Bootstrapping UMR Annotations for Arapaho from Language Documentation Resources

Matt Buchholz, Julia Bonn, Claire Benét Post, Andrew Cowell, Alexis Palmer
University of Colorado Boulder, Department of Linguistics
{matthew.buchholz, julia.bonn, benet.post, james.cowell, alexis.palmer}@colorado.edu

Abstract

Uniform Meaning Representation (UMR) is a semantic labeling system in the AMR family designed to be uniformly applicable to typologically diverse languages. The UMR labeling system is quite thorough and can be time-consuming to execute, especially if annotators are starting from scratch. In this paper, we focus on methods for bootstrapping UMR annotations for a given language from existing resources, and specifically from typical products of language documentation work, such as lexical databases and interlinear glossed text (IGT). Using Arapaho as our test case, we present and evaluate a bootstrapping process that automatically generates UMR subgraphs from IGT. Additionally, we describe and evaluate a method for bootstrapping valency lexicon entries from lexical databases for both the target language and English. We are able to generate enough basic structure in UMR graphs from the existing Arapaho interlinearized texts to automate UMR labeling to a significant extent. Our method thus has the potential to streamline the process of building meaning representations for languages without existing large-scale computational resources.

Keywords: Uniform Meaning Representation, Polysynthetic Languages, Arapaho, Interlinear Gloss Text, Bootstrapping, Endangered Languages, Low-resource Languages

1. Introduction

Uniform Meaning Representation (UMR) is a graph-based semantic labeling formalism that follows in the footsteps of Abstract Meaning Representation (AMR) (Banarescu et al., 2013). Compared to AMR, UMR has an enhanced sensitivity to cross-linguistic diversity, since UMR was developed in close conjunction with several field linguists working with a typologically diverse set of indigenous languages (Van Gysel et al., 2021). The UMR labeling system can be time-consuming to execute, especially if annotators are starting from scratch. In this paper, we focus on methods for bootstrapping UMR annotations for a given language from existing resources, and specifically from typical products of language documentation work, such as lexical databases and interlinear glossed text (IGT). Language documentation resources typically are small in scale and rich in linguistic detail, representing hundreds or thousands of hours of previous analysis and annotation work.

As an example, Figure 1 shows, for one Arapaho sentence, the semantic graph that represents the core participants of the denoted event and their relationships to one another. The full UMR representation for that same sentence appears in example (3) (with the parallel English UMR in (1)), and the IGT in Table 1. The UMR representation encodes information about semantic relationships that does not appear in the IGT.

Using Arapaho as our test case, we present and evaluate a bootstrapping process that automatically generates UMR subgraphs using information from labels in the IGT. The approach requires specification of some language-specific mappings. Additionally, we describe and evaluate a method for bootstrapping valency lexicon entries from lexical databases for both the target language and English. We automatically generate a rough set of language-specific semantic predicate argument structures (i.e. rolesets) using lexical information from the Arapaho Lexical Database (Cowell, 2010) as well as valency structure information bootstrapped from VerbNet (Schuler, 2005; Brown et al., 2022) and the English PropBank Lexicon (Palmer et al., 2005; Bonial et al., 2014; Pradhan et al., 2022). Neither component produces complete UMR graphs. Instead, both are intended to support manual production of graphs, using existing resources to dramatically reduce the amount of annotation effort required.

Our contributions are: a) a method for semi-automatic generation of roleset files; b) a method for semi-automatic generation of partial UMR graphs; c) application of these methods to a polysynthetic language, and evaluation of the materials produced; and d) the foundation for a less language-specific, more generalizable approach.

After providing background on UMR annotation (2.1), IGT (2.2), and the Arapaho language and datasets (3), we describe and evaluate our methods for bootstrapping rolesets (4) and graphs (5). We conclude with limitations and next steps (6).

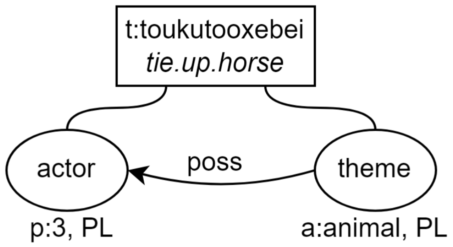

Figure 1: Partial UMR for one Arapaho sentence.

2. Background

2.1. Uniform Meaning Representation

Like AMR before it, UMR represents sentence semantics in the form of directed, rooted, graphs made up of nested predicate argument structures (Gysel et al., 2021). Unlike AMR, UMR aims to do this in a cross-linguistically uniform way, with special accommodations for languages with few to no existing lexical resources. Both AMR and UMR graphs abstract away from variations in morphosyntactic form. In other words, the predicate argument structures used in the graphs are agnostic for part of speech. Graphs for the utterances *“he rode his horse to Shoshone”* and *“his horse-ride to Shoshone”* are treated as logically equivalent, meaning the same predicates, arguments, and relations should be used in both.

For each sentence in a document, a **sentence-level graph** is created in which tokens from the sentence are implemented as concepts (nodes) and the semantic relations between concepts are represented as roles (edges). Graph concepts take the form *(s1c / concept)*, where *s1c* is a variable that uniquely identifies the concept in the document. A **document-level graph** is added for each sentence in order to track temporal dependencies, modal dependencies, and coreference relations. Concept variables are shared across the sentence- and document-level graphs, such as the variable `s2h` for the horse in (1). The top part of (1) is the sentence graph, and the bottom is the document graph.

(1) *“Then they tied up their horses.”*

```
(s2t / tie-up-04
    :ARG1 (s2p / person
        :refer-person 3rd
        :refer-number plural)
    :ARG2 (s2h / horse
        :refer-number plural
        :poss s2p)
    :temporal (s2t2 / then)
    :aspect performance
    :modal-strength full-affirmative)

  (s1s0 / sentence
    :temporal ((DCT :before s2t))
    :modal ((AUTH :full-affirmative s2t)))
    :coreference ((s2p :same-entity ...)
            (s2h :same-entity ...)))
```

UMR provides an inventory of **general semantic roles** such as *:temporal* and *:poss*, which are shown in the example above. It provides a similar inventory of **abstract concepts**, such as *(s2p / person)*, which can be used for uniform handling across languages, or concepts not explicitly expressed in the sentence. In (1), *person* is used as part of UMR’s annotation strategy for pronominal elements. An appropriate abstract concept serves as the head node, with *:refer-person* and *:refer-number* attribute roles marking person and number. Note that the variable *s2p* representing the pronoun ‘they’ is re-entered into the graph under the *:poss* relation, showing that ‘they’ are the possessors of the horses.

Turning to the **document-level graph**, the variable for the *tie-up* predicate is included in the temporal dependency, placed *:before* the document creation time (DCT). This event is also assigned a modal strength value (fully affirmative, attributed to the author– an expansion of the same relation first marked in the sentence-level graph). Graphs for other sentences in the document might show ‘them’ or ‘their horses’ as involved in coreference relations with mentions in other sentences, using the *:same-entity* coreference role.

Like AMR and PropBank (Palmer et al., 2005), UMR uses an inventory of predicate argument structures called **rolesets**. Rolesets provide

TX	ne’toukutooxebei3i’			hinit	neeheyeiniihi’	
MB	ne’-	toukutooxebei	-3i’	hinit	neeheyein-	iihi’
GE	then-	tie up horse	-3PL	right there	nearby-	ADV
PS	PREF-	VAI.INCORP	-INFL	PART	PREF-	DERIV
FT	*“Then they tied their horses right there nearby”*					

Table 1: IGT for Arapaho with the following tiers: word (**TX**), morpheme (**MB**), gloss (**GE**), part of speech (**PS**), and free translation (**FT**).

sense disambiguation for a given event lemma, as well as an associated set of numbered roles that correspond to the event's semantically essential participants. The different morphosyntactic expressions that can be used for the event are also listed, called **aliases**. In (1), we saw the English roleset *tie-up-04* in use; it is presented in more detail in (2). Note that this particular roleset can be used for instances of 'tying-up' events expressed as verbs or as nouns.

(2) **tie-up-04:** *bind with rope*
ALIASES: tie_up-*v*, tying_up-*n*
:ARG0 agent
:ARG1 entity tied
:ARG2 the rope

UMR for new languages. Language-specific rolesets like this form the basis of UMR graphs, for languages that have such rolesets. For languages like Arapaho that do not have an existing inventory of rolesets, UMR suggests that annotators use unmodified surface form tokens as graph predicates. Argument relations then come from a set of general semantic participant roles (e.g., *:actor*, *:theme*, *:instrument*). Creation of a roleset lexicon is costly and time consuming– a problem that can be catastrophically prohibitive for low resource languages.

In this work, we explore the potential of language documentation resources like interlinear glossed text to bootstrap the UMR annotation process.

2.2. Interlinear Glossed Text

Interlinear glossed text (IGT), shown in Table 1, is a richly-annotated data format often produced as part of language documentation projects. In this format, a series of tiers are provided for an utterance that list information such as a phonetic transcription, corresponding orthographic form, word segmentation, morphological glossing, part of speech labels, and free translation.

IGT has long been an integral data format for linguistics; its exploration in NLP is relatively recent. One line of research seeks to develop models to produce IGT, or to speed up its production (Palmer et al., 2009; Georgi et al., 2015; Moeller and Hulden, 2018; Zhao et al., 2020; Barriga Martínez et al., 2021; Shandilya and Palmer, 2023; He et al., 2023; Ginn et al., 2023, among others). A second major research thread extracts various types of linguistic information from IGT; the current paper follows this direction. Some examples are IGT for morphological paradigm induction (Moeller et al., 2020), for learning grammar specifications (Bender et al., 2014), improving dependency parsing (Georgi et al., 2012), or extracting typological features (Lewis and Xia, 2008).

3. Arapaho data

Arapaho is a polysynthetic and agglutinating Algonquian language, featuring extensive head-marking morphology on the verb stem, and free word order (Cowell and Moss Sr, 2011). These properties of the language make it an especially challenging test case for meaning representations, and we discuss some of these complexities in sections 4 and 5.

3.1. Arapaho IGT

The data used here come from the Arapaho Text Database (Cowell, 2024) and consist of Arapaho narratives, recorded in audio and/or video by Andrew Cowell on the Wind River Reservation over the last 20 years. The data was transcribed and translated by Cowell working with consultants, and then interlinearized using Toolbox software. There are over 90,000 sentences in the text collection.

The Arapaho data are labeled using a common protocol for Algonquian languages. Nouns are either NA or NI (grammatically animate or inanimate), and verbs are VAI, VII, VTI or VTA (intransitive, with either animate or inanimate subject; or transitive, with either animate or inanimate object). Labeling includes tense, aspect and modality marking, as well as person and number of pronominal affixes. In addition, the Arapaho IGT includes some more fine-grained labels that are useful for determining valency: VAI.R (reflexive/reciprocal), VAI.PASS (passive), VTA.D (ditransitive), VAI.INCORP (noun incorporation), and others.

A notable feature of all Algonquian languages is proximate/obviative marking: when two or more third-person participants occur in a discourse, one must be selected as most important, and all others are marked with an obviative suffix (if NA), and have obviative agreement markers on associated verbs (for both NI and NA forms). This marking is independent of subject, agent or undergoer status, or syntactic position in a clause. Algonquian verbs include special marking to clarify whether the proximate or obviative participant is the semantic agent in the clause, and the database includes labeling on both the verbs and the nouns to allow this disambiguation to be done automatically.

Natural discourse data varies greatly in complexity, and while many sentences in Arapaho consist of a single verb, some sentences include subordinate clauses of various types, and thus multiple verbs. UMR graphs for complex sentences can use a number of different strategies. In this work, we focus on generating subgraphs for individual verbs. For sentences with this type of subordination, we generate multiple subgraphs and leave it to the annotator to nest them appropriately.

3.2. UMRs for Arapaho

The gold standard UMR graphs we use for planning and comparison in this work come from the UMR 1.0 data release (Bonn et al., 2023).

A UMR graph for the Arapaho sentence corresponding to the English ex. (1) is shown in (3).

(3) *" Ne'toukutooxebei3i'. "*
(TR: "Then they tied up their horses.")

```
(s53t / toukutooxebei-00
    :actor (s53p / person
        :refer-person 3rd
        :refer-number plural)
    :theme (s53a / animal
        :refer-number plural
        :poss s53p)
    :aspect performance
    :modal-strength full-affirmative)

  (s1s0 / sentence
    :temporal ((DCT :before s53t)
            (... :after s53t))
    :modal ((AUTH :full-affirmative s53t)))
    :coreference ((s53p :same-entity ...)
            (s53a :same-entity ...)))
```

The structure of the Arapaho graph is quite similar to the graph for the parallel English sentence. One difference is that the temporal element meaning 'then' is encoded as a prefix in Arapaho, and it does not get its own node in the graph. However, the semantics are still captured in the document-level graph by indicating that the tying up event, *s53n*, occurs *:after* whatever event occurred in the previous sentence.

In this Arapaho sentence, both participants are expressed via the verb. The actors, *-3i'* ('they'), appear as a pronominal index marked on the verb, while the 'horses' are lexically encoded as part of the verb stem through noun incorporation (underlying incorporated noun form *-ôoxew-*). As this morphological component cannot be extracted from the stem, and because it differs from the form used as a stand-alone noun (*'woxhoox'*), we use an abstract *(a / animal)* concept in the graph to represent the horses.

4. Bootstrapping Rolesets from Existing Lexical Resources

4.1. Motivations

For polysynthetic and agglutinating languages like Arapaho, a single event predicate may include many morphological components that cover the entirety of the event's semantics. These may include multiple participants, adjunct arguments, tense, modal and aspectual modifiers, spatial modifiers, valency changing affixes, and more. If annotators treat every single surface word form generated by these complex morphological processes as a unique UMR predicate,predicates will proliferate to the point of absurdity, and the corpus will consist entirely of predicates that occur exactly once. Rolesets (see section 2.1) help to abstract away from morphosyntactic variation, but producing a roleset lexicon for a new language is an enormous undertaking. Here we present a methodology for automatically generating rough rolesets that cluster derivationally-related forms under a uniform predicate label.

As an example, Table 2 shows just a fraction of the derivationally-related forms that have to do with the morpheme *3i'oku* ('sit') in Arapaho.

Derived Form	Definition	IGT label
3i'oku-	sit, be sitting	vai
3i'okuut	sitting	ni
tees3i'oku-	sit on top of s.t.	vai
3ei3i'oku-	sit under s.t.	vai
3i'okuutoneihiinoo-	sat.at-	vii.pass-
3i'okuutooni-	people sit	vii.impers
3i'okuuton-	sit at object	vta
3i'okunooo	one sitting at/with	na.deppart
3i'okuu3oo	thing sat upon	ni.deppart
3i'okuh-	make sit down	vta
3i'okuno'oobe-	sleep sitting up	vai
hou3i'oku-	sit perched	vai
ko'ein3i'oku-	sit in a circle	vai
tei'3i'oku-	sit strongly	vai
3io'kuuto'o	chair	ni
hinen toh3i'okut	OldMan Mt.	placename
Wox 3i'ok	Sitting Bear	persname

Table 2: Derivations for Arapaho verb stem *3i'oku*.

The top section of this list shows verbs that would ideally be clustered into a single roleset. These derivations include a passive, an impersonal form (with a generic 'sitter'), an eventive noun, and various forms in which some spatial landmark is made explicit. The next section includes forms that name participants of the event rather than the event itself (a 'sitter' and a 'thing sat on'). These would utilize the same roleset as the verbs.

The third and fourth sections of the list should not be clustered with the previous sections. Forms in the third section include additional semantics that change the nature of the event, such as causation. The last section contains place and person names that do not directly reference events at all.

4.2. Methods

We use the following approach to produce a reasonable–if not exact–inventory of rolesets from existing resources (in this case, a traditional lexical database and an IGT corpus).

1. **Identify seed verb classes.** The objective here is to identify a small set of verb classes, based both on semantic similarity and on the structure of their core arguments. A linguist familiar with the target language is best suited to this task.

2. **Associate verb classes with semantically similar English verbs, and retrieve rolesets for those English verbs from existing resources.** We leverage the rolesets developed for English VerbNet and PropBank.[1]

3. **Verify fit of English roleset for seed verb classes, modifying the rolesets as necessary.** With some IGT data, sentences containing verbs of a candidate verb class can be sampled, and the argument structure in those sentences can be compared to the English verb's roleset Once more, a linguist familiar with the target language is a useful resource in this task.

4. **Collect aliases for each roleset.** The objective here is to extract morphosyntactic variants which describe the same event as the verb root, as demonstrated above in the top section of Table 2. Here, traditional lexicons and IGT are useful resources.

4.3. Arapaho case study

1-3. Seed verb classes and their English rolesets. For Arapaho, we identified six seed verb classes. Here we discuss two examples. One seed class contains 'give'-type verbs, all of which express some type of transfer; representative samples include:

(4) *biin-* 'give s.t. to s.o.'
tou3e'ein- 'give s.t. as gift to s.o.'
neeceenohoo3- 'give s.t. ceremonially to s.o.'
bexoow- 'bestow s.t. on s.o.'

With this seed verb class in mind, we can bootstrap arguments from a semantically similar verb class in VerbNet or a representative roleset from PropBank. Cross-referencing both VerbNet and PropBank is useful because these resources offer slightly different takes on semantic argument structures. PropBank rolesets may be more nuanced than their associated VerbNet classes and have more roles. In some cases, this nuance is helpful for bootstrapping to a similar verb in another language, but in other cases, the nuance is English-specific. PropBank roles are also numbered according to their syntactic primacy. *:ARG0* is reserved for proto-agents, *:ARG1* is reserved for proto-patients, and direct objects and goals tend to appear as *:ARG2*, etc. (Bonial et al., 2012).

In Arapaho, 'give'-type verbs tend to have the same basic argument structure as VerbNet (*give-13.1*) and PropBank (*give-01*) – an Agent, a Patient, and a Recipient. Thus, the 'give'-type transfer verbs are a case where an English roleset can map cleanly to the Arapaho class.

As another example, we define a group for simple weather verbs, such as those in (5); all are inanimate, intransitive verbs which typically do not take any arguments but a dummy 'it'.

(5) *hoosoo-* 'it is raining'
beeci- 'it is snowing'
wo'wu3oonoosoo- 'it is hailing'

In the case of the Arapaho weather verbs, we first reference the VerbNet class *weather-57*. This class includes a single thematic role, THEME, which applies to arguments like 'cats and dogs' in 'it's raining cats and dogs'. This roleset clashes with the Arapaho verbs, which are all intransitive and cannot take a THEME argument. However, corresponding PropBank rolesets such as *rain-01* and *snow-01* frequently include an argument for the location in which the weather occurs. This argument, while oblique, is frequently present with these verbs in Arapaho as well, so we include it as the sole numbered argument for bootstrapping Arapaho weather verb rolesets.

In the case that a given seed class has no suitable English roleset to use as a template, one could still hand-generate a roleset for this process. We suspect this would be a rare occurrence, and were able to find reasonable English rolesets for each of our seed verb classes.

4. Aliases. To identify aliases, we scan the existing Arapaho lexicon, comparing the morphology of the root verb to the morphology of other entries in the lexicon. We leverage the part-of-speech labeling of the lexicon-—which includes valency information-—in determining whether a potential alias should belong in the roleset of the current root or whether it belongs in its own roleset (e.g. for forms with increased valency).

4.4. Arapaho evaluation

We generate rolesets for six seed verb classes. We evaluate our automatically generated rolesets (and their associated aliases) in two stages, an initial informal analysis (as a sanity check) and a later more structured evaluation.

[1] Unified Verb Index, access to VerbNet and PropBank: https://uvi.colorado.edu/

In the informal evaluation, an Arapaho expert checked one generated roleset for each seed class against several dozen occurrences of the main predicate in the text database. These looked reasonable, prompting the second evaluation, in which the rolesets for 12 verbs from the six seed classes were checked against 10 randomly-selected corpus sentences for each verb.

The newly-generated Arapaho frame files correctly capture the argument structure in 109 of the 115 Arapaho sentences, almost 95% of the time.[2]

Frequency of selected verbs in the dataset. While we picked only six classes of verbs to evaluate in this work, those classes were strategically chosen based on structural and semantic features. A post hoc analysis shows that of the approximately 88,000 verb tokens in the Arapaho Text Database, the 12 verbs evaluated in this study account for about 3,000 tokens; thus, generating frame files for those twelve verbs provides coverage of around 3.4% of the verb tokens in the database. It should be noted that the 12 verbs evaluated were not picked via a prior analysis of their frequency; however, the top five most frequent verbs from each of the six verb classes proposed and analyzed here account for some 5,000 tokens in the dataset. Therefore, defining even a few classes of verbs for roleset bootstrapping can quickly provide argument structures for many sentences, greatly speeding annotation.

Discussion. The six failures noted in the larger evaluation all occurred with verbs meaning roughly 'procure something for someone'. The English rolesets include roles for thing procured, person from whom the thing was procured, and the recipient/benefactee. But in Arapaho, verbs of this type similarly index a thing procured and a recipient/benefactee, with the third role of procurer (rather than source). There were six Arapaho sentences in which an explicit noun occurred, referring to the procurer; these sentences all showed a mismatch with the English-based role sets.

One other problem was noted in the informal analysis: the original English frame files sometimes lack sub-sets for argument structures which actually exist in both English and Arapaho. For example 'kick' in English does not account for a meaning and argument structure of 'kick an object to a person/place' as opposed to 'kick something/-someone.' While this type of sentence did not occur among the 10 randomly selected Arapaho sentences with this verb, the broader informal analysis revealed two examples in the text database with this role set structure. In this case, it was not our methodology of transferring English role sets to Arapaho which failed, but the gap in the original English rolesets. Otherwise, the frame file transfer was highly successful (though admittedly, for verb stems which were predicted to be parallel to English in role set structure).

The only major problem encountered was in the determination of aliases for the head verb. The generated rolesets include many forms which should be excluded from the aliases, most notably secondary verb stems with a different valence from the head verb.

4.5. Generalizability of technique

While Arapaho's morphological complexity will require ongoing refinement of this process, we believe that generating even a rough set of aliases can greatly expedite the UMR annotation process. Our relative success in bootstrapping rolesets for Arapaho verbs suggests that this technique could generalize well to other languages, given the extreme typological differences between English and Arapaho with respect to verbal morphology and morphosyntax.

5. Bootstrapping graphs from IGT

5.1. Motivation

Building a language resource such as an IGT corpus requires an enormous amount of effort and analysis. Further, UMR representations encode much of the same information as a glossed text. For linguists who have already spent time building IGT, the task of "re-annotating" their dataset in a new format (UMR) can be daunting and hard to justify. We therefore seek to expedite the process of UMR graph annotation by leveraging existing IGT labels. With a little bit of scripting, we can partially automate the process of UMR graph generation using existing gloss labels and simple heuristics.

5.2. General approach

Given an IGT dataset, we take the following approach to quickly generate partial UMR graphs for each sentence:

1. **Identify gloss labels corresponding to verbs.** These verbs are the root nodes of any predicate subgraphs in the UMR representation, so correctly extracting them from the glossed text is critical.

2. **Identify gloss labels corresponding to event participants.** Depending on the language, these could be explicit noun phrases, pronouns, verbal agreement markers, etc.

[2] There were only 5 occurrences in the text database for one verb, so we evaluated 115 sentences in total.

3. **Define heuristics mapping participants to predicate roles.** This could include specific gloss labels (e.g. case or verbal agreement markers), word order cues (e.g. constituent order), etc. The focus here is on building heuristics which work well enough *most of the time*, not on defining a process to exhaustively capture all participants.

4. **Develop simple scripts to parse IGT and apply the above rules.** In other words, operationalize the above steps with code.

In this work, we concentrate on the above procedure, with the aim of establishing the "skeleton" (top-level predicate subgraph) of a UMR representation. Naturally, IGT may be rich with other information corresponding to nodes or edges in a UMR representation. For example, in the Arapaho text database, determiners are glossed `DET`, and, in UMR, determiners are represented with a `:mod` edge and a node for the specific determiner. Given the predictability of the determiner's position in an Arapaho noun phrase, representing the determiner in a corresponding UMR graph is straightforward. We largely leave the extraction of such features to future work.

5.3. Arapaho case study

This section describes how we apply this approach to the Arapaho IGT dataset; results appear in 5.4.

1. Verbs. Arapaho has four grammatical verb classes, based on their transitivity and the animacy of the arguments they accept; the corresponding gloss labels, with their expected argument structures, appear in Table 3.

Label	Subject	Object
VII	Inanimate	
VAI	Animate	
VTI	Animate	Inanimate
VTA	Animate	Animate

Table 3: Verb part-of-speech labels with expected argument structure. Rows without data in the **Object** column are intransitive verbs.

This step approximates the argument structure information found in a roleset inventory.

2. Participants. In Arapaho, explicit nouns can appear as independent tokens, but, depending on discourse context, participants may also be omitted. In either case, participants can be marked on the verb. The following table captures both participant-marking strategies as they appear in the database.[3]

Label	Meaning
NI	Inanimate noun
NA	Animate noun
1,2,3 or 4	Person marking
S	Singular (with person marking)
PL	Plural (with person marking, noun)

Table 4: Part-of-speech labels for identifying participants. Note a cumulative affix can encode both agent and patient, e.g. `-1PL/3PL` encodes first-person plural agent, third-person plural patient.

3. Mapping. Now that we have an expected argument structure and a list of potential participants for each predicate, we can define simple but effective rules to map participants to roles.

If any pronominal participants are marked on the verb, these must correspond to the subject and/or object. However, they may be "overwritten" by explicit nouns. Explicit nouns that match both the argument structure expectations in animacy and the number encoded in verbal affixes can fill corresponding subject or object roles.

Given the relatively free word order of Arapaho grammar, some cases are potentially ambiguous. For example, a clause with two explicit singular animate nouns and a `-4S/3S` or `-3S/4S` verbal suffix could allow two different argument structure mappings. Wherever possible, we use obviative markings on the explicit nouns to map arguments to the fourth-person role.[4]

If any of the expected participant roles are unfilled at this point, we fill them with inferred arguments. When the predicate expects an animate participant, we insert a generic third-person singular entity. When an inanimate participant is expected, we add a generic `thing` participant. This strategy handles cases where the participants are clear from context, and the speaker has omitted them. Finally, any participants not mapped to the subject or object roles (e.g. due to disagreement with animacy or plurality) we add to the predicate in a generic `OBLIQUE` argument.

4. Parsing. The IGT database uses simple whitespace separation to align the morpheme, gloss, and part-of-speech lines. As a preprocessing step, we collect all IGT in the database where there are different numbers of items between the morpheme, gloss, and part-of-speech lines. By hand, we re-annotate these texts so that

[3] VII verbs are obligatorily marked for agreement with the subject, but, as an artifact of how the database was annotated, this agreement marker was not always glossed. So we omit it here. Also note that for VTI verbs, no overt marker distinguishes singular or plural patients.

[4] In some sentences in the database, the obviative is either not marked (by speaker omission) or not glossed (by annotator omission). In these cases, there is no straightforward way to map an argument to the correct role.

each morpheme corresponds to exactly one gloss item and also to exactly one part-of-speech item (e.g. re-annotating the gloss for the morpheme *notii3ei* from “look for things” to “look.for.things”, as is standard across the database). Then, for each sentence, we build a partial graph as follows.

First, the script scans the part-of-speech line looking for labels which match those from Table 3; when one is found, a predicate is created with the corresponding morpheme token instantiated as the head node. Next, all possible participants are extracted from the sentence; explicit nouns are instantiated with nodes using the corresponding morpheme, while pronominal arguments are instantiated as required by UMR (e.g. a `-3PL` gloss receives a `person` node in the graph, with appropriate `refer-person` and `refer-number` attributes.) Finally, we map our candidate participant nodes to roles in the expected argument structure for the predicate.

We use the Penman Python library (Goodman, 2020) for building and manipulating the graphs.

A sample glossed text from our database follows, alongside the corresponding human-annotated graph and the automatically-generated UMR graph:

(6) Ci’he’ih’iitounowuu hitiicetino.
ci’-he’ih’ii-toun-owuu hi-icetino
too-NARRPAST.IMPERF-hold-3PL 3S-hands
PROC/PART-PREF-VTI-INFL INFL-NI.OBLPOSS.PL
“Also they were holding their hands.”

Gold standard graph:

```
(s18t / toun-00
    :actor (s18p / person
        :refer-person 3rd
        :refer-number plural)
    :undergoer (s18h / hitiicetino
        :part-of s18p)
    :aspect activity
    :modal-strength full-affirmative)
```

Auto-gen graph:

```
(t / toun
    :SUBJECT (p / person
        :refer-person 3rd
        :refer-number plural)
    :OBJECT (i / icetino))
```

In this case, the generated subgraph aligns with the core components of the gold standard graph. It does not capture aspect or modality.

5.4. Arapaho evaluation

To analyze the performance of our graph generation approach for Arapaho, we perform a manual qualitative evaluation, focusing on 3 key criteria.

The UMR data release contains gold-standard UMR graphs for approximately 400 Arapaho sentences. From these, we filter out sentences with more than one verb, and a few with no verb indicated in the IGT. From the remaining 216 sentences, we randomly select 98 sentences for analysis.[5] We ask two linguists familiar with Arapaho (two of the authors) to compare our generated graphs to the gold-standard graphs and answer the following three questions:

Criterion #1: Have we correctly identified the head verb? The auto-generated graphs correctly locate the head verb in 98/98 test cases, and thus correctly generate the root node of the UMR graph. This particular task is artificially simplified, since we exclude sentences with more than one verb. That said, it is not a trivial task. Various pseudo-verbal forms occur, as well as lexicalized imperatives and verbal nouns.

Criterion #2: Are the subject and object pronominal affixes on the verb correctly translated into the graph structure? 78 of the sentences are graphed correctly, while 20 are incorrect, an 80% success rate.

Criterion #3: Are all overt nouns correctly linked to the head verb, and correctly recognized as subject, object, or oblique forms? There are 35 overt nouns in the 98 sentences, and 26 of these are graphed with the proper argument structure, a 74% success rate. As would be expected, sentences with two overt nouns pose the greatest challenge.

Discussion. According to this evaluation, the subgraphs generated from IGT capture the core argument structure of simple sentences with a high degree of accuracy. The generated graphs are suitable foundations for full UMR graphs, and we expect to extract additional properties from the IGT. That said, there will clearly be challenges in expanding the current approach to the full range of complexity found in the sentences in the database. One challenging task is to recognize incorporated nouns. Fortunately, in this database, all verbs having incorporation are labeled `.INCORP`. Similarly, many Arapaho verbs are structurally intransitive, but semantically transitive. Again labeling indicates this, via secondary `.O` (indefinite object) and `.T` (pseudo-transitive object) labels. Subordinate clause verbs can likewise be automatically recognized based on the presence of a number of subordinating prefixes or inflectional markers. Finally, we can make improvements to the handling of quotations embedded in citational verbs. The algorithm correctly locates the citational verb as the head of the sentence, but this then leaves the argument structure of the citation itself untreated.

[5] We originally selected 100 sentences; 2 were determined to have non-trivial errors in the original IGT.

5.5. Generalizability of technique

While this work uses Arapaho as a case study, we believe the technique of mapping IGT labels to UMR graph components is broadly applicable to many languages.

To illustrate the applicability of our technique to another language, we use Quechua as an example. Quechua, an indigenous language spoken primarily in Peru, Bolivia, and Ecuador, is characterized by its highly agglutinative nature (Adelaar, 2020). Consider the glossed text in (7), alongside a sentence-level UMR representation of that text (Bonn, p.c.).[6]

(7) Mana, mama: awashts.
no mother-1.POSS weave-PAST.R3-NEG
"No, my mother didn't weave."

UMR Graph:

```
(a / away-00
    :actor (p / person
        :ARG0-of (k / kinship-91
            :ARG1 (p2 / person
                :refer-person 1st)
            :ARG2 (m / mama)))
    :polarity -
    :modal-strength negative
    :time (b / before
        :op1 (n / now)))
```

As seen in the Quechua gloss, certain basic units correspond to specific subgraphical structures in the UMR graphs. While in our Arapaho case study we largely focused on mapping participants to predicate roles, there are clear mappings from the gloss to graphical substructures. A `NEG` marker in the gloss maps to a negative `polarity` in the UMR graph; a past tense marker (`PAST`) allows one to automatically annotate the `:time` subgraph. Both of these substructures attach to the predicate head representing the root word (*away*).

It seems likely that the basic approach outlined for Arapaho could be used to bootstrap partial UMR graphs from IGT for a language like Quechua.

6. Discussion and Conclusion

We've demonstrated how two types of resources–rolesets and partial graphs–can be bootstrapped from documents which are typical products of language documentation work. When annotating a UMR graph, a robust library of rolesets allows annotators to quickly look up expected argument structures for a given predicate; additionally, the inventory of aliases for a given roleset enforces some consistency across annotated predicates. Further, even generating partial UMR graphs from IGT–such as the basic predicate-argument structures demonstrated here–can greatly speed the annotation process by reducing the amount of time an annotator spends on each graph.

Our lightweight approach to building graphs from IGT can be extended in a number of directions. With a more extensive library of rolesets, more specific argument structures for verbs will be available during graph generation. We also aim to make our system more robust and extendable to other languages, and to provide a graphical interface for it. With such a system, a linguist could map gloss labels to UMR meanings through simple heuristic rules. (In the Arapaho case, such rules might read 'nouns are labeled as NA or NI in the IGT'; 'determiners are labeled DET and attach to the nearest noun to the right'; etc.) From this mapping, the system could automatically generate partial UMR graphs from IGT, without the need for language-specific scripting.

For this case study, our approaches benefit from the detail of labeling in the Arapaho database and the availability of a language expert to help with bootstrapping. In future work, we plan to encode language-specific information just once, up front, and let that initial configuration guide generation of both rolesets and UMR subgraphs.

The straightforward methods described in this paper have tremendous potential to speed up annotation of previously-analyzed texts with sophisticated meaning representations. While we do not expect to achieve perfect UMR graphs, we do hope to support researchers by automating most of the simpler parts of UMR annotation, freeing the experts up to do the more interesting work of annotating complex structures.

Acknowledgements

This work is supported by a grant from the CNS Division of National Science Foundation (Award no: NSF_2213805) entitled "Building a Broad Infrastructure for Uniform Meaning Representations." Original Arapaho resource production was supported by grants from the Hans Rausing Endangered Language Project and the NSF/NEH Documenting Endangered Languages program.

Ethics Statement

We are aware of the ethical considerations raised by work on languages spoken by Indigenous com-

[6] Quechua data recorded by CU Boulder Quechua teacher Doris Loayza, working in conjunction with Andrew Cowell. It is original data, from speakers of Southern Conchucos Quechua. Transcriptions and translations from Loayza.

munities, such as Arapaho. Members of these communities should be able to directly benefit from any research done on their language.

One of the goals of this work is to lower the barrier to entry for producing semantic representations for a wide range of languages by making the process faster and easier, and thus making technologies based on structured semantic representations more readily available for diverse languages.

Bibliographical References

Willem FH Adelaar. 2020. Morphology in quechuan languages. In *Oxford Research Encyclopedia of Linguistics*.

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.

Diego Barriga Martínez, Victor Mijangos, and Ximena Gutierrez-Vasques. 2021. Automatic interlinear glossing for Otomi language. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 34–43, Online. Association for Computational Linguistics.

Emily M. Bender, Joshua Crowgey, Michael Wayne Goodman, and Fei Xia. 2014. Learning grammar specifications from IGT: A case study of chintang. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 43–53, Baltimore, Maryland, USA. Association for Computational Linguistics.

Claire Bonial, Julia Bonn, Kathryn Conger, Jena D Hwang, and Martha Palmer. 2014. PropBank: Semantics of New Predicate Types. In *LREC*, pages 3013–3019.

Claire Bonial, Jena Hwang, Julia Bonn, Kathryn Conger, Olga Babko-Malaya, and Martha Palmer. 2012. English propbank annotation guidelines. *Center for Computational Language and Education Research Institute of Cognitive Science University of Colorado at Boulder*, 48.

Susan Windisch Brown, Julia Bonn, Ghazaleh Kazeminejad, Annie Zaenen, James Pustejovsky, and Martha Palmer. 2022. Semantic representations for nlp using verbnet and the generative lexicon. *Frontiers in artificial intelligence*, 5:821697.

Andrew Cowell and Alonzo Moss Sr. 2011. *The Arapaho language*. University Press of Colorado.

Ryan Georgi, Fei Xia, and William Lewis. 2012. Improving dependency parsing with interlinear glossed text and syntactic projection. In *Proceedings of COLING 2012: Posters*, pages 371–380, Mumbai, India. The COLING 2012 Organizing Committee.

Ryan Georgi, Fei Xia, and William Lewis. 2015. Enriching interlinear text using automatically constructed annotators. In *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 58–67, Beijing, China. Association for Computational Linguistics.

Michael Ginn, Sarah Moeller, Alexis Palmer, Anna Stacey, Garrett Nicolai, Mans Hulden, and Miikka Silfverberg. 2023. Findings of the SIGMORPHON 2023 shared task on interlinear glossing. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 186–201, Toronto, Canada. Association for Computational Linguistics.

Michael Wayne Goodman. 2020. Penman: An Open-Source Library and Tool for AMR Graphs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 312–319, Online. Association for Computational Linguistics.

Jens E. L. Van Gysel, Meagan Vigus, Jayeol Chun, Kenneth Lai, Sarah Moeller, Jiarui Yao, Timothy J. O'Gorman, Andrew Cowell, William Croft, Chu Ren Huang, Jan Hajic, 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.

Taiqi He, Lindia Tjuatja, Nathaniel Robinson, Shinji Watanabe, David R. Mortensen, Graham Neubig, and Lori Levin. 2023. SigMoreFun submission to the SIGMORPHON shared task on interlinear glossing. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 209–216, Toronto, Canada. Association for Computational Linguistics.

William D. Lewis and Fei Xia. 2008. Automatically identifying computationally relevant typological features. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*.

Angelina McMillan-Major. 2020. Automating gloss generation in interlinear glossed text. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 355–366, New York, New York. Association for Computational Linguistics.

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.

Sarah Moeller, Ling Liu, Changbing Yang, Katharina Kann, and Mans Hulden. 2020. IGT2P: From interlinear glossed texts to paradigms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5251–5262, Online. Association for Computational Linguistics.

Alexis Palmer, Taesun Moon, and Jason Baldridge. 2009. Evaluating automation strategies in language documentation. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 36–44, Boulder, Colorado. Association for Computational Linguistics.

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

Sameer Pradhan, Julia Bonn, Skatje Myers, Kathryn Conger, Tim O'gorman, James Gung, Kristin Wright-Bettner, and Martha Palmer. 2022. Propbank comes of age—larger, smarter, and more diverse. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 278–288.

Karin Kipper Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. University of Pennsylvania.

Bhargav Shandilya and Alexis Palmer. 2023. Lightweight morpheme labeling in context: Using structured linguistic representations to support linguistic analysis for the language documentation context. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 78–92, Toronto, Canada. Association for Computational Linguistics.

Jens E. L. Van Gysel, Meagan Vigus, Lukas Denk, Andrew Cowell, Rosa Vallejos, Tim O'Gorman, and William Croft. 2021. Theoretical and practical issues in the semantic annotation of four indigenous languages. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 12–22, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jin Zhao, Nianwen Xue, Jens Van Gysel, and Jinho D Choi. 2021. UMR-Writer: A Web Application for Annotating Uniform Meaning Representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 160–167.

Xingyuan Zhao, Satoru Ozaki, Antonios Anastasopoulos, Graham Neubig, and Lori Levin. 2020. Automatic interlinear glossing for under-resourced languages leveraging translations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5397–5408, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Language Resource References

Julia Bonn. p.c. UMR Pilot: Southern Conchucos Quechua.

Julia Bonn, Chen Ching-wen, James Andrew Cowell, William Croft, Lukas Denk, Jan Hajič, Kenneth Lai, Martha Palmer, Alexis Palmer, James Pustejovsky, Haibo Sun, Rosa Vallejos Yopán, Jens Van Gysel, Meagan Vigus, Nianwen Xue, and Jin Zhao. 2023. Uniform meaning representation. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Cowell, Andrew. 2010. *A Conversational Database of the Arapaho Language in Video Format*. Endangered Languages Archive. [link].

Cowell, Andrew. 2024. *A Conversational Database of the Arapaho Language in Video Format*. Endangered Languages Archive. [link].