The Problem of Polysynthesis in UMR Annotations:

Complexities in Handling Preverbal Modification and Noun Incorporation in Arapaho

Julia Bonn & Andrew Cowell

DSNA 24
June 3, 2023

What are we doing and why are we doing it?

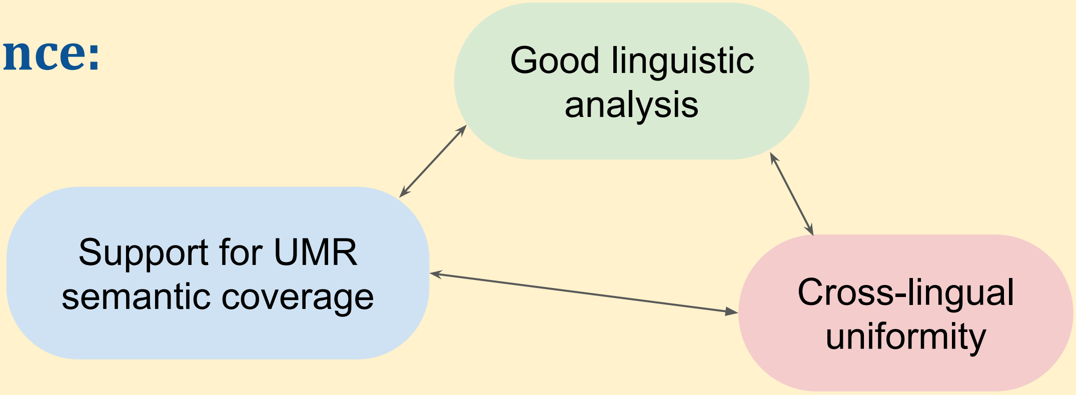
Goal of Annotation Project

- Semantic (SRL+) annotation of typologically-diverse languages for NLP **(Uniform Meaning Representation (UMR))**
- Simultaneous development of a valency lexicon when annotating low-resource languages **(PropBank rolesets)**

Focus of this Talk

- Challenges in annotating Arapaho, as a polysynthetic/agglutinating language
  - Particularly within lexical resource development:
    - Lexical entry structuring
    - Lemma selection

Striking a Balance:



Outline of presentation

- Introduce UMR
- Introduce PropBank Frame Files and Rolesets
- Introduce Arapaho
- Present three special cases demonstrating the issues, then solutions:
    - Tense/Aspect/Modality affixes on verbs
    - Path vector-adding verb stem elements
    - Noun incorporation

Uniform Meaning Representation (UMR): an Overview

- **Cross-lingual semantic annotation**
- **Nested predicate argument structures**
  - Using PropBank rolesets
  - or not-- create lexicon during annotation
- **Document-level annotation:**
  - Coreference
  - Temporal
  - Modal

Rolesets:

say-01: *speech act*

:ARG0-PAG	speaker
:ARG1-PPT	utterance
:ARG2-GOL	audience

arrive-01: *reach destination*

:ARG1-PPT	entity in motion
:ARG2-EXT	extent
:ARG3-SRC	start-point
:ARG4-GOL	destination

Martin said** that the **package had probably already arrived yesterday.

```
(s1s / say-01
   :ARG0 (s1p / person  :name (s1n / name :op1 "Martin"))
   :ARG1 (s1a / arrive-01
      :ARG1 (s1p / package)
      :temporal (s1y / yesterday)
      :QUOT s1s
     :ASPECT Performance
     :MODSTR PrtAff)
   :ASPECT Performance
   :MODSTR FullAff)

(s1 / sentence
   :coref ((s1p :same-entity s2p))
   :temporal ((PAST_REF :contained s1s)
              (DCT :before s1y)
               (s1s :before s1a))
   :modal ((AUTH :FullAff s1p)
           (AUTH :FullAff s1s)
           (s1p :PrtAff s1a)))
```

What do rolesets let us do?

- **Identify senses/lexemes**

  (“*steal* a million dollars” vs “*steal* into the night”)

- **Document thematic roles associated with a lexeme**
  - Explicit & implicit

- **Cluster lexemes and their morphosyntactic forms**
  - Leave, left, lefthand, leaving, on the left, take_leave, etc.

- **Show annotated corpus examples**

- **Give links to other lexical resources for each sense, role**

- *** Document lexical semantics at a level not covered by the UMR schema itself ***

Lexeme Clustering in English: sample Frame File

leave.xml (selected rolesets)

```
<frameset>
 <predicate lemma="leave">
  <roleset id="leave.11" name="depart from a place">
   <aliases>
    <alias framenet="Departing" pos="v" verbnet="">leave</alias>
    <alias framenet="" pos="n" verbnet="-">leaving</alias>
    <alias framenet="" pos="l" verbnet="-">take_leave</alias>
   </aliases>
   <roles>
    <role descr="entity in motion" f="pag" n="0"/>
    <role descr="starting point, location vacated" f="dir" n="1"/>
    <role descr="destination (must also be a location)" f="gol" n="2"/>
   </roles>
  </roleset>
  <roleset id="leave.02" name="bequeath, as in a will">
   <aliases>
    <alias framenet="Giving" pos="v" verbnet="13.3 13.4.1">leave</alia
    <alias framenet="" pos="n" verbnet="">leaving</alias>
   </aliases>
   <roles>
    <role descr="bequeather, donor" f="pag" n="0"/>
    <role descr="thing given" f="ppt" n="1"/>
    <role descr="benefactive / given-to" f="gol" n="2"/>
   </roles>
 </predicate>
```

```
<predicate lemma="left">
 <roleset id="left.20" name="be located on the left side">
  <aliases>
   <alias framenet="" pos="j" verbnet="">left</alias>
   <alias framenet="" pos="j" verbnet="">lefthand</alias>
   <alias framenet="" pos="p" verbnet="">to_the_lefthand_side_of</alias>
   <mwp-descriptions id="to_the_lefthand_side_of">
    <syntaxdesc slots="A B C D E">
     <token arg="" dep="" head="" pos="P" slot="A">to</token>
     <token arg="" dep="" head="D" pos="DET" slot="B">the</token>
     <token arg="" dep="" head="D" pos="JJ" slot="C">lefthand</token>
     <token arg="" dep="" head="A" pos="NN" slot="D">side</token>
     <token arg="" dep="" head="D" pos="P" slot="E">of</token>
    </syntaxdesc>
   </mwp-descriptions>
  </aliases>
  <roles>
   <role descr="theme, entity located on the left" f="ppt" n="1"/>
   <role descr="to the left of" f="loc" n="2"/>
  </roles>
 </roleset>
</predicate>
</frameset>
```

UMR: Annotation Stages for Low-resource Languages

Stage 1: annotation without rolesets

- Graph predicate = surface form predicate
- Arguments annotated with general participant roles provided by UMR

Martin told his boss that the package had probably already left the warehouse.

```
(s2t / told-00
    :agent (s2p / person
         :name (s2n / name :op1 "Martin"))
    :theme (s2l / left-00
        :theme (s2p / package)
        :start (s2w / warehouse)
        :QUOT: s2t
       :ASPECT Performance
       :MODSTR PrtAff)
    :recipient (s2p2 / person
         :ARG1-of (s2h / have-rel-role-92
               :ARG2 s2p
               :ARG3 (s2b / boss)))
    :ASPECT Performance
   :MODSTR: FullAff)
```

UMR: Annotation Stages for Low-resource Languages

Stage 3 annotation: with refined, unified rolesets

- Graph predicate = lemma form
- Numbered arguments, from rolesets

Martin told his boss that the package had probably already left the warehouse.

```
(s2t / told-00
   :agent (s2p / person
        :name (s2n / name :op1 "Martin"))
   :theme (s2l / left-00
      :theme (s2p / package)
      :start (s2w / warehouse)
      :QUOT: s2t
     :ASPECT Performance
     :MODSTR PrtAff)
   :recipient (s2p2 / person
        :ARG1-of (s2h / have-rel-role-92
            :ARG2 s2p
            :ARG3 (s2b / boss)))
  :ASPECT Performance
  :MODSTR: FullAff)
```

```
(s2t / tell-01
   :ARG0 (s2p / person
        :name (s2n / name :op1 "Martin"))
   :ARG1 (s2l / leave-11
      :ARG0 (s2p / package)
      :ARG1 (s2w / warehouse)
      :QUOT: s2t
     :ASPECT Performance
     :MODSTR PrtAff)
   :ARG2 (s2p2 / person
        :ARG1-of (s2h / have-rel-role-92
            :ARG2 s2p
            :ARG3 (s2b / boss)))
  :ASPECT Performance
  :MODSTR: FullAff)
```

Rolesets:

tell-01: *speech act*

aliases: tell-v
telling-n

:ARG0 speaker
:ARG1 utterance
:ARG2 recipient

leave-11: *depart*

aliases: leave-v
leaving-n
take_leave-lvc

:ARG0 entity in motion
:ARG1 start-point
:ARG2 destination

UMR: Annotation Stages for Low-resource Languages

Stage 1: annotation without rolesets

- Graph predicate = surface form predicate
- Arguments annotated with general participant roles provided by UMR

Stage 2: roleset development as part of annotation effort

- Graph predicate = ?
- Arguments from UMR, but what happens to arguments and other modification incorporated into the verb?

Stage 3: annotation with fully refined rolesets (as with English)

- Graph predicate = lemma form
- Numbered arguments, from rolesets

They were just doing sign language back and forth.

```
*(b / beni’beebee3sohowuuneti3i’-00
        :agent (p / person
            :refer-person 3rd
            :refer-number Plural)
        :ASPECT Activity
        :MODSTR FullAff)
```

Lexeme Clustering in English vs Arapaho:

We have guidelines for organizing lexemes into rolesets and frame files in English.

But, how do we organize frame files and rolesets for a polysynthetic and agglutinating language like Arapaho?

Arapaho:

- A Plains Algonquian language
- Fewer than 100 speakers
- Corpus approaching 100,000 sentences
- Has a comprehensive grammar and lexical database

Many ‘words’ are made up of lexicalized elements and are partly syntactic in nature:

- complex secondary derivational finals
- complex preverbal elements
- subj/obj indexing on verbs
- Noun incorporation of participants within verb stems

Many of these elements are designed to be separated into different graph nodes in UMR

TAM Affixes

toonniiciibeetei'inou'u

toon-nii-cii-beet-**ei’in**-ou’u

toon-	nii-	cii-	beet-	**ei'in**	- ou'u
INDEF-	IMPERF-	NEG-	want_to-	**know**	- 3PL
proclitic-	prefix-	prefix-	prefix-	**vti+pl**	- infl

“They don’t want to know [the language].”

```
(s1h / hei’in-00
   :POLARITY -
   :experiencer (s1p / person
      :refer-person 3rd
      :refer-number Plural)
   :stimulus [implicit]
   :ASPECT State
   :MODSTR PrtNeg)

(s1 / sentence
   :temporal ((DCT :overlap s1h))
   :modal ((AUTH :FullAff s1p)
           (AUTH :FullAff s1h)
           (s1p :NeutAff s1a)))
```

- Easily separable from the verb stem
- Not included as part of verb lexical entry in Arapaho dictionary

Roleset solution:

- Drop these from lemmas at any level in the frame files

Graph solution:

- Aspectual affixes:
  - Use :ASPECT attribute annotation
- Negating affixes:
  - Use :polarity attribute annotation
- Other modal affixes:
  - Use :MODSTR attribute annotation

Path Vector-adding Verb Stem elements:

nouutohwoo-	**nouut**-**ohwoo**-	‘dance **out of a place**’ (vector component)
ciitohwoo-	**ciit**-**ohwoo**-	‘dance **into a place**’ (vector component)
oosohwoo-	hoos-**ohwoo**-	‘do a fancy dance’ (modified concept, but no vector component)
beteee-	**beteee**-	‘dance’ (used when concept is unmodified)

- not lexically decomposable, but morphologically decomposable.

Issues:

- There’s no way to create a roleset just for ‘dance’ with the (**ohwoo-**) stem
  - must include vector component

- No good linguistic lexicon would split theses stems up morphologically

- Rolesets have not included morphological breakdown of aliases
  - risk of semantic loss inside roleset

- English/Arapaho graphs are at risk of looking very different:
  - English path info typically separate token, separate graph node
  - But in Arapaho, no separate token, no obvious separate node.
  - loss of ability to track path-related coreference in graphs

Path Vector-adding Verb Stem elements:

Roleset Solutions:

- (**beteee**-) and verbs with the (**ohwoo**-) stem belong in different frame files
  - not etymologically related

- Stems with (**ohwoo**-) clustered in same frame file (*ohwoo.xml*)
  - As different rolesets
  - Same base argument structure (dancer)
  - unique arguments dictated by vector components (path arguments)

nouutohwoo-01: *dance out of a place* **ciitohwoo-02**: *dance into a place* **hoosohwoo-03**: *dancy fancy*

:ARG0-agent dancer **:ARG0-agent** dancer

:ARG0-agent dancer

:ARG1-start start location **:ARG1-start** start location

:ARG2-goal destination **:ARG2-goal** end location

Graph Solution:

- Add 'stub' node for vector to graph
- Allows coreference tracking at document level

nihnouutohwoot

nih- nouutohwoo -t

PAST- dance.out.of.a.place -3S

'He danced out of there.'

Standard UMR:

```
(n / nouutohwoo-01
   :ARG0 (p / person
      :refer-person 3rd
      :refer-number Singular)
   :ASPECT Activity
   :MODSTR FullAff)
```

UMR with stubs:

```
(n / nouutohwoo-01
   :ARG0 (p / person
      :refer-person 3rd
      :refer-number Singular)
   :ARG1 (p2 / place)
   :ARG2 (p3 / place)
   :ASPECT Activity
   :MODSTR FullAff)
```

New additions to PropBank Lexical Resource for Polysynthetic Languages

Morpheme Inventory:

```
<morpheme lemma=”nouut-”>
 <sense key=”1” gloss=”out of a place” pos=”vstem.elem” sem=”vector” slot=”stem-initial” >
  <allomorphs />
  <graph>
   <role role=”:start” roleval=”(VV / place)” head=”” type=”direct-translation” />
   <role role=”:goal” roleval=”(VV / place)” head=”” type=”projected” />
  </graph>
 </sense>
</morpheme>
<morpheme lemma=”ohwoo-”>
 <sense key=”1” gloss=”dance” pos=”vstem.root” sem=”motion event” slot=”stem-core” >
  <allomorphs />
  <graph>
   <role role=”:agent” roleval=”” head=”” type=”argument” />
   <roleset file=”ohwoo.xml” />
  </graph>
 </sense>
</morpheme>
<morpheme lemma=”hoos-”>
 <sense key=”1” gloss=”fancy” pos=”vstem.elem” sem=”manner” slot=”stem-initial” />
  <allomorphs />
  <graph />
</morpheme>
<morpheme lemma=”-t”>
 <sense key=”1” gloss=”3S” pos=”INFL” sem=”animate” slot=”verb final”/>
  <allomorphs />
  <graph>
   <subgraph head=”verb-core”>
    (VV / person    :refer-person 3rd    :refer-number Singular)
   </subgraph>
  </graph>
</morpheme>
```

Goal:

- Resource can be referenced in rolesets for more complete semantic coverage of complex preds

Include:

- IGT info from traditional lexical database
  - gloss, pos
- More detailed semantic and syntactic info from database/grammar
  - allomorphs
  - semantic category
  - token slot
- Mappings to UMR graph elements
  - arguments
  - roleset/frame file mapping

New additions to PropBank Lexical Resource for Polysynthetic Languages

Morphologically-complex Aliases in Rolesets:

```
<predicate lemma="nouutohwoo">
  <roleset id="nouutohwoo.01" name="dance out of a place">
   <aliases>
    <alias pos="vai" uform="nó.uutóhwoo-" >nouutohwoo</alias>
    <alias pos="ni.participle" uform="nó.uutohwóot" >nouutohwoot</alias>
    <mcp-descriptions alias="nouutohwoo">
     <mb morphemes="nouut-ohwoo" slots="A-B" />
     <morphdesc>
      <morpheme key="1" arg=":start" head="B" pos="vec.elem" slot="A">nouut</token>
      <morpheme key="1" arg="" head="" pos="stem.root" slot="B">ohwoo</token>
     </morphdesc>
    </mcp-descriptions>
    <mcp-descriptions alias="nouutohwoot">
     <mb morphemes="nouut-ohwoo" slots="A-B" />
     <morphdesc>
      <morpheme key="1" arg=":start" head="B" pos="vec.elem" slot="A">nouut</token>
      <morpheme key="1" arg="" head="" pos="stem.root" slot="B">ohwoo</token>
      <morpheme key="1" arg="" head="" pos="stem.root" slot="B">ohwoo</token>
     </morphdesc>
    </mcp-descriptions>
   </aliases>
```

Noun Incorporation in Verb Stems

ciito’ohnii-	**ciit**-**o’ohn**-**ii**-	‘put on **shoes**’
neeto’ohnii-	**neet**-**o’ohn**-**ii**-	‘take off **shoes**’
ciitotoohee-	**ciit**-**otooh**-**ee**-	‘put on **pants**’
neetotoohee-	**neet**-**otooh**-**ee**-	‘take off **pants**’

- Morphologically decomposable, not lexically decomposable

Issues:

- There’s no way to create a roleset just for ‘put on’ or ‘take off’ with without the incorporated noun
- No good linguistic lexicon would split theses stems up morphologically
- Accounting for these elements even more critical here, as they are core event participants, not oblique
    - **Complication:**
        - Can’t just add general UMR concept stub nodes for these-- need specific referent concept.
        - But, incorporated form doesn’t always match standalone form-- what to add?

Noun Incorporation inside Verb Stems

Roleset Solutions:

- Frame file clusters rolesets around the predicating elements, not the incorporated noun
- Predicate lemma captures the predicating element(s) without the noun
- Roleset lemmas *do* include the incorporated noun

Frame File: ***ciit.xml***

ciito'ohnii-01: *put on one's shoes*		ciitotoohee-02: *put on one's pants*	
:ARG0-agent	dresser	:ARG0-agent	dresser
:ARG1-theme	shoes	:ARG1-theme	pants

Frame File: ***neet.xml***

neeto'ohnii-01: *take off one's shoes*		neetotoohee-02: *take off one's pants*	
:ARG0-agent	undresser	:ARG0-agent	undresser
:ARG1-theme	shoes	:ARG1-theme	pants

Frame File: ***hoxesiini.xml***

hoxesiini-01: *be dusty, dirty*
:ARG1-theme dusty thing

Graph Solution:

- Add node for the noun
    - Allows coreference tracking at document level

nihneeto'ohniit.
nih- neeto'ohnii -t
PAST- take.off.ones.shoes -3S
PREFIX- vai.incorp -INFL
'He took off his shoes.'

```
(s1c / ciito'ohnii-01
   :ARG0 (s1p / person
      :refer-person 3rd
      :refer-number Singular)
   :ARG1 (s1w / wo'oh
      :refer-number Plural)
   :ASPECT Performance
   :MODSTR FullAff)
```

nihhoxesiini3i'.
nih- hoxesiini -3i'
PAST- be.dusty -3PL
PREFIX- vii -INFL
'They were dusty.'

```
(s2h / hoxesiini-01
   :ARG1 (s2t / thing
      :refer-number Plural)
   :ASPECT State
   :MODSTR FullAff)
(s2 / sentence
   :coref ((s1w :same-entity s2t)
```

Up Next:

Roadmap of Phases:

1. **Roleset Design: deep but narrow**
   a. Deep investigation into the grammar and ~600 sentence of Arapaho text

2. **Automatization: shallow but broad**
   a. Use existing lexical database to automatically generate as many rough rolesets as possible

3. **Refining, refining, refining**
   a. Continued fleshing out of rolesets in tandem with annotation

4. **Final resource: deep and broad**