Theoretical and Practical Issues in the Semantic Annotation of Four Indigenous Languages

Jens E. L. Van Gysel [1], Meagan Vigus[1], Lukas Denk [1], Andrew Cowell[2], Rosa Vallejos[1], Tim O'Gorman[3], and William Croft[1]
[1]Department of Linguistics, University of New Mexico
[2]Department of Linguistics, University of Colorado, Boulder
[3]Thorn
{jelvangysel,mvigus, ldenk,rvallejos,wcroft}@unm.edu,
tim.ogorman@wearethorn.org, james.cowell@colorado.edu

Abstract

Computational resources such as semantically annotated corpora can play an important role in enabling speakers of indigenous minority languages to participate in government, education, and other domains of public life in their own language. However, many languages – mainly those with small native speaker populations and without written traditions – have little to no digital support. One hurdle in creating such resources is that for many languages, few speakers would be capable of annotating texts – a task which requires literacy and some linguistic training – and that these experts' time is typically in high demand for language planning work. This paper assesses whether typologically trained non-speakers of an indigenous language can feasibly perform semantic annotation using Uniform Meaning Representations, thus allowing for the creation of computational materials without putting further strain on community resources.[1]

1 Introduction

Over the last few decades, there has been a call to enable speakers of indigenous minority languages to participate in government, education, and other domains of public life in their own language. Computational resources can play an important role in such efforts (Probst et al., 2002). For example, semantically annotated corpora for minority languages can be used for information extraction to obtain situational awareness in disaster situations (Griffitt et al., 2018), to link unstructured text in various languages to structured knowledge bases (Zhang and Rettinger, 2014), and as scaffolding for machine translation into these languages . As of 2019, only 1705 out of the 7795 languages in Simons and Thomas (2019), 22%, had any digital support. Even for languages with large native speaker populations and considerable political standing such as Farsi, computational resources are often limited (Feely et al., 2014).

The limited availability of (digital) data in such minority languages is only one hurdle to the creation of computational resources. Wherever data are available, they need to be provided with semantic annotations in order to be made maximally useful for the purposes described above.

Semantic annotation allows unstructured text to be linked to representations such as Abstract Meaning Representations (AMR, Banarescu et al., 2013) or Discourse Representation Structures (DRS, Kamp and Reyle, 2013; Bos et al., 2017). Such annotation schemes have become more cross-linguistically informed over the years. The DARPA Low Resource Languages for Emerging Incidents project (LORELEI), for one, has conducted shared annotation tasks with languages such as Tagalog, Yoruba and Somali (Griffitt et al., 2018). The Uniform Meaning Representation project, on the other hand, aims to make English-based AMR cross-linguistically applicable (Van Gysel et al., 2021).

In practice, however, current annotation workflows have little chance of being applied to truly "no-resource" languages. Semantic annotation is typically done by speakers of the target language, as it is assumed that (native) speaker intuitions are necessary to make judgments required for semantic annotation. This may be feasible for "low-resource" languages with millions of speakers such as Oromo, Tigrinya, Uyghur, and Ukrainian – the "incident languages" in Griffitt et al. (2018). For many others, including most of the 1500 languages with fewer than 1000 speakers (Eberhard et al., 2020), such annotators are unlikely to be available for several reasons (see section 2). This paper therefore has two main goals. Firstly, it assesses whether the structure of UMR indeed makes it scalable to languages with a different typological profile than traditionally well-represented languages in NLP such as English and Mandarin. Secondly, it assesses whether

[1]The sixth author was affiliated with the University of Massachusetts Amherst at the time this work was done.

Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop, pages 12–22

non-speakers of an indigenous language trained in typological linguistics can successfully perform UMR semantic annotation of such languages based on morpheme-level glosses, utterance-level free translations, grammars, and dictionaries.

Specifically, we present quantitative results of two annotation experiments using UMR to annotate texts in Kukama (Tupían, Peru), and Arapaho (Algonquian, US), and qualitative results of initial annotation efforts with Sanapaná (Enlhet-Enenlhet, Paraguay) and Navajo (Athabaskan, US). These four languages were chosen because (1) they represent a range of resource availability from no-resource (Sanapaná) to low-resource (Arapaho, see Sections 2.1-2.4), (2) they are typologically diverse, representing more isolating (Kukama), agglutinating (Sanapaná), and polysynthetic (Arapaho, Navajo) types, and (3) co-authors of this paper have significant expertise in them. In section 2, the advantages of a workflow using linguistically trained non-speakers as annotators are laid out, and its necessity is illustrated through sociolinguistic sketches of the four languages at hand. Section 3 introduces UMR. Section 4 presents an overview of theoretical issues relating to the UMR guidelines encountered during the annotation of these four languages. Sections 5-6 present the inter-annotator agreement and adjudication results of the Kukama and Arapaho annotation experiments. Section 7 presents an overview of difficulties with the annotation workflow used in these two experiments.

2 Limitations on current semantic annotation procedures

Regardless of how large a corpus is available and how typologically informed an annotation scheme is, trained and qualified annotators are vital to successful annotation. These are often hard to come by in low-resource language contexts. Linguists working on such languages have the training required to familiarize themselves with semantic values and concepts used by annotation schemes. Since they are often still in the process of studying the language, however, their semantic judgments may be less than reliable. Native speakers can provide the most accurate interpretations of the meanings of forms in their language. From this point of view, they make "better" annotators. However, they vary with respect to various "literacies".

According to Eberhard et al. (2020), 3982 of the world's languages – slightly over half – have a writing system. However, only 15-20% of languages have an actual written tradition (Borin, 2009). Therefore, native speaker consultants often have limited literacy in their own language, although they may have literacy in regional or national linguae francae. Many low-resource languages are spoken in remote areas where people may not yet have easy access to or familiarity with digital technology. They are also likely to have more limited access to formal linguistics training than the national societies surrounding them.

In contexts of advanced language shift (which many indigenous language-speaking communities face), these factors may form correlated continua, and may be related to age. There may be a continuum of multilingualism so that younger speakers are more likely to be dominant in a lingua franca, while older generations are near-monolingual in the local language. Similarly, the speakers most fluent in an endangered language are likely to have gone through less formal education, be less literate, and be less comfortable with digital tools.

These factors conspire to create a situation where few speakers of low-resource indigenous languages are currently well-equipped to conduct semantic annotation projects on their language. Furthermore, wherever there are qualified local language experts, their time is typically in high demand for language planning projects such as curriculum development and teaching, which are typically given higher priority in language maintenance efforts. The following brief sociolinguistic sketches of Kukama, Arapaho, Sanapaná, and Navajo illustrate these points.

2.1 Kukama (Tupian, Peru)

Among the Kukama, most of the around 1,000 native speakers are over 60, and live scattered across many remote villages, limiting their opportunities to use the language with each other. Many have not received formal education in either Kukama or Spanish. Adults in their 40s and 50s may have passive knowledge of Kukama. Through revitalization projects, young people have started learning Kukama as a heritage language and training as bilingual teachers. Several orthographies have been proposed, but the implementation and consolidation of an official one is still ongoing. It is unlikely that the heritage speakers have sufficient command of the language to perform semantic annotation by themselves, or that the elderly native speakers can be sufficiently trained in linguistics to do so.

2.2 Arapaho (Algonquian, USA)

Among the Northern Arapaho (Wyoming), there are about 100 native speakers, all over 65, only a handful of whom have strong literacy skills in Arapaho. Several hundred have native passive knowledge, and several hundred second language learners exist, though only a handful of younger people have acquired significant (but non-fluent) productive knowledge. Some second language learners with good literacy and technology skills could do annotation using the existing semantically labeled and translated corpus, but may need to collaborate with native speakers to capture all nuances.

This context may be better described as low-resource than no-resource, and is one of the few indigenous language contexts where annotation might happen without direct participation of a linguist. A linguist would likely still have to train annotators. Even here, the very small numbers of literate native speakers and reasonably productive second-language speakers poses issues: there are high demands on the time of both groups to engage in basic teaching and curriculum development.

2.3 Sanapaná (Enlhet-Enenlhet, Paraguay)

Sanapaná is spoken by around 1000 people in Paraguay. There is still intergenerational transmission in two communities. In the other Sanapaná communities, most people have shifted to Paraguayan Guaraní or Spanish for daily communication. Even where children still learn Sanapaná in the home, they go through formal education in Spanish and Guaraní, and tend to be more literate in these national languages than in Sanapaná. Young people often have smartphones, but few people have experience working with computers. No native speakers have formal training in linguistics, so it is unlikely that even most of the younger generation of native speakers would be able to perform semantic annotation independently.

2.4 Navajo (Athabaskan, USA)

With around 170.000 speakers, Navajo is the least endangered Native North American language. Although younger generations are becoming less fluent, their language and culture are part of curricula in high schools and colleges across the reservation (e.g. Diné College; Navajo Technical University), and various pedagogical and linguistic materials are available to promote research on the language. Furthermore, the University of New Mexico offers a graduate degree in Navajo Linguistics and employs native speakers as professors and instructors. Navajo scholars from this environment would be able to do semantic annotation of their language.

2.5 Takeaways

For communities whose language is threatened, the role of dictionaries, teaching materials, and text collections in language development projects may seem more obvious than that of semantic annotation. Therefore, the few speakers who would be qualified for annotation (multilingual, literate, linguistically trained native speakers; field linguists with some degree of proficiency in the language), as well as speech communities as a whole, may prefer to spend their limited time and resources in ways that contribute more directly to these tangible goals. An annotation workflow where non-speakers of the language (e.g. research assistants who are unable to spend time in the field) can do semantic annotation would facilitate the creation of computational resources without drawing on the limited and valuable time of local experts.

3 Uniform Meaning Representation (UMR)

Uniform Meaning Representation (UMR) is a semantic annotation scheme designed to meet the needs of a wide range of NLP applications which require intermediate meaning representations. UMR is based on Abstract Meaning Representation (Banarescu et al., 2013). AMR captures the meanings of natural language sentences as single-rooted, directed, node- and edge-labelled graphs, and focuses on predicate-argument structure, named entities, and word sense disambiguation. It was mainly designed for the annotation of English. The usefulness of AMR and its amenability to machine learning have been repeatedly proven (Cai and Lam, 2020; Li et al., 2020).

3.1 Goals of UMR annotation

Potential downstream uses of UMR include applications of interest to the NLP community, and applications of use to speech communities. One the one hand, the use of AMR as an intermediate representation for e.g. inferencing for question answering (Sachan and Xing, 2016; Kapanipathi et al., 2021) and human-robot interaction (Bastianelli et al., 2013) has been proven. So far, inferencing from AMRs has necessarily remained

limited to sentence-level information, or has been based on ad-hoc meaning representations spanning multiple sentence-level AMRs. UMR's document-level structure (see Section 3.2) is expected to aid such applications. Supervised machine learning systems for morphological analysis can be trained on small datasets (as small as 3000 annotated word tokens, Moeller and Hulden, 2018; Moeller, 2021) collected during language documentation. Application of such methodologies to semantic annotation could allow for the expansion of such downstream applications to low-resource language contexts.

On the other hand, UMR-annotated corpora would be useful to more typical language documentation endeavors. UMR is designed as a "road map" to help research teams develop PropBank-style frame files during the process of annotation. The annotation process may thus help enrich lexicons with argument structure information (Croft and Sutton, 2017). The inclusion of such information in lexicons can facilitate the learning of the language by new speakers, especially in contexts of advanced language endangerment where most learners are adults and have limited opportunities to acquire usage patterns through immersion.

3.2 Structure of UMR

UMR nodes represent concepts (e.g. word senses, named entity types, attribute values), while edges between nodes represent relations between those concepts (e.g. participant roles, attribute types such as aspect, person, and number, and general semantic relations). Sample UMRs for Kukama sentences can be found in the supplementary materials.

UMR builds on AMR by extending it to new semantic domains, such as aspect, coreference, and temporal and modal dependencies (Van Gysel et al., 2021). Most semantic domains are captured in a sentence-level graph structure, while co-reference, temporal relations, and modality are captured in a document-level graph. This paper deals only with the sentence-level structure: we annotated (and calculated IAA for) predicate-argument structure, aspect, and other non-participant role relations within the phrase and the clause (e.g. possession, quantification). UMR specifications for these semantic domains can be found in Van Gysel et al. (2021).[2] Temporal and modal relations were not annotated, nor was cross-sentence co-reference. UMR also aims to extend AMR to new languages by basing its structure soundly on insights of linguistic typology (Van Gysel et al., 2019), and by making its annotation workflows amenable to low-resource contexts often encountered in work with the world's endangered languages (Vigus et al., 2020).

One way in which UMR was made flexible for low-resource languages is a roadmap approach (Van Gysel et al., 2021). Languages without existing computational resources start at "Stage 0". Annotation here operates at the word level, so that annotators do not have to morphologically decompose words - they do, however, have to consider the semantics of inflectional morphology. Since different semantic domains annotated by UMR are somewhat independent, annotators for "Stage 0" languages may choose to only annotate categories for which they are confident in their analysis - likely categories expressed by independent words. They may, for instance, choose to annotate argument structure and tense, but not aspect. Annotators are more likely to use coarse-grained categories on annotation lattices (Van Gysel et al., 2019). "Stage 1" of UMR annotation is based on the existence of a lexicon with PropBank-style frame files and advanced semantic analysis of the language. There is somewhat of a continuum between the two stages: as semantic analysis of the language advances, more categories will be able to be annotated, and they will be annotated at a more fine-grained level, moving the language closer to Stage 1. For the annotation task described in this paper, annotators used Stage 0 UMR as the languages did not have existing lexicons with frame files.

4 UMR Applied to Four Indigenous Languages

The UMR annotation scheme and its guidelines were designed with cross-lingual variation in mind. We tested UMR on narrative texts in two indigenous languages, Navajo and Sanapaná. 107 lines of Sanapaná text were annotated by the first author, amounting to 332 Sanapaná words and 600 words in the free Spanish translation. 261 lines of Navajo text were annotated by the third author, amounting to 2044 Navajo words and 5020 words in the free English translation. In addition, issues that arose in the Kukama and Arapaho experiments (Sections 5-6) also tested the UMR annotation scheme. UMR adds semantic structure to AMR annotation, specifically aspect, general participant roles, quantifi-

[2]The current UMR guidelines can be found at https://umr4nlp.github.io/web.

cation, scope, and temporal and modal structure (including polarity). We focus on the annotation of those categories here.

4.1 Extensions to Current UMR

Many semantic categories remain unannotated in the current version of UMR (or standard AMR). The annotators of the four indigenous languages encountered some of these categories, even in the short texts they annotated. These categories are targets for future development of UMR.

These categories pertain to discourse-level phenomena, including interactional categories. UMR includes coreference relations to entities and events. However, it does not annotate the relative discourse prominence of referents. Arapaho partly encodes discourse prominence through a grammatical distinction between proximate and obviative noun phrases. UMR does not currently annotate coreference between a pronoun and a section of prior discourse (also known as "discourse deixis") either, as with the Sanapaná anaphoric construction in (1), where *apkeleyvoma enyatav'a* 'how our ancestors lived/the lives of our ancestors' is equated to a whole paragraph of preceding discourse.

(1) *ahlta=n-t-em-ak*
PHOD=2/3F-COP-PST/HAB-V2.NFUT
apk-el-eyv-om-a
2/3M-DSTR-live-PST/HAB-NMLZ
en-yata-v'a
1PL-grandfather-PL
'That is how our ancestors lived.'

Another category not currently annotated is clause-level information structure, specifically the distinction between topic-comment, thetic and identificational (focus) sentences (Lambrecht, 1994). The last sentence type includes focus operators such as 'just' and 'only' as in (2) from Sanapaná.

(2)
apk-el-v-ay'-aye=hlta
2/3M-DSTR-arrive-PST/HAB-V1.NFUT=PHOD
vanhla' valayo sokhoye'
only Paraguayan at.first
'Only the Paraguayans arrived at first.'

Finally, interactional grammatical constructions such as speech act constructions and vocatives occurred in reported speech in the narratives.

UMR has adopted AMR's large set of named entity categories. However, the set needs to be extended to include named entities not found in large industrialized societies. In the annotated texts, named entities were found referring to clans (Navajo), age-grade societies (Arapaho) and supernatural beings (Kukama).

4.2 The Sentence-Level to Document-Level Annotation Pipeline

Issues in the application of UMR to specific constructions in the four languages raised more general issues in semantic annotation across languages.

The first issue is the relation between sentence-level and document-level annotation (across sentences). The standard pipeline is to annotate individual sentences before document-level annotation. Aspect is annotated at the sentence level, while modal dependencies are annotated at the document level, as hypothetical "worlds" or mental spaces (Fauconnier, 1985) containing unrealized events, that may be referred to across sentences. For unrealized events, such as the Arapaho imperatives in (3), we annotate aspect at the sentence level as the event would be realized in the mental space of the speaker's command.

(3) *wohei cei-te'e be! cei-koohu!*
okay to.here-this.side friend to.here-run
'Wohei come this way, friend! Run this way!'

However, the annotation of the mental space as the unrealized space of the speaker's command is done at the document level. Following this standard pipeline means that document-level context is not considered during sentence-level annotation, which is somewhat counterintuitive to linguist annotators.

4.3 Semantic Annotation and Lexical Semantic Differences

Another broad issue arose with the annotation of general participant roles, which are used for arguments of predicates before frame files are developed for a language. This allows for the annotation of basic clause structure at the outset, since developing frame files for a lexicon is a long-term task. The general participant roles are defined in UMR by broad semantic event classes: mental events have Experiencer and Stimulus roles, motion/location events have Theme and Location/Goal/Source roles, change of state events have Actor and Undergoer roles, and transfer events have a Recipient role as well as Agent and Theme roles. But some roles vary within and across languages with respect to the agentivity of participants.

For example, the subject of ‘stay’ is annotated as Actor in UMR since the subject of English *stay* is more active/agentive than *be (at)*. Kukama *yuti* is translated as‘stay’, but is more stative, so Actor is not an appropriate label. Conversely, the subject of ‘know’ is annotated as Experiencer in UMR, because the subject of English *know* is not agentive. The Arapaho verb *he'in-* is typically loosely glossed ‘know’; however, it can also mean ‘find out about’ and ‘remember, keep in mind’, and it can be used in imperatives as ‘know/remember it!’. This verb could be better glossed as ‘actively remember/understand and act thereon’ and involves agency on the part of the subject. In contrast, Arapaho *hee3obee-* means ‘know’ or ‘feel’ in the sense of information acquired via sensory input of any kind, and the subject is not agentive.

The first issue here is that participant roles are event-specific, and fixing a general participant role for events such as ‘stay’ or ‘know’ is arbitrary to some extent. But the cross-lingual issue is that events are categorized or conceptualized in different ways in different languages: Kukama *yuti* is construed as less active than English *stay*, and Arapaho has two different verbs covering some of the same set of events as English *know*, but not precisely. Semantic annotation of certain grammatical categories, including participant roles and aspect, is closely tied to lexical semantics. We cannot assume that verbs in other languages are semantically identical to their English translation equivalents. Yet research in lexical semantic typology is much less far advanced than in the semantic typology of grammatical categories (Koptjevskaja-Tamm and Vanhove, 2012).

A solution is the development of verb-specific frame files, with enough semantic detail to capture differences as well as similarities in meaning across languages. As these verb-specific frame files are developed, every verb-specific participant role will be linked to the relevant general participant role previously used in this language, so as to not render “Stage 0” participant role annotations obsolete. As noted above, that is a major task, and in the meantime, general participant role annotation must be taken with a large grain of salt.

4.4 Preserving Language-Particular Semantic Subtleties

Finally, a general observation of the language experts that adjudicated the annotation experiments described in sections 5 - 6 is a concern that cross-lingual semantic annotation may be too coarse-grained to capture important semantic subtleties encoded in particular languages. For example, in Arapaho, *heyeih-* is a prefix meaning ‘almost [to a location], almost [finished], etc.’ with the action still ongoing and the assumption that it will be or at least could be completed successfully. In contrast, *too-* is a prefix meaning ’almost [did it, but not quite], ‘almost [but just missed], etc.’ with the action now ended but unsuccessful. English *almost* does not distinguish these senses, and a semantic representation based on *almost* does not capture the meaning of each Arapaho translation equivalent. We suggest extending the idea of lattices of semantic values proposed by Van Gysel et al. (2019) to semantic categories such as that expressed by *almost*, *heyeih-* and *too-*, to capture language-particular semantic subtleties while retaining cross-lingual comparability.

5 Annotation Experiment 1: Kukama

5.1 Annotation procedure and materials

60 lines of Kukama text were taken from a traditional narrative collected by the fifth author of this paper, a linguist with extensive knowledge of Kukama, but little UMR experience. This dataset amounted to 223 words of Kukama text, and 360 words in the English free translation. These data were annotated according to the current UMR guidelines (Van Gysel et al., 2021) by the first and second author of this paper, both of whom have extensive experience with UMR and training in typological and/or descriptive linguistics, but little to no previous knowledge of Kukama.

The first ten lines were annotated by both annotators separately as a trial, after which they convened and adjudicated their disagreements. This adjudicated version was then discussed with the expert linguist. The following 50 lines were then annotated by both annotators independently, and constitute the sample over which inter-annotator agreement was calculated. Annotators took an average of 5.8 and 6.7 minutes per line for these 50 lines, respectively. Annotators had at their disposal a transcription of the Kukama text, a morpheme breakdown, morpheme-level glosses in English and Spanish, and intonation unit-level free translations in English and Spanish. They could also draw on a Kukama grammar (Vallejos, 2016), and a dictionary (Vallejos and Amías, 2015).

5.2 Evaluation procedure

Inter-annotator agreement (precision, recall, and F-score) between the two annotators was calculated using Smatch (Cai and Knight, 2013). Two sets of agreement scores are presented in section 5.3: the first one represents the exact agreement between the sets of annotations, while the second one represents the "compatibility" of the annotations on the typological lattices in which the UMR annotation categories are organized; see Van Gysel et al. (2019). If one annotator chooses a fine-grained value (e.g. Performance, for aspect) that is a subvalue of the value chosen by the other annotator (e.g. Process), this triple counts as a disagreement for the set of identity scores, but as an agreement for the set of compatibility scores.

In addition to IAA scores, we present accuracy scores for both annotators. A gold standard annotation was constructed by discussing the annotation of each of the 50 lines of data with the language expert. First, the two annotators established an adjudicated graph for each line between themselves, which was subsequently adjusted as deemed necessary by the language expert. Both annotators' original graphs were then compared with this established gold standard using Smatch. Once again, both identity and compatibility scores are reported.

5.3 Annotation results

The IAA scores of the Kukama annotation experiment can be seen in the first two lines of Table 1. IAA is rather high, approaching the 0.8 threshold for precision, recall, and F-score. Treating compatible-but-not-identical annotations as agreement increases precision by 0.01, but does not affect recall and F-score. This result can be interpreted as indicating that both annotators were able to apply the UMR annotation guidelines to Kukama in rather systematic and consistent ways, and were able to use annotation categories at the same level of granularity, even though neither of them had significant previous knowledge of the morphosyntactic or morphosemantic characteristics of Kukama.

Lines 3-6 show both annotators' agreement with the gold standard established after adjudication with the language expert. In both cases, precision surpasses the threshold of 0.8, while recall approaches it, and F-score surpasses it for one annotator. The compatibility scores show only a marginal increase in F-score, and only for one annotator.

	Precision	Recall	F-score
Ann. 1 - Ann. 2 Identity	0.77	0.79	0.78
Ann. 1 - Ann. 2 Compatibility	0.78	0.79	0.78
Ann. 1 - Gold Identity	0.81	0.76	0.78
Ann. 1 - Gold Compatibility	0.81	0.76	0.78
Ann. 2 - Gold Identity	0.85	0.78	0.81
Ann. 2 - Gold Compatibility	0.85	0.78	0.82

Table 1: Kukama IAA and Accuracy scores

6 Annotation Experiment 2: Arapaho

6.1 Annotation procedure and materials

35 lines of Arapaho text were taken from a traditional narrative collected by the fourth author of this paper, a linguist with extensive knowledge of Arapaho. This dataset amounted to 128 words of Arapaho text, and 310 words in the English free translation. The data were annotated by the same two annotators as experiment 1, following the same procedure. The first 10 lines were again used as a pilot, followed by annotation of 25 lines for the purposes of IAA calculation. Annotators took an average of 11 and 8.2 minutes per line for these 25 lines, respectively. Annotators had similar resources at their disposal as for Kukama: transcription, morpheme breakdown, and morpheme and sentence-level translations in English. They could draw on an online Arapaho-English lexical database (Cowell, 2021), and a grammar (Cowell and Moss Sr, 2008).

6.2 Evaluation procedure

Inter-annotator agreement was calculated in the same way as described for Kukama in 5.2, and is reported in Table 2. Again, gold standard annotations were constructed for each line by first adjudicating between the two main annotators, and subsequently discussing every annotation with the language expert, who adjusted this adjudicated annotation as necessary. Both for IAA calculation and comparison with the gold standard, Smatch was used.

6.3 Annotation results

The IAA scores of the Arapaho experiment are listed in the first two lines of Table 2. The scores for identity are somewhat lower than those for Kukama, even though they are still in the mid 0.7 range.

	Precision	Recall	F-score
Ann. 1 - Ann. 2 Identity	0.79	0.73	0.76
Ann. 1 - Ann. 2 Compatibility	0.81	0.74	0.77
Ann. 1 - Gold Identity	0.79	0.74	0.76
Ann. 1 - Gold Compatibility	0.80	0.75	0.77
Ann. 2 - Gold Identity	0.90	0.92	0.91
Ann. 2 - Gold Compatibility	0.90	0.92	0.91

Table 2: Arapaho IAA and Accuracy scores

Considering compatible annotations as agreements has a more pronounced impact on IAA than in Kukama, bringing them up into the same 0.8-region as Kukama IAA. This somewhat lower agreement score, at least before factoring in compatibility, together with the higher average annotation time per line, indicates that the annotators had more difficulty annotating Arapaho than Kukama and were more cautious, opting for more coarse-grained annotation values more often. Nevertheless, the fairly high precision, recall, and F-scores show that both annotators were again able to apply the UMR guidelines to Arapaho in consistent ways.

Lines 3-6 of Table 2 again show the Smatch scores comparing each annotator's initial result with the final gold standard annotation. Once again, accuracy is high, with scores approaching 0.80 for one annotator, and surpassing this threshold for the other. This discrepancy may be at least partially explained by the fact that the higher-scoring annotator took significantly more time per line – more thorough consultation of the reference materials may, unsurprisingly, correspond to higher annotation accuracy.

7 Discussion: Implications for Annotation Workflows

This paper set out to assess whether a typologically trained linguist can semantically annotate languages they do not speak or otherwise have expertise in, in order to facilitate the creation of computational resources for indigenous languages without having to further burden field linguists or communities. The IAA scores for the experiments discussed in sections 5-6 suggest that this is relatively possible: we were able to reach fairly high agreement between annotators, and between each annotator and an expert in the target language, with a minimal amount of practice. In other words, Kukama and Arapaho morpheme-by-morpheme glosses and free translations, together with reference materials such as a dictionary and grammar, allowed annotators to interpret the meanings of sentences in these languages fairly accurately. However, the types of disagreements between annotators, and between annotators and language experts, point towards a number of ways in which there is still room for improvement regarding the workflow described here.

On the one hand, a fair number of disagreements between annotators, and inaccuracies on their part, stemmed from inconsistencies and/or gaps in the available materials. For example, in the Kukama sentence in Figure 1, the Spanish and English free translations are not internally consistent – the English translation does not contain the word 'almost', and in the Spanish version, the word for 'wanted' is between brackets to indicate the wanting is implied. Even though Vallejos (2016) details the use of *iyara* as both an adverb meaning 'almost' and a desiderative marker, these inconsistent translations would lead annotators to make different decisions.

Issues like these are likely to be quite frequent when working from materials collected in language documentation projects. These materials are often collected over multiple years and by multiple people, and analyses often change throughout these periods and are not always automatically updated in materials collected earlier.

Other disagreements and inaccuracies stemmed from notational and content-based common practice in language description and grammar writing. For instance, meanings of derived stems or compounds are often not transparently clear from the sum of the glosses of the component parts: the Kukama stem *ɨtsɨ-kaka* is glossed in the annotated text as 'be scared-REC', but is listed in Vallejos (2016, p. 202) as one of a set of verbs with which the reciprocal marker takes on an inchoative function ('become scared'). Since the transcriber/translator of the text opted for 'idiomatic' translations on the utterance-level rather than a more 'literal' translation ('my daughters call scared,' rather than something like 'my daughters call, having become scared'), such nuances of meaning slip between the cracks of the morpheme-level gloss and the utterance-level translation. This lost information can have important ramifications for annotation choices, in this case for the aspect annotation of this predicate.

Figure 1: Kukama, Line 2 of annotated text

Regarding content, many Arapaho noun incorporation and associated motion constructions were difficult to annotate, since UMR guidelines for event and participant identification rest on syntactic tests that are not always discussed in detail in reference grammars, such as the ability of a verb with incorporated object to co-occur with an independent NP co-referential with this object, (Vigus et al., 2020).

Even though it is unrealistic to expect field linguists to gear their grammar-writing practices specifically towards creating materials with maximal usefulness for semantic annotation, there may be ways in which the impact of these issues can be limited. For example, the issue of meanings slipping between the cracks of morpheme-level glosses and utterance-level translations could partially be remedied by providing an intermediate level of granularity in translations – translations at the word level. Many field linguists provide the texts they collect with word-level translations in the early stages of their analysis (before they have a thorough understanding of the morphological structure of the language). For certain types of languages (e.g. languages with complex and irregular morphophonemics, or languages that make extensive use of derivation and compounding with idiosyncratic semantics for word-formation), linguists may continue producing these word-level translations in their databases even at more advanced stages of analysis. These word-level translations are usually not exported when texts are published or cited as examples in linguistics publications. However, providing them to semantic annotators would be likely to significantly improve inter-annotator agreement and overall accuracy.

For those aspects of descriptive linguistic practice which are harder to influence, it may be easier to compromise on the side of the annotation workflow. A workflow where annotators are given time to read a grammar of the language and familiarize themselves with its structure before annotating, rather than using the grammar only as a reference work to look up constructions and their functions in an ad-hoc way during annotation is likely to result in higher agreement and accuracy scores. And, rather than hoping that annotators can reach a high enough accuracy to establish gold standard annotations by themselves, allowing for consultation and adjudication with a language expert is likely to remain necessary.

8 Summary and conclusions

This paper has aimed to raise awareness among computational linguists that the languages currently treated in NLP research and shared tasks as low-resource languages, and especially no-resource languages lacking any digital presence to speak of, do not just present problems of scale in comparison to major world languages. They also present qualitatively different challenges, especially regarding the recruitment of qualified annotators. The numerous languages with small speaker numbers and limited resources, which are often spoken in more remote areas, pose unique challenges for creating digitally robust and cross-linguistically comparable semantic annotation projects. Nevertheless, this paper showed that workflows can be designed which allow for semantic annotation without placing an undue burden on the limited resources of native speaker communities. Semantically annotated corpora developed in this way, and computational resources derived from them, have the potential to both provide benefits to indigenous communities, and to allow us to progress in our understanding of lexical and grammatical semantics.

9 Ethical Considerations

In recent years and decades, indigenous and other scholars have cautioned against the "mining" of indigenous language data – the use of such data for research purposes without consultation with the community that speaks the language and without sharing benefits with that community. Even though this paper shows that non-speakers of a language can be recruited to do semantic annotation, we strongly recommend that such projects be designed and conducted in collaboration with community members.

References

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.

Emanuele Bastianelli, Giuseppe Castellucci, Danilo Croce, and Roberto Basili. 2013. Textual inference and meaning representation in human robot interaction. In *Proceedings of the Joint Symposium on Semantic Processing. Textual Inference and Structures in Corpora*, pages 65–69, Trento, Italy.

Lars Borin. 2009. Linguistic resources for the languages of the world. Paper presented at the GF Summer School, Gothenburg, 26 August 2009.

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.

Deng Cai and Wai Lam. 2020. AMR parsing via graph-sequence iterative inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1290–1301, Online.

Shu Cai and Kevin Knight. 2013. "Smatch: An evaluation metric for semantic feature structures". In *"Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)"*, pages 748–752, Sofia, Bulgaria. Association for Computational Linguistics.

Andrew Cowell. 2021. Arapaho Lexical Database, Version 2021. Boulder, CO: University of Colorado. Available at https://verbs.colorado.edu/arapaho/public/view_search).

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

William Croft and Logan Sutton. 2017. Construction grammar and lexicography. In Patrick Hanks and Gilles-Maurice de Schryver, editors, *International Handbook of Modern Lexis and Lexicography*. Springer, Berlin.

David M Eberhard, Gary F. Simons, and Charles D. Fennig. 2020. "Ethnologue: Languages of the World. Twenty-third Edition". http://www.ethnologue.com.

Gilles Fauconnier. 1985. *Mental Spaces: Aspects of Meaning Construction in Natural Language*. Cambridge University Press, Cambridge.

Weston Feely, Mehdi Manshadi, Robert E Frederking, and Lori S Levin. 2014. The CMU METAL Farsi NLP approach. In *LREC*, pages 4052–4055.

Kira Griffitt, Jennifer Tracey, Ann Bies, and Stephanie Strassel. 2018. Simple semantic annotation and situation frames: Two approaches to basic text understanding in LORELEI. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

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.

Pavan Kapanipathi, Ibrahim Abdelaziz, Srinivas Ravishankar, Salim Roukos, Alexander Gray, Ramón Fernandez Astudillo, Maria Chang, Cristina Cornelio, Saswati Dana, Achille Fokoue, Dinesh Garg, Alfio Gliozzo, Sairam Gurajada, Hima Karanam, Naweed Khan, Dinesh Khandelwal, Young-Suk Lee, Yunyao Li, Francois Luus, Ndivhuwo Makondo, Nandana Mihindukulasooriya, Tahira Naseem, Sumit Neelam, Lucian Popa, Revanth Gangi Reddy, Ryan Riegel, Gaetano Rossiello, Udit Sharma, G P Shrivatsa Bhargav, and Mo Yu. 2021. Leveraging Abstract Meaning Representation for knowledge base question answering. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3884–3894, Online. Association for Computational Linguistics.

Maria Koptjevskaja-Tamm and Martine Vanhove. 2012. New directions in lexical typology. *Linguistics, Special issue*, 50.3.

Knud Lambrecht. 1994. *Information Structure and Sentence Form. Topic, Focus, and the Mental Representation of Discourse Referents*. Cambridge University Press, Cambridge.

Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, David R Mortensen, Graham Neubig, Alan W Black, et al. 2020. Universal phone recognition with a multilingual allophone system. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8249–8253. IEEE.

Sarah Moeller. 2021. *Integrating Machine Learning into Language Documentation and Description*. Ph.D. thesis, University of Colorado at Boulder.

Sarah Moeller and Mans Hulden. 2018. Automatic glossing in a low-resource setting for language documentation. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, pages 84–93, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Katharina Probst, Lori Levin, Erik Peterson, Alon Lavie, and Jaime Carbonell. 2002. MT for minority languages using elicitation-based learning of syntactic transfer rules. *Machine Translation*, 17(4):245–270.

Mrinmaya Sachan and Eric Xing. 2016. Machine comprehension using rich semantic representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 486–492.

Gary Simons and Abbey Thomas. 2019. Assessing digital language support as a factor in language vitality. *6th International Conference on Language Documentation and Conservation. University of Hawai'i at Manoa, 28 February - 3 March 2019.*

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

Rosa Vallejos and Rosa Amías. 2015. *Diccionario kukama-kukamiria castellano*. FORMABIAP.

Jens E. L. Van Gysel, Meagan Vigus, Jayeol Chun, Kenneth Lai, Sarah Moeller, Jiarui Yao, Tim O'Gorman, Andrew Cowell, William Croft, Jan Huang, Chu-Ren Hajič, James H. Martin, Stephan Oepen, Martha Palmer, James Pustejovsky, Rosa Vallejos, and Nianwen Xue. 2021. Designing a uniform meaning representation for natural language processing. *Künstliche Intelligenz*, pages 1–18.

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. Association for Computational Linguistics.

Meagan Vigus, Jens E. L. Van Gysel, Tim O'Gorman, Andrew Cowell, Rosa Vallejos, and William Croft. 2020. Cross-lingual annotation: a road map for low- and no-resource languages. In *Proceedings of the Second International Workshop on Designing Meaning Representations*, pages 30–40, Barcelona Spain (online). Association for Computational Linguistics.

Lei Zhang and Achim Rettinger. 2014. X-LiSA: Cross-lingual semantic annotation. *Proceedings of the VLDB Endowment*, 7(13):1693–1696.