Cross-lingual annotation: a road map for low- and no-resource languages

Meagan Vigus[1], Jens E. L. Van Gysel[1], Tim O'Gorman[2], Andrew Cowell[3], Rosa Vallejos[1], and William Croft[1]
[1]Department of Linguistics, University of New Mexico
[2]College of Information and Computer Sciences, University of Massachusetts Amherst
[3]Department of Linguistics, University of Colorado, Boulder
{mvigus,jelvangysel,rvallejos,wcroft}@unm.edu,
togorman@cs.umass.edu, james.cowell@colorado.edu

Abstract

This paper presents a "road map" for the annotation of semantic categories in typologically diverse languages, with potentially few linguistic resources, and often no existing computational resources. Past semantic annotation efforts have focused largely on high-resource languages, or relatively low-resource languages with a large number of native speakers. However, there are certain typological traits, namely the synthesis of multiple concepts into a single word, that are more common in languages with a smaller speech community. For example, what is expressed as a sentence in a more analytic language like English, may be expressed as a single word in a more synthetic language like Arapaho. This paper proposes solutions for annotating analytic and synthetic languages in a comparable way based on existing typological research, and introduces a road map for the annotation of languages with a dearth of resources.

1 Introduction: Cross-linguistically informed semantic annotation

In recent years, there has been a surge of interest in annotation schemes that allow natural language texts to be parsed into semantic representations usable for information extraction, machine translation, and other downstream purposes. Good results have been achieved in automatically parsing natural language texts into Abstract Meaning Representations (Banarescu et al., 2013, AMR) and Discourse Representation Structures (Kamp and Reyle, 2013; Bos et al., 2017, DRS), among others, as demonstrated in various shared annotation tasks (Abzianidze et al., 2020; May and Priyadarshi, 2017).

However, most efforts in developing and testing such annotation schemes have focused on a restricted set of (typically Indo-European) languages with large native speaker populations. For example, the shared annotation tasks reported on in Abzianidze et al. (2020) and May and Priyadarshi (2017) were all based on English. Large AMR corpora exist, to our knowledge, only for English and Mandarin - both morphologically isolating languages, with comparatively little inflectional and derivational morphology. PropBank (Palmer et al., 2005) has been extended to large-scale languages with derivational morphology, such as Hindi and Arabic. But the annotation of derivational morphology relies on a thorough documentation of its role in the language, which often isn't available for low-resource languages. For morphosyntactic annotation, this bias is less apparent: the Universal Dependencies project (de Marneffe et al., 2014) has annotated treebanks from 96 languages, representing 20 linguistic families. However, Indo-European languages are disproportionately represented (53/96 languages). In terms of native speaker populations, 62 out of the 96 UD languages have more than 1 million native speakers, belonging to the largest 6% of languages in the world (Eberhard et al., 2020). Only 24 UD languages have relatively small native speaker populations (10 UD languages are ancient languages).

This apparent bias in languages represented in computational linguistic work likely has consequences for the structure of annotation schemes. The World Atlas of Language Structures chapter on the morphological structure of verbs (Bickel and Nichols, 2013) looks at a sample of 145 languages and finds that, on average, languages express 5.52 inflectional categories within the verb. The 19 UD languages

Proceedings of the 2nd International Workshop on Designing Meaning Representations, pages 30–40
Barcelona, Spain (Online), December 13, 2020

with more than 10 million native speakers that are also included in Bickel and Nichols' (2013) sample express only an average of 4 inflectional categories within the verb - exemplifying a known correlation between morphological complexity and demographic factors (Lupyan and Dale, 2010). The annotation schemes developed in the context of high-population, typically Indo-European languages, may therefore not carry over well to smaller-scale, often more morphologically complex languages. Many smaller-scale languages do not have a long history of linguistic analysis, and therefore understanding of their structure may be progressing in tandem with annotation efforts.

Expanding annotation efforts to such morphosyntactically diverse languages may, apart from simply expanding typologically sound coverage of annotation efforts, improve the overall utility of annotation schemes. Cross-linguistic annotation schemes must incorporate a certain amount of flexibility in order to deal with differences in conventionalized semantic distinctions, and the morphosyntactic expression of these distinctions. This flexibility in design can also benefit monolingual annotation, by allowing for flexibility with annotators who have different levels of linguistic training.

For that reason, this paper proposes solutions for extending AMR to as many languages as possible, including a "road map" for languages with few existing resources. This provides a starting point for flexible but consistent annotation of a number of semantic categories. It also describes how the cross-linguistic diversity in morphosyntax, paired with pre-existing linguistic analyses and resources, can inform the design of a flexible annotation process. This road map provides steps towards more detailed semantic annotation, as the linguistic analysis of the language progresses and computational resources are created, in order to eventually arrive at the same level of specificity in annotations as in high-resource languages. The creation of a comparable cross-linguistic semantic annotation scheme is of course a larger topic than can be covered in a single paper; this paper sets forth a general approach for dealing with differences in linguistic properties and resource availability (the road map), and specific annotation solutions for certain semantic categories and morphosyntactic phenomena (e.g., synthesis).

In this paper, examples are drawn from three no- or low-resource languages: Sanapaná, Kukama, and Arapaho. Sanapaná (Enlhet-Enenlhet) has about 1000 native speakers living in Paraguay. Aside from an ongoing documentation project (Van Gysel, 2020), there are only exploratory analyses of the morphology (Gomes, 2013; Van Gysel, 2017). Kukama (Tupian) has about 1000 native speakers living in Peru. Existing linguistic resources include a descriptive grammar (Vallejos, 2016), a Kukama/Spanish bilingual dictionary (Vallejos and Amías, 2015), translated and morphologically analyzed texts (Vallejos, 2014), and some pedagogical materials. Arapaho (Algonquian) is spoken by two communities in the United States, the Northern Arapaho in Wyoming and the Southern Arapaho in Oklahoma. Among the Northern Arapaho, there are around a hundred native speakers, and several hundred with passive knowledge of the language. Linguistic resources include a grammar (Cowell and Moss Sr, 2008), an online lexical database with detailed part-of-speech labelling and argument structure information, and an annotated text database of nearly 100,000 sentences with accompanying audio and/or video.

2 Cross-linguistic annotation: Typological issues

Certain typological issues arise when constructing a semantic annotation scheme that can, in theory, be applied to any language. Three general types of issues are described here.

First, some types of morphosyntactic differences do not hinder the annotation of semantic information, and can therefore largely be ignored in a semantic annotation scheme. For example, languages may indicate grammatical roles via constituent order or case affixation of argument phrases, but argument phrases in both types of languages can be annotated for their semantic roles in the same way.

Next, there are major typological differences in the conventionalized semantic distinctions that languages make in their grammar, i.e. how languages 'carve up' conceptual space. For example, some languages distinguish only SINGULAR from NON-SINGULAR nominal number, other languages distinguish SINGULAR, DUAL, TRIAL, and PLURAL (more than three), still other languages have a FEW (including singular) vs. MANY nominal number system (Corbett, 2000, chapter 2). For these types of semantic differences, the use of lattices of category values has been proposed to allow flexible but consistent annotation (Van Gysel et al., 2019); we adopt this approach and incorporate it into the road map in §3.

Finally, languages differ in terms of how concepts are packaged into words and sentences. As discussed in §1, languages that are more synthetic, packaging many morphemes/concepts into a single 'multiconcept' word, have not been well-represented in past annotation efforts. This also presents a practical issue: for languages at an earlier stage of documentation, it may not be possible for annotators to morphologically decompose multiconcept words. Therefore, the issue of how to maintain consistent annotation across both more analytic and more synthetic languages will be the main focus of this section.

Even 'word' does not have a consistent definition across languages. Most languages have a language-internal concept of 'word', at least as a cognitively salient unit of the language (Bolinger, 1963). But these units do not share consistent linguistic traits across languages, nor is there a widely-accepted definition of what should constitute a word across languages (Dixon et al., 2002), but see Zingler (2020).

2.1 One predicate instead of two

In many languages, a single verb with derivational morphology may express what is expressed by two verbal words (e.g.., main verb, complement, auxiliary) in English and other analytic languages. In general, we treat derivational morphology as a single predicate along with the verb to which it attaches. Derivational morphology may express phasal aspect, as in 1 from Arapaho.[1] The aspectual marking, whether an affix or a separate word, is not annotated as a separate predicate, since it selects a phase of the event.[2] Derivational morphology may also express an external causing event, shown in 2 from Kukama. For causatives, either a single event with causative semantics is identified or two events are identified, one for the causing event and one for the caused event. This is based on whether negation can apply to the causing event and caused event separately. For derivational morphology, as in Kukama, negation would scope over both events, meaning that it is construed as a single event and annotated as such. In English, causative auxiliaries can be negated separately from the caused event (e.g., *Grandmother didn't make the kid drink / Grandmother made the kid not drink*); therefore, two events are identified.

(1) ceesisnoo'oebiicitiit.
ceesis-noo'oe-biicitii-t
IC.begin-around-bead.s.t.-3S
'She is starting to bead around it.'

(c / biicitii 'bead s.t.'
:Actor (a / '3S')
:Undergoer (u / '3S')
:aspect Activity
:modstr Aff)

(b / bead s.t.
:Actor (s / she)
:Undergoer (i / it)
:aspect Activity
:modstr Aff)

(2) nai kurata-**ta** churan=ui uni=pu
grandmother drink-**CAU** kid=PST water=INS
'Grandmother made the kid drink the water.'

(k / kuratata 'make drink'
:Causer (n / nai 'grandmother')
:Actor (c / churan 'kid')
:Undergoer (u / uni 'water')
:aspect Performance
:modstr Aff)

(d / drink
:Cause (m / make
:Actor (g / grandmother)
:aspect Performance
:modstr Aff)
:Actor (k / kid)
:Undergoer (w / water)
:aspect Performance
:modstr Aff)

For modality, as shown in 3 from Arapaho, we apply semantic criteria to determine whether a single predicate or multiple predicates are identified; see §3.3 for a discussion of the modal annotation. If the

[1] We present examples with annotations for predicate-argument structure, modal strength and polarity, and aspectual structure; temporal annotations have been omitted. The annotations make use of the general 'Stage 0' participant roles; §3 explains the relevant annotation categories in more detail. Abbreviations used in glosses are the following: 2 = second person; 3 = third person; ALLAT = allative; APPL = applicative; APPRX = approximative; CER = certainty; CAU = causative; DEF = definite; DISTR = distributive; IC = initial change; IMPERF = imperfective; INF = inferred; INS = instrumental; LOC = locative; M = masculine; NARR = narrative; PAS = passive; PL = plural; PST = past; REDUP = reduplication; S = singular; SBJ = subjunctive.

[2] The aspect indicated by the morphology is reflected in the aspect annotation. Inceptive phasal aspect is annotated as ACTIVITY to reflect that the event may be ongoing.

modal can itself be modalized (i.e., appear under the scope of another modal), then it is annotated as its own predicate. Since English allows this (e.g., *they **might want** to take it...*), *want* is annotated as a predicate. But, in the Arapaho, this is not possible, and therefore the modal is annotated in the same predicate as the main verb. While this criterion generally correlates with the expression of the modal as a complement-taking predicate versus an affix on the verb, it relies on semantic criteria that can be applied across languages. In both cases, the modal informs the modal strength annotation of the verb.

(3) xonouu niibeetwon3eiinein.
xonouu
immediately
nii-**beet**-won-3eiin-ein
IMPERF-**want.to**-ALLAT-put.inside.a.place-3S/2S
'Right away he wants to go and put you in jail.'

(n / beetwon3eiin 'want to go and put s.t. inside a place'
:Actor (a / '3S')
:Theme (t / '2S')
:aspect Habitual
:modstr Neut)

(w / want
:Experiencer (h / he)
:Stimulus (g / go
:Actor (h)
:aspect Habitual)
:Stimulus (p / put
:Actor (h)
:Theme (y / you)
:Goal (j / jail)
:aspect Habitual)
:aspect Habitual
:modstr Aff
:modal g
:modal p)

Associated motion is treated similarly. Whether or not motion events are considered a single predicate with the verb or a separate predicate depends on whether locative or directional expressions that occur in the clause correspond to arguments of the motion event (as opposed to arguments or circumstantial locatives modifying the main event). When they are arguments of the motion event, it is identified as a separate predicate; when they are not, it is considered a single predicate with the verb. In the Sanapaná example in 4, no *arrive* predicate is identified in the annotation. The associated motion morphology *-angv-akm* indicates that the seeing event occurs after arriving at a location other than the deictic center. A locative expression can occur with this construction, but there is no evidence that this is an argument of the motion event rather than a circumstantial locative of the *see* predicate. Therefore, only a single predicate is annotated in Sanapaná. In English, this location can be expressed as an unambiguous argument of the motion event (e.g., *we arrived **home** and saw...*), so a separate *arrive* predicate is annotated.

(4) netamen apk-el-vet-**angv**-ay-**akm**-e' hlema nenhlet, ang-kelvana.
afterwards 2/3M-DISTR-see-**LOC**-PST/HAB-**APPRX**-V1.NFUT one person 2/3F-woman
'Afterwards, they arrived and saw a person, a woman.'

(v / engvetangvayam 'arrive and see'
:Experiencer (a1 / apk-el- '3PL.M')
:Stimulus (n / nenhlet 'person'
:mod (a2 / angkelvana 'woman')
:quant 1)
:aspect State
:modstr Aff)

(a / arrive
:Actor (t / they)
:aspect Performance
:modstr Aff)
(s / see
:Experiencer (t)
:Stimulus (p / person
:mod (w / woman)
:quant 1)
:aspect State
:modstr Aff)

2.2 One word containing predicate and arguments

Languages can also package together concepts that cut across the event-participant distinction that is fundamental to semantic annotation schemes that rely on predicate-argument structure, such as AMR. For these types of multiconcept words, namely pronominal indexation and noun incorporation, both a predicate and an argument are identified at all stages of the road map.

In many languages, participants are indexed on the verb; this is often called agreement or pronominal affixation. In certain constructions, participants are signalled only through indexation and not expressed elsewhere in the clause. We treat the indexed participants as pronouns and identify both a predicate and an argument (or arguments) for a single word. This can be seen in examples 1, 3, and 4 above.

Noun incorporation involves a word that expresses both a predicate and an argument. Mithun (1984) identifies four types based on their structure and function across languages. These types of noun incorporation exist on a grammaticalization cline, with languages that exhibit the more grammaticalized types also exhibiting the less grammaticalized types. Example 5 shows Type I incorporation, the least grammaticalized, and 6 shows Type IV incorporation, the most grammaticalized, both from Arapaho.

(5) he'ih'iixooxookbixoh'oekoohuutoono'

He'ih'ii-xoo-xook-
NARR.PST.IMPERF-REDUP-through-
bixoh'oekoohuutoo-no'
act.so.that.hand.appears.quickly-PL

'they were sticking their hands right through them [the ghosts] to the other side'

(b/ bixoh'oekoohuutoo 'stick hands through'
:Actor (a1 / '3PL')
:Theme (t / 'hands')
:Undergoer (g / '[ghosts]')
:aspect Endeavor
:modstr Aff)

(6) hoono' nuhu' tihciinii'eihiinit, he'ih'etoocein nuhu' hitiine' nuhu' hoote.

hoono' nuhu' tih-cii-nii'eihiini-t
not.yet this when.PST-NEG-be.eagle-3.S
he'ih-'et**oocei**n nuhu'
NARR.PST-pull.**rope-like.thing.**out this
hi-tiin-e' nuhu' **hoote**
3S-mouth-LOC this **sinew**

'At the [time] when he wasn't yet an eagle, he took [it] out of his mouth, the sinew.'

(e / 'etoocein 'pull rope-like thing out'
:Actor (a / '3S')
:Theme (h1 / hoote 'sinew')
:Material/Source (h2 / hi-tiin-e' 'his mouth'
:part-of a)
:Temporal (h3 / have-role-91
:ARG0 (a)
:ARG1 (n / nii'eihiini 'be eagle')
:aspect State
:modstr Neg)
:aspect Performance
:modstr Aff)

Type I noun incorporation doesn't allow the addition of a syntactic argument that corresponds to the incorporated noun. Type IV noun incorporation, often called classificatory constructions, incorporates a more general noun into the verb, whose referent can be made more specific by the addition of a syntactic argument in the clause. In the less grammaticalized types of noun incorporation (Types I-III), both a predicate and arguments are identified, as in 5. The more grammaticalized types of noun incorporation, as in 6, are treated like derivational morphology and only a predicate is identified.[3]

2.3 Nonverbal clauses: Different packaging of "predicate" and arguments

Nonverbal clauses, such as locative, possessive, object, and property predication, and equational clauses, vary across languages in terms of how concepts are packaged into words (Stassen, 1997; Stassen, 2009). There are three nonverbal clause strategies, two of which are problematic for the predicate-argument structure of AMR.[4] These strategies are shown in 7 and 8 from Kukama. In 7, the theme participant and the noun 'shaman' each correspond to a single word, but the predication does not map to a specific word, though it is inherent in the construction. This poses a problem in annotating the "predicate" of the clause. In 8, the possessum and the predication correspond to the same word, that is, an "argument" is predicativized. Like participant indexation and noun incorporation, these types of constructions pose a problem for the annotation of predicate-argument structure. From a semantic perspective, it's important that the different strategies receive comparable annotations, since they have the same meaning.

[3] Due to space limitations, the English translation annotations for these examples are included in the supplementary material.

[4] The third strategy is the use of a verb separate from either participant, such as *have* in the English translation of 8, or the copula in the translation of 7.

These two different problematic strategies require different solutions. In the case of predicativized arguments as in 8, we use the same solution as for pronominal affixes and less-grammaticalized noun incorporation: both a nonverbal clause function and argument are identified and annotated separately. When there is no predicate, as in 7, then we assume that the annotator is able to recognize the type of nonverbal clause function, and use an abstract predicate in the annotation.

(7) ajan kunumi tsumi
this young.man shaman
'This young man is a shaman.'

(h / have-role-91
:ARG0 (k / kunumi 'young man')
:ARG1 (t / tsumi 'shaman')
:aspect State
:modstr Aff)

(8) Mijiri-tin ɨara-yara
Miguel-CER canoe-owner
'Miguel does have a canoe.' (Lit. 'Miguel is a canoe-owner')

(e / ɨara-yara 'has canoe'
:ARG0 (m / Mijiri 'Miguel')
:ARG1 (i / ɨara 'canoe')
:aspect State
:modstr Aff)

(h / have-03
:ARG0 (m / Miguel)
:ARG1 (c / canoe)
:aspect State
:modstr Aff)

Some of the nonverbal clause functions have specialized predicates in AMR, but not all; we propose additional predicates for those functions (see Table 1; ARG0 is always an argument, but ARG1 may be predicativized). The first four types in Table 1 describe possession and location. Possession and location may be predicated of the possession and the spatial figure, as in *This bicycle belongs to my brother* and *The bicycle is in the garage*. However, possession and location may be used in a context in which the information is presented as 'thetic' or 'all-new' in the terms of Lambrecht's (1994) theory of information structure (cf. the contrast between 'have' and 'belong' possession in Heine (1997)). One common thetic function is presentational, as in *I have one brother* or *In the garage was a single bicycle*. AMR has predicates for thetic possession (HAVE-03) and predicative location (HAVE-LOCATION-91); we add predicates for thetic location (EXIST-91) and predicative possession (BELONG-01).

The predication of properties (*Susan is smart*) and object categories (*Susan is a professor*) can be distinguished straightforwardly. AMR uses HAVE-MOD-91 for property predication and some types of object predication; we propose to restrict it to property predication. Other types of object predication are expressed in AMR with HAVE-REL-ROLE-91 or HAVE-ORG-ROLE-91; we propose a superordinate predicate HAVE-ROLE-91 that covers all object predication clauses. Finally, equational sentences (*He is the father of the bride*), corresponding to Lambrecht's identificational information structure, are challenging to distinguish from object predication in context (see Stassen (1997, 106-111)). Where this can be done, we propose to use the predicate IDENTITY-91.

Clause type	**Predicate**	**ARG0**	**ARG1**
thetic/presentational possession	have-03	*possessor*	*possession*
predicative possession	belong-01	*possession*	*possessor*
thetic/presentational location	exist-91	*location*	*theme*
predicative location	have-location-91	*theme*	*location*
property predication	have-mod-91	*theme*	*property*
object predication	have-role-91	*theme*	*object category*
equational	identity-91	*theme*	*equated referent*

Table 1: Nonverbal clause predicates

3 The road map

Section 2 covered solutions to typological issues that are raised by the inclusion of low- and no-resource languages in semantic annotation efforts. This section puts forth a "road map" approach to annotation, which synthesizes the typological solutions with practical solutions for the inclusion of languages with few existing computational or documentary resources.

	Stage 0	Stage 1
Annotation targets	treat derivational morphology as single word, separate inflectional morphology	indicate derivational morphological relations in lexicon
Aspect	coarse-grained categories on lattice	fine-grained categories on lattice
Modal strength	annotate with only MODSTR and placeholder values; no conceivers	use modal lexicon with modal strengths, fill in remaining unspecified modal strengths
Participant roles	general semantic roles	lexicalized roles, annotate implicit roles

Table 2: Road map annotation stages

The road map approach both ensures comparability across diverse languages, and allows for flexibility in the annotation of any one language. The road map specifies a starting point for languages with few resources (Stage 0), the end point for fully specified annotation (Stage 1), and a process for moving between these, defined for each annotation category. These are not discrete annotation stages, and languages will move gradually from the Stage 0 to Stage 1 annotation. Where a language begins on the road map for each annotation category is determined by the typological features of its grammar, its state of documentation, and the computational resources developed thus far.

The road map allows for flexibility across languages and annotation categories. Languages with a paucity of linguistic or computational resources can still begin annotation efforts. Languages with typological features that complicate the annotation of certain semantic categories can still be annotated for those categories, albeit at a less detailed level. Within a language, different annotation categories may be annotated at different stages, depending on the language's typological features and existing resources.

The road map approach also ensures comparability across languages, even when languages are at different stages, because annotation values retain their meaning across the road map stages. This also ensures that different-stage annotations for the same language are compatible. As annotation and documentation efforts continue, the annotation of a language may progress along the road map. But, the annotations done at the beginning stages are still accessible and comparable to the later stage annotations. For languages that have limited resources in terms of time investment by speakers and/or field linguists, having this type of compatibility built into the annotation scheme is critical.

The remainder of this section will demonstrate how the road map approach functions with regard to a number of annotation categories: annotation targets, participant roles, aspect, and modal strength and polarity. The road map for these categories is summarized in Table 2.

3.1 Annotation targets

The main cross-linguistic issues with the identification of annotation targets (i.e., predicates and arguments) are the types of multiconcept words covered in §2. The annotation of multiconcept words is the same throughout the stages of the road map; however, their representation in the lexicon builds up in complexity. For example, verbs with derivational morphology are first treated as different words than their non-derived counterparts in the lexicon. As the understanding of the language progresses, multiconcept words are morphologically decomposed and morphological relations are added to the lexicon.

The identification of a span of text for each annotation target is determined by the language experts for each language, since what is considered the 'citation form' of a word differs across languages. Fusional morphology, such as that for pronominal indexation in Arapaho (see 3, 5, 6 above), cannot be split apart at any stage of the road map and therefore a span of text is not indicated for those arguments.

3.2 Aspect

Aside from the multiconcept word issues with regard to aspectual morphology discussed above, the main issue with aspect annotations cross-linguistically is that languages differ widely in terms of which aspectual distinctions are conventionalized in their grammar. In order to resolve these differences, we utilize the aspectual lattice from Van Gysel et al. (2019), shown in the supplementary material. It ranges from the most coarse-grained categories of IMPERFECTIVE and PROGRESSIVE, to ATELIC PROCESS and PERFECTIVE, to the 'basic' level of STATE, ACTIVITY, ENDEAVOR, and PERFORMANCE, and finally, very fine-grained categories, such as POINT STATE or DIRECTED IRREVERSIBLE ACHIEVEMENT.

For a language at an earlier stage of linguistic analysis, it may not be clear to the annotator which of the more fine-grained aspect values should apply. Therefore, annotators may select a more coarse-grained category on the lattice. For example, the linguistic analysis of aspect in Sanapaná, in 9, is still under way. The aspectual implications of the suffixal morphology (specifically, the passive *-akp* which also functions as a reciprocal, and the subjunctive *-o*), are not yet fully understood. Therefore, the more coarse-grained ATELIC PROCESS value is used, instead of an ACTIVITY or ENDEAVOR value.

(9) tenyo
then
apk-ehl-pa'met-kes-**akp-o**=hla
2/3M-DISTR-talk-APPL-**PAS.M-SUBJ**=INF
ap-yavokhoho.
2/3M-all
'Then they were all talking to each other.'

(p / ehlpa'metkesamma'ap 'speak to each other'
:Actor (a / apyavokhoho 'all')
:Recipient (a / apyavokhoho 'all')
:aspect Atelic Process
:modstr Aff)

Stage 2 of the aspect annotation uses the more fine-grained categories on the aspect lattice. Example 5 above from Arapaho expresses an event that is aspectually similar to 9. Since Arapaho has a longer history of linguistic study, the more fine-grained annotation of ENDEAVOR can be applied.

3.3 Modal strength and polarity

We follow Vigus et al. (2019) in representing modal strength and polarity as a dependency structure. The nodes are events or conceivers (i.e., a source, an entity whose perspective on an event is modeled in the text). The edges in the dependency structure correspond to epistemic strength and polarity values; event nodes are the children of either conceivers or other events on whom they depend for their modal value.

Like aspect, languages differ in the modal strength distinctions that are conventionalized in their grammar and therefore we use a typological lattice, shown in the supplementary materials. This lattice is based around a FULL vs. PARTIAL vs. NEUTRAL modal strength distinction; the coarse-grained categories are NON-FULL and NON-NEUTRAL; the finer-grained categories include WEAK PARTIAL, STRONG NEUTRAL, etc. These combine with an AFFIRMATIVE/NEGATIVE polarity distinction.[5]

The Stage 0 annotation involves the underspecification of some parts of the modal dependency structure. Events are annotated for their modal strength (MODSTR) using the lattice, but conceivers are unspecified. Some event types receive special annotations; two of these are events under the scope of a modal predicate, and events under the scope of a reporting/speech predicate.[6] A placeholder MODAL value is used for modal predicates; and a QUOT value is used for reporting predicates. Events under the scope of modals don't receive a MODSTR value; reported events do receive a MODSTR value in the same way as other predicates. This way, events under the scope of other predicates in the modal dependency receive a consistent annotation, while annotators avoid the complexity of annotating the full dependency structure. This annotation for modal predicates is shown above in the English translation of 3.

The MODAL and QUOT values can be automatically converted into an underspecified dependency structure; the participant role annotation can also be leveraged to specify conceivers (e.g., the EXPERIENCER of a modal predicate is its conceiver). The modal strength imparted by modal predicates is unspecified in the dependency structure at Stage 0. Stage 1 involves adding this information to the lexicon entries for modal predicates (e.g., *want* imparts a NEUT strength on its complement) and filling in other unspecified values to reach a fully specified modal dependency structure.

3.4 Participant roles

Semantic role annotation is one category where issues related to typological differences and resource disparities intersect. As has been noted in the verbal semantic literature (Croft, 2012; Hartmann et al., 2014), semantic roles, such as AGENT or PATIENT are difficult to apply consistently across languages.[7]

[5] In this paper, we use the default level annotations to yield six modal strength values: full affirmative AFF, partial affirmative PARTAFF, neutral affirmative NEUTAFF, neutral negative NEUTNEG, partial negative PARTNEG, and full negative NEG.

[6] Conditionals and purpose clauses also receive special placeholder annotation values, COND and PURP respectively.

[7] For example, transfer constructions can realize either the giver as subject (*I gave the cat some wet food*), or the recipient as subject (*the cat received her wet food*). This varies both within and across languages, making it unclear which participant should receive the AGENT semantic role.

Central roles	Actor, Undergoer, Theme, Recipient, Force, Causer, Experiencer, Stimulus
Peripheral roles	Instrument, Companion, Material/Source, Place, Start, Goal, Affectee
Roles for entities and events	Cause, Manner, Reason, Purpose, Temporal, Extent

Table 3: UMR non-lexicalized roles

Therefore, both typological research (Hartmann, 2013; Malchukov and Comrie, 2015) and semantic annotation (e.g., PropBank) have moved away from general semantic roles and towards microroles, or lexicalized semantic roles. Roles are defined for each verb (e.g., *eat* has an EATER and FOOD); this allows for valid cross-linguistic comparison in typology and consistency in semantic annotation.

The major drawback of this approach is that it requires an existing lexicon complete with lexicalized roles for the verbs in a language. For languages that do not have this, the creation of such a resource is a rather large hurdle to overcome in order to begin annotation. Therefore, the road map moves from more general semantic roles at Stage 0 to lexicalized microroles at Stage 1. For languages that have PropBank-style frame files created, annotation can begin at Stage 1. For languages that do not, annotation begins with general semantic roles at Stage 0, while simultaneously building up a lexicon of frame files.

Stage 0 of the road map involves selecting a label for each participant from a set of general (i.e., non-lexicalized) semantic roles, shown in Table 3. This inventory is largely an extension of the AMR inventory of non-core roles, with roles added for core arguments such as STIMULUS. These additions are based upon the cross-linguistic argument realization patterns in the ValPaL database (Hartmann, 2013); this ensures that the labels reflect distinctions that are common in the grammatical systems of the world's languages. At Stage 0, implicit (i.e., unexpressed) participants are not annotated; this is shown in the Arapaho example in 3 above, where the goal participant is not annotated, as it is not overly expressed.

In order for a language to progress along the road map with regard to participant roles, Stage 0 also involves beginning to set up a lexicon with frame files. Within each frame file, the mapping between a lexicalized semantic role and its non-lexicalized counterpart is indicated. This way, annotations at different stages of the road map will be comparable with each other. As frame files are created, annotators use the lexicalized roles for verbs that have them; for other verbs, the general semantic roles are used. At Stage 1, the lexicalized roles are used; this is shown below for example 3 in §2.1. In Arapaho, the existing lexical description with argument structure information can be leveraged in annotation to create frame files like the one shown below. Stage 2 also involves the annotation of implicit roles, based on the frame files; therefore, the goal (ARG2) participant for example 3 is annotated.

predicate: BEETWON3EIIN
arguments:
ARG0: putter → ACTOR
ARG1: put thing → THEME
ARG2: putting goal → GOAL

(n / beetwon3eiin 'want to go and put s.t. inside a place'
:ARG0 (a / '3S')
:ARG1 (t / '2S')
:ARG2 (g / 'jail')
:aspect Habitual
:modstr Neut)

4 Conclusion

This paper recognizes issues not previously dealt with in the annotation of cross-linguistic semantic information: multiconcept words and no-resource languages. As multiconcept words are more common in languages with a smaller speech community, they have not been dealt with in past annotation schemes. We present solutions for extending AMR across languages, including the annotation of multiconcept words; these depend on the semantic category of the concept. We have also outlined a road map approach to beginning annotation on very low or no-resource languages, ensuring that the annotation is truly cross-linguistic in terms not only of typological diversity but of resource availability as well.

5 Credits

We gratefully acknowledge the support of the National Science Foundation Award Nos. 1764091 to the University of New Mexico and 1764048 to the University of Colorado (Collaborative Research: Building a Uniform Meaning Representation for Natural Language Processing).

References

Lasha Abzianidze, Rik van Noord, Hessel Haagsma, and Johan Bos. 2020. The first shared task on discourse representation structure parsing. *arXiv preprint arXiv:2005.13399*.

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, August. Association for Computational Linguistics.

Balthasar Bickel and Johanna Nichols. 2013. Inflectional synthesis of the verb. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Dwight Bolinger. 1963. The uniqueness of the word. *Lingua*, 12(2):113–136.

Johan Bos, Valerio Basile, Kilian Evang, Noortje J Venhuizen, and Johannes Bjerva. 2017. The groningen meaning bank. In *Handbook of linguistic annotation*, pages 463–496. Springer.

Greville G. Corbett. 2000. *Number*. Cambridge University Press.

Andrew Cowell and Alonzo Moss Sr. 2008. *The Arapaho language*. University Press of Colorado.

William Croft. 2012. *Verbs: Aspect and causal structure*. Oxford University Press, Oxford.

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and D. Manning, Christopher. 2014. Universal stanford dependencies: A cross-linguistic typology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4585–4592. European Language Resources Association (ELRA).

Robert MW Dixon, Alexandra Y Aikhenvald, et al., 2002. *Word: A cross-linguistic typology*, chapter Word: a typological framework, pages 1–41.

David M Eberhard, Gary F. Simons, and Charles D. Fennig. 2020. Ethnologue: Languages of the world. twenty-third edition. http://www.ethnologue.com.

Antonio Almir Silva Gomes. 2013. *Sanapaná uma lingua maskoy: Aspectos gramaticais*. Ph.D. thesis, Universidade Estadual de Campinas.

Iren Hartmann, Martin Haspelmath, and Michael Cysouw. 2014. Identifying semantic role clusters and alignment types via microrole coexpression tendencies. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"*, 38(3):463–484.

Martin Taylor Bradley (eds.) Hartmann, Iren Haspelmath, editor. 2013. *Valency Patterns Leipzig*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Bernd Heine. 1997. *Possession: Cognitive sources, forces, and grammaticalization.* Cambridge Studies in Linguistics. 83. Cambridge University Press, Cambridge.

Hans Kamp and Uwe Reyle. 2013. *From discourse to logic: Introduction to modeltheoretic semantics of natural language, formal logic and discourse representation theory*, volume 42. Springer Science & Business Media.

Knud Lambrecht. 1994. *Information structure and sentence form: Topic, focus, and the mental representations of discourse referents.* Cambridge studies in linguistics: 71. Cambridge University Press.

Gary Lupyan and Rick Dale. 2010. Language structure is partly determined by social structure. *PloS one*, 5(1):e8559.

Andrej Malchukov and Bernard Comrie. 2015. *Valency classes in the world's languages*. Walter de Gruyter, Berlin/Boston.

Jonathan May and Jay Priyadarshi. 2017. Semeval-2017 task 9: Abstract meaning representation parsing and generation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 536–545.

Marianne Mithun. 1984. The evolution of noun incorporation. *Language*, 60:847–94.

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

Leon Stassen. 1997. *Intransitive predication*. Clarendon Press, Oxford.

Leon Stassen. 2009. *Predicative Possession*. Oxford University Press, Oxford, UK.

Rosa Vallejos and Rosa Amías. 2015. Diccionario kukama-kukamiria castellano. *Iquitos: AIDESEP: ISEPL: FORMABIAP*.

Rosa Vallejos. 2014. The kukama-kukamiria documentation project. https://elar.soas.ac.uk/Collection/MPI971108 (accessed: 4 August 2020).

Rosa Vallejos. 2016. *A Grammar of Kukama-Kukamiria: A language from the Amazon*. Brill.

Jens E. L. Van Gysel, Meagan Vigus, Pavlina Kalm, Sook-kyung Lee, Michael Regan, and William Croft. 2019. Cross-linguistic semantic annotation: Reconciling the language-specific and the universal. In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 1–14, Florence, Italy, August. Association for Computational Linguistics.

Jens E. L. Van Gysel. 2017. Temporal predicative particles in Sanapaná and the Enlhet-Enenlhet language family (Paraguay): A descriptive and comparative study. MA Thesis, Universiteit Leiden.

Jens E. L. Van Gysel. 2020. A documentation of historical narratives amongst the Sanapaná (Enlhet-Enenlhet) of the Paraguayan Chaco. https://elar.soas.ac.uk/Collection/MPI1234837 (accessed: 31 October 2020).

Meagan Vigus, Jens E. L. Van Gysel, and William Croft. 2019. A dependency structure annotation for modality. In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 182–198.

Tim Zingler. 2020. *Wordhood issues: typology and grammaticalization*. Ph.D. thesis, University of New Mexico.