, and Ting Liu. 2018. An AMR aligner tuned by transition-based parser. In *Proc. of EMNLP*, pages 2422–2430, Brussels, Belgium.

Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. 2004. The NomBank project: an interim report. In *Proc. of the Frontiers in Corpus Annotation Workshop*, pages 24–31, Boston, Massachusetts, USA.

Teruko Mitamura, Zhengzhong Liu, and Eduard Hovy. 2015. Overview of TAC-KBP 2015 Event Nugget Track. In *Proc. of TAC*, Gaithersburg, Maryland, USA.

Nafise Sadat Moosavi and Michael Strube. 2016. Which coreference evaluation metric do you trust? A proposal for a link-based entity aware metric. In *Proc. of ACL*, pages 632–642, Berlin, Germany.

Anna Nedoluzhko, Michal Novák, Silvie Cinková, Marie Mikulová, and Jiří Mírovský. 2016. Coreference in Prague Czech-English Dependency Treebank. In *Proc. of LREC*, pages 169–176, Portorož, Slovenia.

Edward Newell and Jackie Chi Kit Cheung. 2018. Constructing a lexicon of relational nouns. In *Proc. of LREC*, pages 3405–3410, Miyazaki, Japan.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: a multilingual treebank collection. In *Proc. of LREC*, pages 1659–1666, Portorož, Slovenia.

Tim O'Gorman, Michael Regan, Kira Griffitt, Ulf Hermjakob, Kevin Knight, and Martha Palmer. 2018. AMR beyond the sentence: the Multi-sentence AMR corpus. In *Proc. of COLING*, pages 3693–3702, Santa Fe, New Mexico, USA.

Tim O'Gorman, Kristin Wright-Bettner, and Martha Palmer. 2016. Richer Event Description: Integrating event coreference with temporal, causal and bridging annotation. In *Proc. of the 2nd Workshop on Computing News Storylines*, pages 47–56, Austin, Texas, USA.

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

Massimo Poesio and Ron Artstein. 2008. Anaphoric Annotation in the ARRAU Corpus. In *Proc. of LREC*, pages 1170–1174, Marrakech, Morocco.

Massimo Poesio, Roland Stuckardt, and Yannick Versley, editors. 2016. *Anaphora Resolution: Algorithms, Resources, and Applications*. Theory and Applications of Natural Language Processing. Springer, Berlin.

Nima Pourdamghani, Yang Gao, Ulf Hermjakob, and Kevin Knight. 2014. Aligning English strings with Abstract Meaning Representation graphs. In *Proc. of EMNLP*, pages 425–429, Doha, Qatar.

Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. 2014. Scoring coreference partitions of predicted mentions: a reference implementation. In *Proc. of ACL*, pages 30–35, Baltimore, Maryland.

James Pustejovsky, Kiyong Lee, Harry Bunt, and Laurent Romary. 2010. ISO-TimeML: An international standard for semantic annotation. In *Proc. of LREC*, pages 394–397, Valletta, Malta.

Marta Recasens, Zhichao Hu, and Olivia Rhinehart. 2016. Sense anaphoric pronouns: Am I one? In *Proc. of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016)*, pages 1–6, San Diego, California.

Ida Szubert, Adam Lopez, and Nathan Schneider. 2018. A structured syntax-semantics interface for English-AMR alignment. In *Proc. of NAACL-HLT*, pages 1169–1180, New Orleans, Louisiana.

Aaron Steven White, Drew Reisinger, Keisuke Sakaguchi, Tim Vieira, Sheng Zhang, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2016. Universal Decompositional Semantics on Universal Dependencies. In *Proc. of EMNLP*, pages 1713–1723, Austin, Texas, USA.

Amir Zeldes. 2017. The GUM corpus: creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Amir Zeldes. 2018. A predictive model for notional anaphora in English. In *Proc. of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 34–43, New Orleans, Louisiana.

Amir Zeldes and Shuo Zhang. 2016. When annotation schemes change rules help: A configurable approach to coreference resolution beyond OntoNotes. In *Proc. of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016)*, pages 92–101, Ann Arbor, Michigan.

A Detailed Guidelines

A.1 Identifying Mentions

Non-scene-non-participant units. A certain subset of the remaining unit types are considered to be mention *candidates*. This subset is comprised of the categories, *Time*, *Elaborator*, *Relator*, *Quantity*, and *Adverbial*.

Time (T) Absolute or relative time expressions like *on May 15, 1990*, *now*, or *in the past*, which are marked Time (T) in UCCA, are considered mentions. However, frequencies and durations, which are also T units in UCCA, are discarded. In order to reliably distinguish these different kinds of time expressions from each other, they have to be identified manually.

Elaborator (E) Elaborators modifying the Center (C) of a non-scene unit are considered mentions if they themselves describe a scene or entity. This is the case, for example, with (relative) clauses and (prepositional) phrases describing the Center's relation with another entity as in

$$\underline{[\ \textit{the book}_C\ \underline{[\textit{about the dog}_C\ ]_E}\ ]},$$

as well as contingent attributive modifiers, which are stative scenes in UCCA, like *old* in

$$\underline{[\ \textit{the}\ \underline{[\ \textit{old}_S\ (\textit{book})_A\ ]_E}\ \textit{book}_C\ ]}.$$

By contrast, Elaborator units that do not evoke a person, thing, abstract object, or scene are not considered referring, as in

$$\underline{[\ \textit{medical}_E\ \textit{school}_C\ ]},$$

where *medical* is an inherent property and thus non-referring.

In English, this often corresponds to units whose Center is an adjective, adverb, or determiner.[18] Bear in mind, however, that these syntactic criteria are language-specific and should only be taken as rough guidance, rather than absolute rules. Thus, referring non-scene Elaborators should be identified manually. By contrast, E-scenes will be identified as mentions automatically, by the scene unit criterion.

Relator (R) Relators should be marked as mentions if and only if they constitute an anaphoric (or cataphoric) reference *in addition to* their relating function.

As an illustration what we mean by that, consider the two occurrences of **that** in the following example, which are both Relators in UCCA:

I didn't like ***that***$_1$ *he said the things* ***that***$_2$ *he said.*

Here, $that_2$ is an anaphoric reference to *things*, whereas $that_1$ is purely functional and thus should not be identified as referring. In English, this corresponds to the syntactic category of relative pronouns.

Most R units, however, are non-referring expressions like prepositions, so identification of the few referring instances of Relators has to be done manually.

Quantity (Q) Partitive constructions like *one of the 5 books* contain mentions of two distinct referents: *the 5 books* and *one of the 5 books*. According to the v2 UCCA guidelines, these expressions should be annotated as an Elaborator-Center structure with a remote edge:

$$\underline{[[\ \textit{one}_Q\ (\textit{books})_C\ ]_C\ \underline{[\ \textit{of}_R\ \textit{the}_F\ \textit{5}_Q\ \textit{books}_C\ ]_E}\ ]_X}$$

Such an annotation will result in correct identification of the two mentions based on the guidelines given so far (by choosing the E unit and the whole X[19] unit), without the need to identify the Quantifier (Q) unit one_Q. However, in foundational layer annotations made based on the UCCA v1 guidelines the same phrase would receive a flat structure (cf. discussion of Centers above):

$$\underline{[\ \textit{one}_Q\ \textit{of}_R\ \textit{the}_F\ \underline{\textit{5}_Q}\ \textit{books}_C\ ]_X}$$

In this case, we choose the whole X unit as a mention of the one book (respecting semantics rather than morphology), and the Q unit *5* as a mention of the five books.

Adverbial (D) While most Adverbial units (D) are by default not considered to be referring (they describe secondary relations over events), in some cases D units can be identified as mentions (also see **coordinated mentions** in appendix A.2).

One such phenomenon are prepositional phrases like *for another reason* and *in the majority of cases* are annotated as D in the corpus, as they modify scenes, not entities.

Another class of Adverbial units that may be identified as referring are the so-called secondary verbs like *help*, *want* and *offer*, which, according to the UCCA guidelines, modify scenes evoked

[18] According to the UCCA v1 guidelines, articles are to be annotated as Elaborators. In the v2 guidelines, the default category for articles has changed to Function.

[19] We use the placeholder X here as the actual category depends on the context (i.e., sibling and parent units) in which a unit is embedded.

by primary verbs, but do not themselves denote scenes. However, the relations described by them can sometimes be coreferring antecedents independently from the main scene:

[I_A $lost_P$ [10 lbs]$_A$.]$_j$
[I_A was_F $extremely_D$ $happy_S$ [$about_R$ $that_C$]$_A$ $_j$] .
vs.
[She_A $helped_{D\,i}$ me_A $lose_P$ $_j$ [10 lbs]$_A$.]
[I_A $really_D$ $appreciated_P$ $that_{A\,i}$] .

In both examples, *losing weight* is the main scene according to UCCA, however, in the second example the object of appreciation is *helping*. Thus, we do mark secondary verbs as mentions, but *only if* they are referred back to in the way demonstrated above.

A.2 Resolving Coreference

Appositives. Appositives and titles cooccurring with (named) entity mentions are annotated as Elaborators in UCCA and thus automatically included in the entity mention they modify. They should be marked as separate mentions, coreferring with the modified unit.

If a title or occupational noun occurs by itself or as a copular argument, we treat it as a relational noun as described in the next paragraph.

Extensional vs. intensional readings. A coordinated mention of a group of individuals, such as *John, Paul, and Mary* evokes a referent that is distinct from the possibly already evoked referents of *John*, *Paul*, and *Mary*, respectively.

Relational nouns (e.g., "the president [of Y]"), which are instantiated by a specific individual or a fixed-size set of individuals (e.g., "[X's] parents") at any given point in time, should usually be marked as coreferring with their instances, as inferred from context. This corresponds to an *extensional* (or set-theoretical) notion of reference: a distinct referent is identified by the individuals in which this concept manifests (extension).

Only in clearly generic statements like

The president's power is limited by the constitution.

You should always do what your parents tell you to.

should they evoke separate referents from any specific presidents or parents also mentioned in the same discourse. This corresponds to an *intensional* (or indirect) notion of reference: a distinct referent is identified by its general idea or concept (intension), rather than its instances.

Mentions of group-like entities with undetermined size, such as *the committee* or *all committee members*, should always be analyzed intensionally, evoking a referent separate from the possibly mentioned referents for the individuals comprising it.[20]

Negated scenes. Mentions of scenes are referring and should be marked as coreferring with other mentions of the same scene (same process or state and same participants), regardless of whether or not that scene really took place or is hypothetical:

I hoped she liked the pizza$_i$ and was relieved when I learned she did$_i$.

When both a scene and its negation are mentioned, these mentions should evoke separate referents:

I hoped she liked the pizza$_i$ and was surprised when I learned she didn't$_j$.

Coordinated mentions. When entities or events are described in conjunction, they evoke a separate referent for each of the conjuncts, and a third one for the set comprising them. The whole coordination can be explicitly referred to with another (pronominal) mention:

It is likely [that [the shock will dislocate]$_i$ ***and*** *[(the shock) break both your arms]$_j$]$_k$;*
nevertheless this$_k$ is a small price to pay for your life.

I want [Ivy$_i$ ***and*** *William$_j$]$_k$ on my debate team,*
because [both of them]$_k$ are great.

However, if the mentions are presented in *dis*junction, no separate mention for the full disjunction should be marked. If an anaphoric pronoun occurs, there are several options.

For events, a secondary relation (marked D in UCCA) that holds for both conjuncts, if available, should be marked instead:

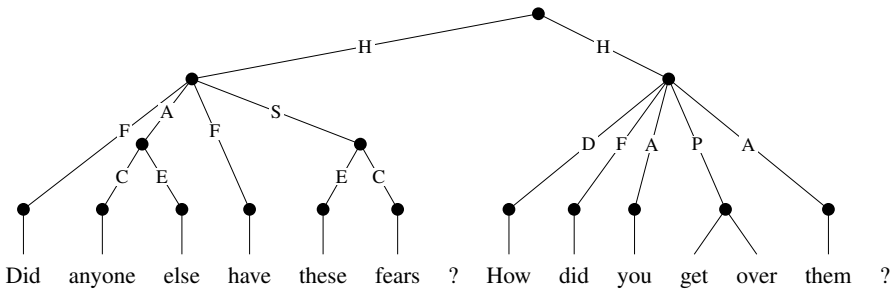
It is likely$_k$ [that [the shock will dislocate]$_i$ ***or*** *[(the shock) break both your arms]$_j$] ;*
nevertheless this$_k$ is a small price to pay for your life.

If such a unit is not available, and for entities, no coreference relation exists:

I want [Ivy$_i$ ***or*** *William$_j$] on my debate team,*
because [both of them]$_k$ are great.

[20] But singular and plural mentions of the same group can corefer (Zeldes, 2018).

Remote edges. Different types of remote edges call for different coreference annotations. *Non*-head (i.e., non-Center, -State, or -Process) remote edges indicate that the same entity/scene modifies or participates in two (potentially also coreferent) unit mentions, namely its primary parent (or the primary parent of the unit it heads) and its remote parent. This corresponds to zero anaphora, or a "core" element that is implicit in one context and explicit in another. *Head* remote edges, however, merely indicate the *category* of entity/event that is shared between a full and an elliptic or anaphoric mention ("sense anaphora"; Recasens et al., 2016). E.g., *books* in "two of the 5 books" is category-shared between *5 books* and *two (books)*, which are separate non-coreferent mentions. Whether the primary and remote parent are coreferent or not is contingent on context.


H
H
F
A
F
S
C
E
E
C
D
F
A
P
A
Did
anyone
else
have
these
fears
?
How
did
you
get
over
them
?