

Subcategorizing Adverbials in Universal Conceptual Cognitive Annotation

Zhuxin Wang Jakob Prange Nathan Schneider

Georgetown University

{zw85, jp1724, nathan.schneider}@georgetown.edu

Abstract

Universal Conceptual Cognitive Annotation (UCCA) is a semantic annotation scheme that organizes texts into coarse predicate-argument structure, offering a broad coverage of semantic phenomena. At the same time, there is still need for a finer-grained treatment of many of the categories. The Adverbial category is of special interest, as it covers a wide range of fundamentally different meanings such as negation, causation, aspect, and event quantification. In this paper we introduce a refinement annotation scheme for UCCA’s Adverbial category, showing that UCCA Adverbials can indeed be subcategorized into at least 7 semantic types, and doing so can help clarify and disambiguate the otherwise coarse-grained labels. We provide a preliminary set of annotation guidelines, as well as pilot annotation experiments with high inter-annotator agreement, confirming the validity of the scheme.

1 Introduction

The UCCA representation scheme is designed to be a multi-layered, cross-linguistically portable and stable semantic annotation scheme based on linguistic typology (Abend and Rappoport, 2013; Dixon, 2010). Among its 14 coarse-grained semantic categories, Adverbial serves as the label for modifiers to a scene’s main predicate. Despite the syntactic-sounding name, Adverbial is a semantic category that cuts across a range of syntactic constructions that provide scene-modifying meanings. These meanings fall into drastically different semantic categories—negations, manner PPs, modals, adverbs are all candidates for the Adverbial category—that change the sentence’s meaning in different ways.

This paper proposes to refine UCCA Adverbials, in order to better understand the semantic functions they encode. By distinguishing different Adverbial functions, we can disambiguate sentences that are

otherwise annotated similarly (if not identically) in UCCA. For example:

- (1) The chicken was [**deliciously** D] [cooked P].
DESCRIPTION
- (2) The chicken was [**not** D] [cooked P].
NEGATION
- (3) The chicken was [**more** D] [cooked P].
COMPARISON
- (4) The chicken was [**already** D] [cooked P].
ASPECTUAL

We propose to add 7 refinement labels for the Adverbial (D) category in the UCCA scheme. The current work is in line with a series of other UCCA refinements (§2), and fills the gap of Adverbials, as subtyping Adverbials in particular can refine UCCA annotation to be more informative and precise (§3). After providing the motivation to study Adverbials, we propose a subcategorization scheme with explanations (§4).¹ Finally we present pilot results of annotation on UCCA data and IAA measures (§5),² followed by discussions of difficulties, complexities, and limitations (§6).

2 Background

Adverbs. It is important to differentiate the *semantic adverbial* category from the morpho-syntactic class of *adverbs* (Dixon, 2010). In UCCA, for example, the current Adverbial (D) material ranges from manner adverbs and expressions of negation to comparatives to semi-contentful light verbs (Hershcovich et al., 2019).

UCCA categories. The annotation of the UCCA scheme is organized into layers. In the foundational layer, each passage is broken down into scenes,

¹A complete annotation manual is available at https://github.com/IvyWang13/uca_adverbials_dmr/blob/master/UCCA_DMR_annotation_guidelines.pdf

²All data used for the pilot annotation can be found at https://github.com/IvyWang13/uca_adverbials_dmr.

which in turn are made up of scene evokers labeled either State (S) or Process (P), core argument units called Participants (A), which may be either non-scenes (i.e., roughly, ‘entities’) or scenes themselves, and modifiers such as Time (T) or Adverbial (D) units (Abend et al., 2020). Adverbials can be understood as denoting secondary predications over a scene’s main State or Process relation.

Other extensions to UCCA. Building upon the UCCA foundational layer, there have been several efforts and proposals to refine annotations of specific constructions such as implicit arguments (Cui and Hershcovich, 2020), numeric expressions (Cui and Hershcovich, 2021), coreference (Prange et al., 2019b), and adpositional phrases (Prange et al., 2019a; Shalev et al., 2019). The latter class in particular partially overlaps with the Adverbial category, as discussed in §6. This paper will accompany these annotation refinements, with a special focus on the Adverbials.

3 Motivation

Adverbial category is relatively broad. As mentioned briefly, Adverbials represent a wide range of semantic functions; this level of diversity is not shared by other foundational layer categories. Consider, for example, the category ‘Time’: it is also a scene-level category, but has a much more specific semantic purpose. The other categories like ‘State,’ ‘Process,’ and ‘Participant,’ also have better-delimited/dedicated functions, whereas Adverbials can target all general non-scene-evoking elements. Some Adverbials are manner adverbs describing an action, such as *He spoke softly*; some are modal auxiliaries such as *should, might*; others involve constructions expressing aspectual information such as *continue to...* Not all Adverbials are syntactically adverbs, nor do they resemble one another in semantic contributions. This leads to challenges when disambiguating the actual semantic content in sentences in UCCA. Consider (1, 2), repeated here as (5, 6):

(5) The chicken was [**deliciously** D] [cooked P].

(6) The chicken was [**not** D] [cooked P].

Clearly, their meanings are drastically different, but without subcategorization of the D-instances, they receive the same semantic annotation in the UCCA foundational layer. Adverbials (negation, quantification, modality, manner, etc.) fulfill fundamentally different functions with respect to *how*

the meaning of the rest of the sentence is interpreted. Because Adverbials (D) cover wide semantic ground, refining the semantic annotation can thus be more informative, along similar lines as refining Participants with semantic roles (Prange et al., 2019a; Shalev et al., 2019).

Foundational layer annotations are inconsistent. Beyond the domain of Adverbial-proper, another key motivation of D-refinements is to identify sources of disagreement or confusion among the foundational layer categories. While the detailed UCCA annotation guidelines serve as a point of reference when choosing between categories, many borderline cases remain. For example, the below D-instance *during the last week* was identified during our refinement annotation as a borderline case between D-Aspectual and T (Time), despite being annotated as D.

(7) Assets ... [grew P] by \$1.5 billion [**during the latest week** D], to \$352.7 billion.

With the proposed refinement, we hope to eliminate such confusion as much as possible, suggesting revisions to the current guidelines as a whole. The goal is to clarify differences between categories such as D-Aspectual vs. Time, D-Possibility vs. Ground, etc., the former being in the extension layer and the latter being in the foundational layer.

Another aspect of foundational layer inconsistency lies in whether or not to treat a unit as unanalyzable, i.e. omitting further internal structure in multiword instances, such as *one by one, have to, no longer, no matter how*, etc. In the current task, certain Adverbial multiword phrases are encountered with different internal structures. While the current refinement does not depend or bear on Adverbial internal structures, this is an opportunity for annotators to flag such inconsistencies. With our proposed refinements added, we thus expect the foundational layer consistency to increase.

4 Proposed Annotation Scheme

4.1 Semantic Types of Adverbials

We propose seven Adverbial refinement labels (Table 1). Below we provide a synopsis of all proposed categories with a few examples attached. The set of labels aims to subdivide the Adverbial category by semantic functions such that it has a comparable level of granularity as its peers. We strive to draw clear lines among the categories, with each category serving one distinct semantic function as

a modifier in the sentence. The origin of this label set is empirical: we examined and refined the Adverbial instances in the UCCA *Twenty Thousand League Under the Sea* (English) corpus (Abend and Rappoport, 2013), the source of many of the enumerated examples throughout the paper. So far the current scheme is specific to English—we leave cross-linguistic generalizations to future research. The ultimate goal for these labels is to equip UCCA with a useful extension with which annotators can pinpoint a specific label rather than grouping distinct modifier items into a catch-all category.

We also add that, in the current proposal, we work with and conform to previous foundational layer categories and do not change the existing analyses, but we expect our annotated data to be useful for detecting and correcting foundational layer inconsistencies in future work (§3).

4.1.1 Aspectual

ASPECTUAL is a label for D-instances that indicate the start-/end-points of a state or event. It also includes an event’s duration or repetition, whether it is temporary or habitual, and event quantification.

- (8) Captain Baker [**at first** D] [thought P] he was in the presence of an unknown reef. ASPECTUAL
- (9) I would not [go P] there [**again** D]. ASPECTUAL

4.1.2 Causal

CAUSAL is reserved for Adverbials conveying causes or conditions. Examples include words that mention cause or reason, and phrasal causatives like the verbs *make*, *cause*, *ensure*, *prevent*.

- (10) There was good [**reason** D] to [stop and think P]. CAUSAL
- (11) I [**let** D] it [wag P]. CAUSAL
- (12) The material [**prevents** D] water from [leaking P] out of the bag. CAUSAL

Note that the word *prevent* is considered CAUSAL because it signifies ‘causing-to-not-happen.’³

4.1.3 Comparison

COMPARISON includes explicit comparative and superlative markers, e.g., *more*, *most*, and *less*.

³However it does not receive the NEGATION label because it does not include additional or explicit negation; see below.

- (13) ...for the world’s [**most** D] [emotionless S] man. COMPARISON
- (14) That a private individual had such a mechanism at his disposal was [**less than** D] [probable S]. COMPARISON⁴
- (15) This is [**tantamount** D] to [saying P]... COMPARISON

4.1.4 Degree

DEGREE is for words and phrases that convey the extent of an action or a state.

- (16) That would [**profoundly** D] [affect P] public opinion. DEGREE
- (17) You [get P] [**just** D] the collision we need to cause the specified catastrophe. DEGREE

4.1.5 Description

DESCRIPTION is used to label additional properties, features, attributes with which an action or state happens.

- (18) ...those sheets of water had been [**violently** D] [churned P]. DESCRIPTION
- (19) ...for [its P...] [**shrewd** D] [management ...P]. DESCRIPTION

Degree vs. Description. Ideally, the semantic functions of DEGREE and DESCRIPTION do not overlap. DEGREE is reserved for adverbials that convey the sheer extent (low or high) of the predicate, or in UCCA’s terms, the State or Process. DESCRIPTION gives one (or more) additional, discrete property to the State or Process. It helps to formulate DEGREE in terms of reinforcing a scale dimension upon the property, as is often done in formal semantics literature (Kennedy and McNally, 2005). A DEGREE adverbial should explicitly give the information of ‘where’ on the scale of property P the target is. The scale of the property P must have already been provided by the context. On the other hand, DESCRIPTION is the material that provides the property, which is then used to modify the State or Process. For instance, in the phrase ‘the big surprise,’ it is more desirable to analyze ‘big’ as DEGREE, whereas in the phrase ‘the big news’

⁴While the word ‘probable’ would qualify as D-POSSIBILITY if it were annotated as Adverbial in the foundational layer (emphasizing the ‘availability of a mechanism’ as the main scene), we conform to the existing annotation of ‘probable’ as the scene’s main (stative) predicate in our pilot study.

Type	Description	Examples
Aspectual	Describing the start-/end-points, duration, or repetition of S/P	<i>later, still, three times, constantly</i>
Causal	Causatives and conditions	<i>had it not been, cause, let, make</i>
Comparison	Involves comparison upon the said S/P	<i>more, most, less</i>
Degree	Describing the extent of S/P	<i>extremely, colossal</i>
Description	Descriptives, bounding S/P to a certain property or dimension or evaluation, semantically contentful	<i>fast, feverishly, (was) busy (doing...)</i>
Negation	Involves uses of negation	<i>without, nothing, no longer</i>
Possibility	Conveys modality or possible events, thoughts, expectations	<i>want, will, actually, must, attempt (to do...)</i>

Table 1: Proposed semantic functions across UCCA Adverbials; S stands for State and P stands for Process

the word ‘big’ is analyzed as DESCRIPTION. In the former, the property of being a surprise is provided by the word ‘surprise,’ and the extent of surprise is given by the word ‘big.’ In the latter, the word ‘big’ is giving a new size attribute to the news, hence the label DESCRIPTION.

- (20) In an instant the frigate’s deck would become [**densely** D] [populated P]. DEGREE

‘Densely’ in (20) is annotated as DEGREE. With the predicate being ‘populated,’ one could roughly ask the question ‘How much is this deck populated?’, in which case ‘densely’ corresponds better to DEGREE than to DESCRIPTION. Meanwhile, DESCRIPTION labels are commonly found with adjectives and manner adverbs. DESCRIPTION can also be a partial verb phrase, when it adds additional information to the scene-evoking unit.

- (21) Several people [**did me the honor of** D] [consulting P] me on the phenomenon in question. DESCRIPTION

A tricky case for English (arguably extensible to other languages as well) is the word ‘well.’ When used as an adverb, ‘well’ can either denote a high degree, or ‘good’ in terms of description—sometimes both. Annotators should read the context carefully to arrive at a decision.

- (22) [**well** D] [received P] in scholarly circles. DESCRIPTION
- (23) [**well** D] [versed P] in the theory of classification. DESCRIPTION
- (24) The frigate [kept P...] [**well** D] [out ... P]. DEGREE
- (25) I was [**well** D] [satisfied S] with my cabin. DEGREE

4.1.6 Negation

The NEGATION label is applied to D-instances with explicit negation morphology.

- (26) [**No** D] business dealings have been [crowned P] with greater success. NEGATION
- (27) It did [**n’t** D] seem due for [resurrection P]. NEGATION

It might be argued that NEGATION is more dependent on morphological than semantic criteria, as least for English. But, we see NEGATION as a necessary category here that should not be subsumed under other labels, because its function indeed changes the meaning of a sentence or utterance, and gives the Adverbial instance a unique, irreplaceable flavor.

4.1.7 Possibility

POSSIBILITY not only includes modal auxiliaries (*would, could, should, must*, etc.), but also all D-instances that convey possible outcomes, attempts, thoughts, and expectations.

- (28) ...the Moravian [**apparently** D] [undamaged P]. POSSIBILITY
- (29) It [turned out to be S] [a foolish business A] [**after all** D]. POSSIBILITY
- (30) In this event I [**would be inclined to** D] [accept P] the existence of a giant narwhale. POSSIBILITY
- (31) Oh [**really** D]? POSSIBILITY

Some D-instances can be interpreted as including a certain level of speaker belief, but they are not annotated as POSSIBILITY in the current schema, if they serve a more salient contextual modification function that assigns them to another category.

- (32) He [took a P...] [**definite** D] [liking ... P] to me. DEGREE

- (33) Our server was [**quite** D] [attentive S] and the food was fantastic. DEGREE
- (34) Some [saw P] it [**purely** D] as a scientific problem to be solved. DEGREE

In (32), *definite* conveys the extent of the liking, and from context the phrase *definite liking* is not immediately read as having strong speaker beliefs. Similarly, in (33) and (34), the most immediate interpretations are factual, rather than intensional or possible. In general, if there is a more “factual” meaning available, annotators are encouraged to avoid over-interpretation and choose the most apparent reading.

Empirically, one may encounter more ambiguous, borderline examples, such as (35),⁵ and in those cases one should exercise their judgment to evaluate whether the element of POSSIBILITY outweighs that from another label.

- (35) At the front door of his office, I [**nearly** D] [turned P] around.

4.2 Multiple labels

A particular feature of the current subcategorization scheme is the ability to add multiple labels to one Adverbial instance. The reason for a multi-label scheme is that there is not necessarily a one-to-one mapping of semantic function and linguistic construction. English shows frequent uses of one construction denoting multiple functions, and this might be even more common in other languages.

Some functions, such as NEGATION and COMPARISON, are easily layered on top of other functions such as DESCRIPTION or DEGREE by means of, e.g., morphology. Meanwhile there are also D-instances that cover both DESCRIPTION and DEGREE domains, for example. In order to accommodate all of these cases, the current scheme proposes that subtypes can be combined.

Co-occurrence with Negation. The label NEGATION can be added in parallel with other labels.

- (36) I [**no longer** D] [left P] the ship’s deck. NEGATION, ASPECTUAL
- (37) ...[**nothing more than** D] to [go forth P]. NEGATION, COMPARISON

However, we only consider phrases with explicit phrasal negation, and do not include scene-evoking

elements that have morphologically or semantically apparent antonyms.

- (38) Master ’s eyes would kindly [**stop** D] [bulging P]. ASPECTUAL
- (39) They were [**unable** D] to [say P]. DESCRIPTION

Co-occurrence with Comparison. Like NEGATION, COMPARISON is often added upon another semantic function. Due to morphological properties of English, some comparatives take a form of suffixation on one word unit, some do not. Annotators should take care and add the label COMPARISON when appropriate.

- (40) You’d be [squashed P] [**as flat** D]. COMPARISON, DESCRIPTION
- (41) No whaling vessel could have been [**better** D] [armed S]. COMPARISON, DESCRIPTION

Other co-occurrences. As mentioned, there are also Adverbials, often phrasal, that combine other categories. For example:

- (42) [Attitude D] of staff was [[**very** E] [**bad** C] D]. DESCRIPTION, DEGREE
- (43) The store is clean, [run P] [[**very** E] [**professionally** C] D] and a pleasure to be in. DESCRIPTION, DEGREE

In (42) and (43), the labels DESCRIPTION and DEGREE are combined because both examples include both labels, as exemplified by the degree modifier ‘very’ and the descriptive words that follow.

It is also worth noting that (42) and (43) both show Adverbial internal structure. Namely *very* is Elaborator, and *bad*, *professionally* are both Center of the Adverbial instances. For the current task, we treat the D-instances flatly, leaving to future research the possibility of potentially refining the internal categories in a similar fashion.

4.3 Verbs

Verbs that are not identified as main predicates form a substantial part of the Adverbial category. These are verbs that precede the main scene-evoking State or Process, or verbs of subevents at the scene level. Verbs constitute a syntactic category that is semantically diverse on its own, so they warrant further explication for ease of annotation. Currently, we have identified several groups of verbs that could be fit into our categories.

⁵In fact this sentence comes from one of our annotation studies, with one POSSIBILITY and two DEGREE labels.

First, verbs relating to event start or end points, or that situate the main action in a temporal dimension, should be labeled ASPECTUAL. Examples include *begin*, *start*, *finish*, *complete*, *continue*, etc.

(44) The day was [**beginning** D] to [break P].

(45) He [**postponed** D] our [meeting P] until next Monday.

Second, verbs of causes or preventions should be CAUSAL, such as *make*, *force*, *cause*, *tempt*; *let*, *permit*, *allow*, *prevent*, *avoid*, *spare*, etc.

(46) Those dogs are not [**allowed** D] to [enter P].

(47) Candy [**helped** D] Ann [bake P] a cake.

Third, verbs of wanting, needing, or trying should be POSSIBILITY.

(48) I never [**expected** D] to [win P] that prize.

(49) She [**pretended** D] to be [looking P] away from the disaster.

Lastly, verbs that do not clearly fit into one of the above categories are likely subevents that involve more concrete actions, and should belong with the DESCRIPTION category.

(50) “What about master’s collections?” Conseil [**ventured** D] to [observe P].

(51) He [**received** D] a [call P].

(52) ...how many times I [**shared** D] the [excitement P] of general staff and crew.

5 Agreement Studies

As a pilot to verify the effectiveness of the subcategorization, three computational linguists familiar with UCCA independently annotated Adverbial subcategories. Two sets of annotation were conducted. The first set involved two annotators. Difficulties encountered were used to revise the annotation guidelines, and then a second data sample was annotated by all three annotators.

	# Annotators	# D-instances
Set 1	2	167
Set 2	3	122
Total	3	289

Table 2: Annotation set statistics.

Category	Annotator 1		Annotator 2		Ann. 3
	Set 1	Set 2	Set 1	Set 2	Set 2
Description	66	32	54	25	19
Negation	15	19	15	19	19
Possibility	26	17	22	18	17
Degree	20	24	25	25	23
Aspectual	27	18	29	20	21
Comparison	1	2	2	4	3
Causal	2	1	10	4	5
Comp.+Desc.	5	3	1	3	2
Neg.+Asp.	4	—	4	—	—
Desc.+Deg.	1	3	2	1	—
Poss.+Neg.	—	—	1	—	—
Comp.+Deg.	—	2	—	—	1
Asp.+Caus.	—	—	—	1	1
N/A	—	1	2	2	11
Total	167	122	167	122	122

Table 3: Subcategory counts for annotation sets 1 & 2. Annotator 3 did not participate in Set 1 annotation. N/A refers to a judgment that the unit should not have been considered an Adverbial in the foundational layer.

5.1 Data

For both annotation sets, we used unseen sentences from UCCA-annotated English Web Treebank and Wikipedia corpora (Abend and Rappoport, 2013). We randomly sampled 100 sentences for each round of annotation, and annotated 289 Adverbial instances out of the 200 sentences (Table 2).

5.2 Results

The count results for two annotation sets are shown in Table 3. In the first annotation set, the largest class by far was DESCRIPTION, which was partially due to the sampled sentences, but partially also due to lack of clarity in the guidelines for other categories. Annotators tended to default to DESCRIPTION when uncertain about a more distinct semantic function. This led to a revision to the guidelines, which then was used in the second round. Among the remaining categories, NEGATION, POSSIBILITY, DEGREE, and ASPECTUAL are all fairly frequent and balanced, whereas COMPARISON and CAUSAL are much rarer. The scheme should be examined on more unseen data in the future to determine how these cases (COMPARISON and CAUSAL) can be better handled as Adverbials or in UCCA more generally. We also add a row for “N/A” instances—D-instances which the annotator(s) deemed to be foundational layer misannotation, or for which the annotator(s) did not find a suitable fit in the current scheme (see §6).

The inter-annotator agreement results for each

Category	Set 1		Set 2
	κ	F1-score	α
Description	0.76	0.85	0.70
Negation	0.97	0.97	1.00
Possibility	0.78	0.82	0.87
Degree	0.81	0.83	0.79
Aspectual	0.85	0.88	0.84
Comparison	0.66	0.67	0.95
Causal	0.32	0.33	0.57
N/A	0.00	0.00	0.26
Avg. with N/A	0.64	0.84	0.74
Avg. without N/A	0.74	0.84	0.82

Table 4: Inter-annotator agreement results for Sets 1 & 2. Since Cohen’s κ does not support multi-label input, we provide average across the categories for it (macro-average) whereas for F1 score we provide micro-average across the instances rather than the categories. For Krippendorff’s α we provide macro-average.

	A1 & A2		A1 & A3		A2 & A3	
Exact Match	104	85.2%	93	76.2%	99	81.1%
+ Partial Match	108	88.5%	98	80.3%	102	83.6%
+ N/A	109	89.3%	99	81.1%	104	85.2%
Cohen’s κ	0.85		0.77		0.82	

Table 5: Pairwise raw agreement for Set 2. **Row 1:** ‘Exact Match’ means two annotators agreed completely, regardless of multi-label or not. Percentage out of all 122 instances. **Row 2:** ‘Partial Match’ indicates the annotators did not disagree completely. This includes the scenario where one annotator assigned one label and the other assigned multi-label and one label overlaps. Percentages out of 122 instances. **Row 3:** We then add the scenario where both annotators assigned N/A to the instance. Percentages out of 122 instances. **Row 4:** Pairwise Cohen’s κ by exact match.

category are shown in Table 4. For the first set with two annotators, we report Cohen’s κ as well as F1 measure, and for the second set, with an additional annotator, we report Krippendorff’s α as one metric for inter-annotator agreement. Table 5 reports more details on the second set, which involved three annotators. We report counts of different kinds of agreement/disagreement, as well as pairwise Cohen’s κ .

Overall we observe good agreement among the annotators. NEGATION has the highest rate of agreement. This is likely due to the fact that, at least in English, negation morphology (*no*, *not*, *n’t*) is easy to identify, even when it co-occurs with another category, such as ASPECTUAL, i.e. *never*. The categories with the next highest agreement are ASPECTUAL, POSSIBILITY, and DEGREE. The CAUSAL label suffers in agreement, which is likely due to its rarity, as well as annotators’ varying willing to assign this label altogether. The revised

guidelines used in Set 2 improve this somewhat.

6 Discussion

Based on the results, we turn to some specific challenges with D-refinement annotation, as well as discussing how our scheme relates to two similar representation frameworks.

6.1 Challenges

Non-applicable instances. In Table 3, some instances have not been annotated with any refinements (‘N/A’) by at least one annotator. Many such instances are either questionable as Adverbial to start with, or are exceptions that are not covered yet by our proposed guidelines, or both.

- (53) [Back D] to my poor rating - I was excepted to medical school and went in to cancel my membership as I was told I could do so since I was moving away.

For example, (53) contains the word ‘back,’ but given the context, ‘back’ has a salient discourse function, arguably Ground in the foundational layer. (54) is another example: the proform ‘one’ construes an event instance (the event being ‘being someone’s favorite’) as selecting an entity out of a set. This is a kind of ‘atemporal’ event quantification, and while our ASPECTUAL label perhaps comes closest, some annotators did not assign any label.

- (54) Little Women was [one D] of Hepburn’s personal [favorites P].

The current guidelines lack a clear direction on how to deal with such cases, as they become empirically discovered. We will keep updating and improving the guidelines in future work.

Verbs as Adverbials. While the guidelines strive to cover many verbs, during our pilot annotation there are still a number of verbs that have caused some disagreement.

- (55) The only thing I [found D] [edible s] were the potato wedges, I finally gave up, he kept trying - he [found D] the fried wantons to be [OK s].

In (55), there are two occurrences of the verb ‘found,’ both of which are Adverbials. The annotators had difficulty classifying these verbs: Annotator 1 labeled the first ‘found’ as DESCRIPTION and

the second as POSSIBILITY, annotator 2 labeled both as POSSIBILITY, while annotator 3 labeled N/A for both.

- (56) The album’s hard rock edge [**included** D] conspicuous guitar [contributions P] from Robert Fripp, ... and Tom Verlaine.

In (56), while annotators 1 and 2 labeled ‘included’ as DESCRIPTION, annotator 3 deemed it N/A.⁶ Future annotators should especially take care with verbal Adverbials, try to paraphrase them if possible—if no categories seem to fit, one should flag the instance as an exception, which will be aggregated for future improvements of the refinement.

Multiple labels. Another source of disagreement is multiple annotation. Interestingly, when two annotators agree that an instance has multiple labels, they tend to also agree on *which* labels to choose (e.g., NEGATION+ASPECTUAL (57)). However, Adverbial instances that require more than one refinement are relatively rare to begin with, and annotators do not always agree on *when* this is the case. For example, consider (58).

- (57) I would [**never** D] [sing P] in public.
NEGATION, ASPECTUAL
- (58) So don’t get taken in, [**keep** D] your eyes [open S] if you choose to shop here.

For ‘keep’ in (58), annotator 1 only labeled ASPECTUAL, while annotators 2 and 3 labeled ASPECTUAL+CAUSAL, citing the reason that ‘keep’ is equivalent to ‘causing to stay.’ In general, annotators should semantically decompose each instance in this way before making a decision. As more of such potentially multi-labeled instances occur, further iterations of the guidelines will add more concrete directions on the extent of semantic decomposition with D-refinements.

Clarifying foundational layer decisions. We find two cases where other foundational layer categories and D are not reliably distinguished. First, we find that D-Aspectual and Time (T) overlap: both have been applied to instances of frequency and duration, e.g. (59, 60). Empirically, D-Aspectual also includes event quantification (61), change-of-state verbs (62), as well as temporally unspecific indicators of recency (or urgency, or lateness) (63).

- (59) John will [come P] [**at around eight** T].
- (60) John will [come P] [**regularly** T].
- (61) John will [come P] [**for the second time** D].
ASPECTUAL
- (62) Information [**becomes** D] more [abundant D].
ASPECTUAL
- (63) You’d [**just** D] been [yanked P]. ASPECTUAL

Second, some foundational layer Adverbials serve similar functions as the GROUND category. According to the UCCA Foundational Layer guidelines, GROUND serves to “relate some unit to its speech event”, i.e., a pragmatic element occurring with the scene. However, there are many edge cases that blur the the line between pragmatics and semantics and in those cases GROUND may overlap with certain (roughly, epistemic or deontic) modal usages of D-POSSIBILITY. For example, the units in (64) and (65) are annotated as GROUND and D-POSSIBILITY respectively.

- (64) You can [go P], [**for all I care** G].
- (65) He [**reportedly** D] [told P] Hepburn...
POSSIBILITY

The similarities between D-ASPECTUAL and TIME, as well as D-POSSIBILITY and GROUND, respectively, invite future investigations into the boundaries between these categories.

6.2 Relation to Other Representations

AMR. In Abstract Meaning Representation (AMR) and its proposed extensions (Banarescu et al., 2013; Donatelli et al., 2018; Bonial et al., 2018), a sentence is labeled with relations and concepts in a rooted, directed, acyclic graph.⁷ The current work has overlaps with AMR concepts—at different levels of the AMR graph—some of which can even be said to be close mappings, but overall our proposed scheme is designed to be compatible with UCCA’s layered annotation structure rather than already-fine-grained AMR graphs. For example, in (66) and (67) we juxtapose the AMR and D-refined UCCA annotations of a sentence, respectively. While D-POSSIBILITY maps to the AMR concept possible-01 and D-NEGATION maps to :polarity -, their structures remain distinct (the main predicate in AMR is possible-01, whereas

⁶Annotator 3 noted in a comment that DESCRIPTION is closest if “N/A” was not an option.

⁷Retrieved from <https://github.com/amrisi/amr-guidelines/blob/master/amr.md>

in UCCA it is [go P]). We leave more thorough comparison of the two schemes to future research.

(66) The boy can't go.

```
(p / possible-01
  :polarity -
  :ARG1 (g / go-02
    :ARG0 (b / boy)))
```

(67) [The boy A] [can D-POSSIBILITY] ['t D-NEGATION] [go P]

SNACS. SNACS is a token-based adpositional sense-disambiguation scheme that is known to integrate well with UCCA Foundational Layer annotation labels (Schneider et al., 2018; Prange et al., 2019a). As a refinement layer, the current scheme is also compatible with SNACS adpositional categories to a certain degree. The overlap is notably seen in adpositional phrases that as a whole are also labeled as D category. For instance, the MANNER category in SNACS would map onto D-DESCRIPTION instances where the instance is a prepositional manner modifier.

(68) Its paddle wheels were [churning P] the sea [with perfect steadiness D]. DESCRIPTION

7 Conclusion

Finding an appropriate level of abstraction is a central problem in meaning representation design. Motivated by the observation that UCCA's Adverbial category encompasses too wide a range of *heterogeneous* semantic functions as a scene-level category, we proposed a semantic refinement annotation scheme consisting of 7 subtypes, allowing multiple annotation to cover a limited amount of compositionality. Pilot annotation studies on corpora previously annotated with the UCCA foundational layer showed that this subcategorization scheme offers a good coverage of the existing Adverbial instances.

Some challenges remain. Though rare, some Adverbial instances were marked by annotators as non-applicable under the scheme. We also found that allowing multiple labels can cause annotator disagreement depending on annotators' extent or style of analysis. As more empirical data accrue, future work should aim to mitigate such sources of disagreement.

Acknowledgements

We sincerely thank Daniel Hershcovich as well as three anonymous reviewers for their constructive criticism and feedback on this paper.

References

- Omri Abend and Ari Rappoport. 2013. [Universal Conceptual Cognitive Annotation \(UCCA\)](#). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 228–238, Sofia, Bulgaria. Association for Computational Linguistics.
- Omri Abend, Nathan Schneider, Dotan Dvir, Jakob Prange, and Ari Rappoport. 2020. [UCCA's foundational layer: Annotation guidelines v2.1](#). *arXiv preprint arXiv:2012.15810*.
- 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 *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.
- Claire Bonial, Bianca Badarau, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Tim O'Gorman, Martha Palmer, and Nathan Schneider. 2018. [Abstract Meaning Representation of constructions: The more we include, the better the representation](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Ruixiang Cui and Daniel Hershcovich. 2020. [Refining implicit argument annotation for UCCA](#). In *Proceedings of the Second International Workshop on Designing Meaning Representations*, pages 41–52, Barcelona Spain (online). Association for Computational Linguistics.
- Ruixiang Cui and Daniel Hershcovich. 2021. [Meaning representation of numeric fused-heads in UCCA](#). *arXiv preprint arXiv:2106.07364*.
- R. M. W. Dixon. 2010. *Basic Linguistic Theory*.
- Lucia Donatelli, Michael Regan, William Croft, and Nathan Schneider. 2018. [Annotation of tense and aspect semantics for sentential AMR](#). In *Proceedings of the Joint Workshop on Linguistic Annotation, Multitext Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 96–108, Santa Fe, New Mexico, USA.
- Daniel Hershcovich, Omri Abend, and Ari Rappoport. 2019. [Content differences in syntactic and semantic](#)

[representation](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 478–488, Minneapolis, Minnesota. Association for Computational Linguistics.

Christopher Kennedy and Christopher Louise McNally. 2005. Scale structure, degree modification, and the semantics of gradable predicates. *Language*, 81:345–381.

Jakob Prange, Nathan Schneider, and Omri Abend. 2019a. [Made for each other: Broad-coverage semantic structures meet preposition supersenses](#). In *Proceedings of CoNLL*, pages 174–185, Hong Kong, China. Association for Computational Linguistics.

Jakob Prange, Nathan Schneider, and Omri Abend. 2019b. [Semantically constrained multilayer annotation: the case of coreference](#). In *Proceedings of DMR*, pages 164–176, Florence, Italy.

Nathan Schneider, Jena D. Hwang, Vivek Srikumar, Jakob Prange, Austin Blodgett, Sarah R. Moeller, Aviram Stern, Adi Bitan, and Omri Abend. 2018. [Comprehensive supersense disambiguation of English prepositions and possessives](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 185–196, Melbourne, Australia. Association for Computational Linguistics.

Adi Shalev, Jena D. Hwang, Nathan Schneider, Vivek Srikumar, Omri Abend, and Ari Rappoport. 2019. [Preparing SNACS for subjects and objects](#). In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 141–147, Florence, Italy.