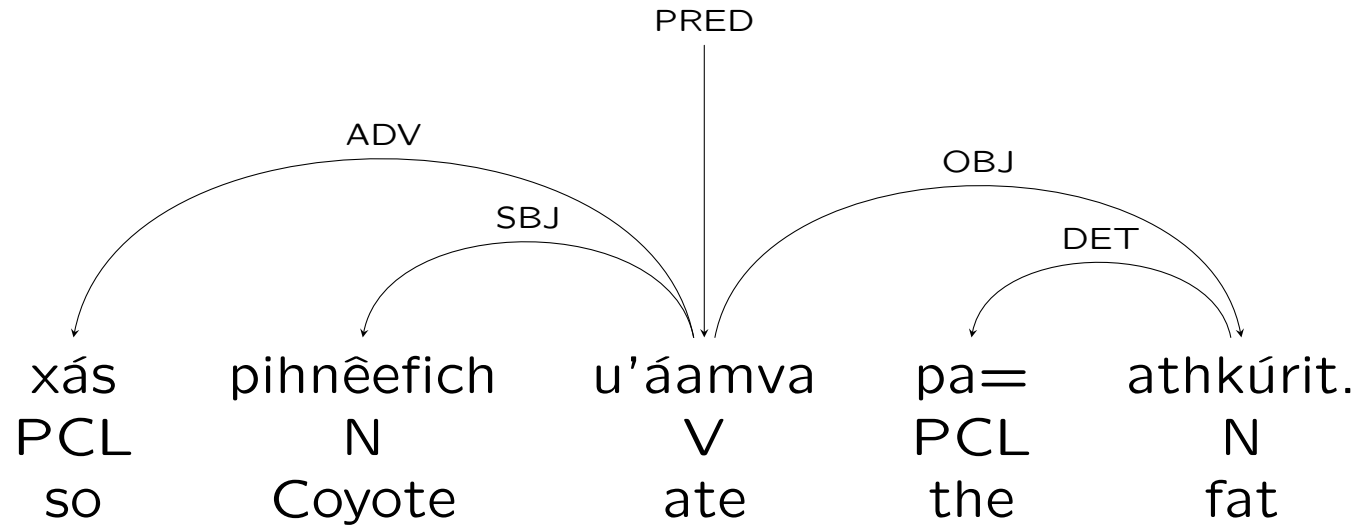

PRED
ADV
SBJ
OBJ
DET
xás
PCL
so
pihnêefich
N
Coyote
u'áamva
V
ate
pa=
PCL
the
athkúrit.
N
fat

Developing the Karuk Treebank

Andrew Garrett, Clare Sandy,
Erik Maier, Line Mikkelsen, Patrick Davidson

Fieldwork Forum, November 13, 2013, UC Berkeley

1. What is a treebank?

2. What is Karuk?

3. Why a Karuk treebank?

4. Analytic approach and annotation scheme

5. Process and implementation

6. Acknowledgements

What is a treebank?

- a syntactically annotated corpus; typically electronic and large
  - tree = syntactic tree diagram
  - bank = repository
- annotation is exhaustive, systematic, and single-parse

Treebank uses

- gold standard for parsing (e.g. Penn Treebank for English)

- human language technologies, automatic content extraction, cross-lingual information retrieval, information detection, . . .

- **empirical basis for linguistic analytic research**

  – search and extract syntactic patterns

Existing treebanks

- 80 treebanks for 38 languages
  - e.g. Classical Arabic, Catalan, Czech, English (16), Estonian, Ancient Greek, Hebrew, Hindi, Icelandic, Japanese, Korean, Persian, Russian, Spanish, Thai, Turkish, Urdu, Vietnamese.
- size (# of sentences): 1,216 (Hebrew) to 69,537 (English)
- analytic approaches:
  - Phrase Structure Grammar, **Dependency Grammar**
  - HPSG, Case grammar, Combinatory Categorial Grammar

The Karuk language

- traditionally spoken along the middle course of the Klamath river in northern California; isolate within Hokan group

- 1800-2700 speakers at contact; currently around 6-12 first-language speakers and 20-50 second-language speakers

- non-configurational syntax; [New/Contrast V Old]
  - **free argument order**
  - **free argument omission**
  - **free argument split**

Why a Karuk treebank?

1. to improve understanding of Karuk syntax

   - Bright (1957:119-142): valuable but incomplete structuralist description

   - Macaulay (1989–2010): targeted analyses of clitics, agreement, alignment, applicatives, evidentials, and the discourse particle *káruma*

   - much of Karuk syntax remains uncharted terrain, including clausal organization, word order, NP-internal structure, subordination, anaphora and cataphora, non-verbal predication, valence and valence changing processes, quantification, modification, coordination . . .

2. critical for effective language teaching and revitalization

3. extensive linguistic fieldwork (1853 to present)
   $\rightarrow$ substantial text corpus (est. $>$ 10,000 sentences)

   - large enough to generate reliable information about Karuk syntax

   - too large to effectively extract this information by examining raw texts

4. harness existing lexical and morphological annotation (4,800 sentences annotated)

5. proof of concept: first treebank of endangered language without premodern written history

Analytic approach

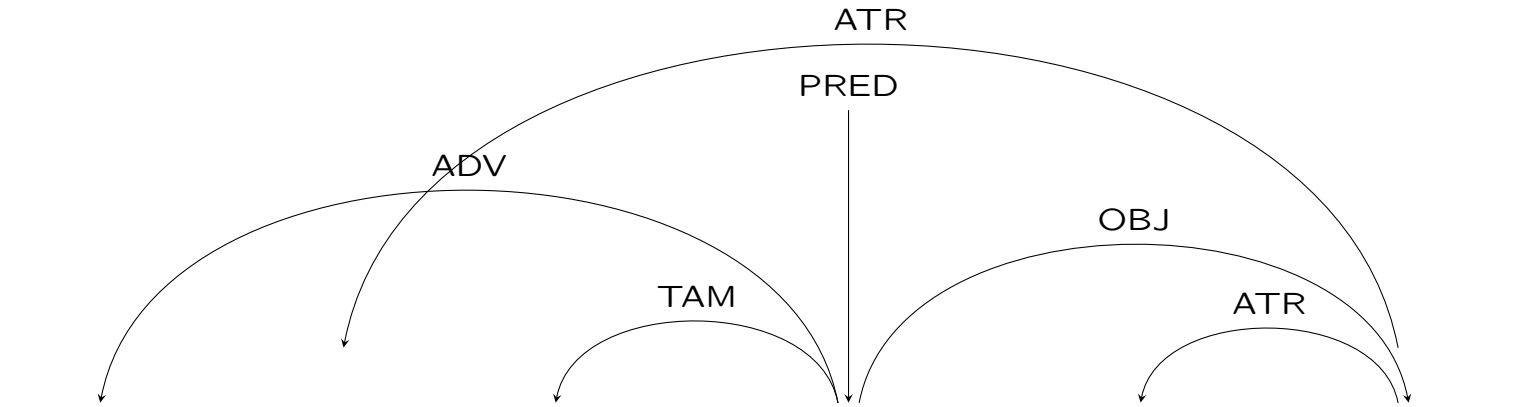
- Catch 22 of Karuk Treebanking
  - to understand Karuk syntax we need to build the treebank
  - to build the treebank we need to understand Karuk syntax

- what to do?
  - adopt an analytic framework that is **surface oriented** and **flexible**
  - develop analytic principles to **guide** annotation and **adjudicate** between competing analyses

Dependency Grammar

- central notion is dependency relation between head and dependent, e.g. a verb and its object

- what makes it different from Phrase Structure Grammar:
  - no intermediate nodes; relations hold only between syntactic words
  - "constituents" need not be contiguous
  - no null elements



So they dug a lot of brodiaeas.

- **surface oriented**:
  - discontinuous constituents can be coded without commitment to an underlying position or structure
  - omitted arguments can be left uncoded
- **flexible**: dependency is independent of word order

Annotation Scheme

Each word is annotated for three attributes:

- head: "the word that I am a dependent of"
- syntactic relation: "the syntactic relation I bear to that head"
- part of speech: "my part of speech" (from dictionary)

Word order is coded implicitly in the xml document.

Annotation guidelines

1. Headedness:

   (a) the predicate is the head of the clause

   (b) the noun is the head of NP

   (c) the postposition is the head of PP

2. Inventory of 15 syntactic relations (e.g. SBJ, OBJ, POSS) and when to use them.

3. Dossier of specific constructions (e.g. clefts, comparatives, gapping, possessive clauses, vocatives) and how to annotate them.

Adjudication Principles

1. Be faithful to Karuk.

2. Don't exoticize.

3. Attach high.

4. Same string gets same annotation.

5. Be faithful to translation.

Process

- **Phase I** (2011): deciding on analytic approach and identifying texts for 500-sentence pilot corpus.

- **Phase II** (2012): morphological annotation of pilot corpus, first syntactic annotation of pilot corpus, development of general analytic principles.

- **Phase III** (2013): second annotation of pilot corpus, development of annotation guidelines, third annotation of pilot corpus, xml mark-up, development of annotator interface.

- **Phase IV** (2014–): large-scale annotation w. automatic xml encoding, development of search interface.

Implementation

Goal is treebank of 10,000 sentences $\rightarrow$ team effort!

- Linguistic expertise
  - Karuk Study Group (volunteer-based)
  - Research assistants (paid)
  - URAP (units)
  - Ling 170 and Ling 175 (group assignments and final projects)
- Funding: NSF-DEL grant November 2011–November 2014

- Computational expertise: Patrick Davidson and Ronald Sprouse
  - best xml practices
  - version control
  - system administration
  - software development

$\rightarrow$ **The Dependency Grammar Graph Builder and Viewer!**

Yôotva

- Karuk elders, teachers and activists: Lucille Albers, Tamara Alexander, Sonny Davis, Susan Gehr, Julian Lang, Crystal Richardson, Nancy Richardson, Bud Smith, Vina Smith, Florrine Super, Arch Super, and Charlie Thom, Sr.

- NSF (award #1065620 Karuk [kyh] and Yurok [yur] syntax and text documentation); The Undergraduate Research Apprentice Program.

- UC Berkeley students: Nico Baier, Shane Bilowitz, Kayla Carpenter, Anna Currey, Erin Donnelly, Kouros Falati, Matt Faytak, Nina Gliozzo, Morgan Jacobs, Erik Maier, Karie Moorman, Olga Pipko, Melanie Redeye, Clare Sandy, Jeff Spingeld, Tammy Stark, Whitney White and 11 other students in Linguistics 170, Spring 2012.

- IT gurus: Patrick Davidson and Ronald Sprouse.

- Jonathan Kummerfeld for assistance with treebank information

Syntactic relations used in the Karuk Treebank

PRED (Predicate): The label given to the predicate of a main clause.

TAM (Tense/aspect/mood): The dependency relation between a predicate and a tense, aspect or mood particle.

NEG (Negation):The dependency relation between a predicate and a negative particle.

SBJ (Subject): The dependency relation between a predicate and its subject.

OBJ (Direct object): The dependency relation between a predicate and its direct object.

IO (Indirect object): The dependency relation between a predicate and its indirect object

POSS (Possessor): The dependency relation between a noun and its possessor.

ATR (Attribute): The dependency relation between a noun and a modifier or determiner.

COMP (Complement): The dependency relation between a head and its complement.

Syntactic relations used in the Karuk Treebank (cont.)

QUOT (Quotative complement): The dependency relation between a quote-introducing predicate and the head of its quote.

ADV (Adverbial): The dependency relation between a predicate and a modifying adverbial.

SPRED (Secondary predicate): The dependency relation between a predicate and a resultative or depictive predicate.

APOS (Apposition): The dependency relation between two appositional elements.

SUB (Subordinator): The dependency relation between a predicate and its subordinator pa=

COORD (Coordination): The dependency relation between two coordinated elements.