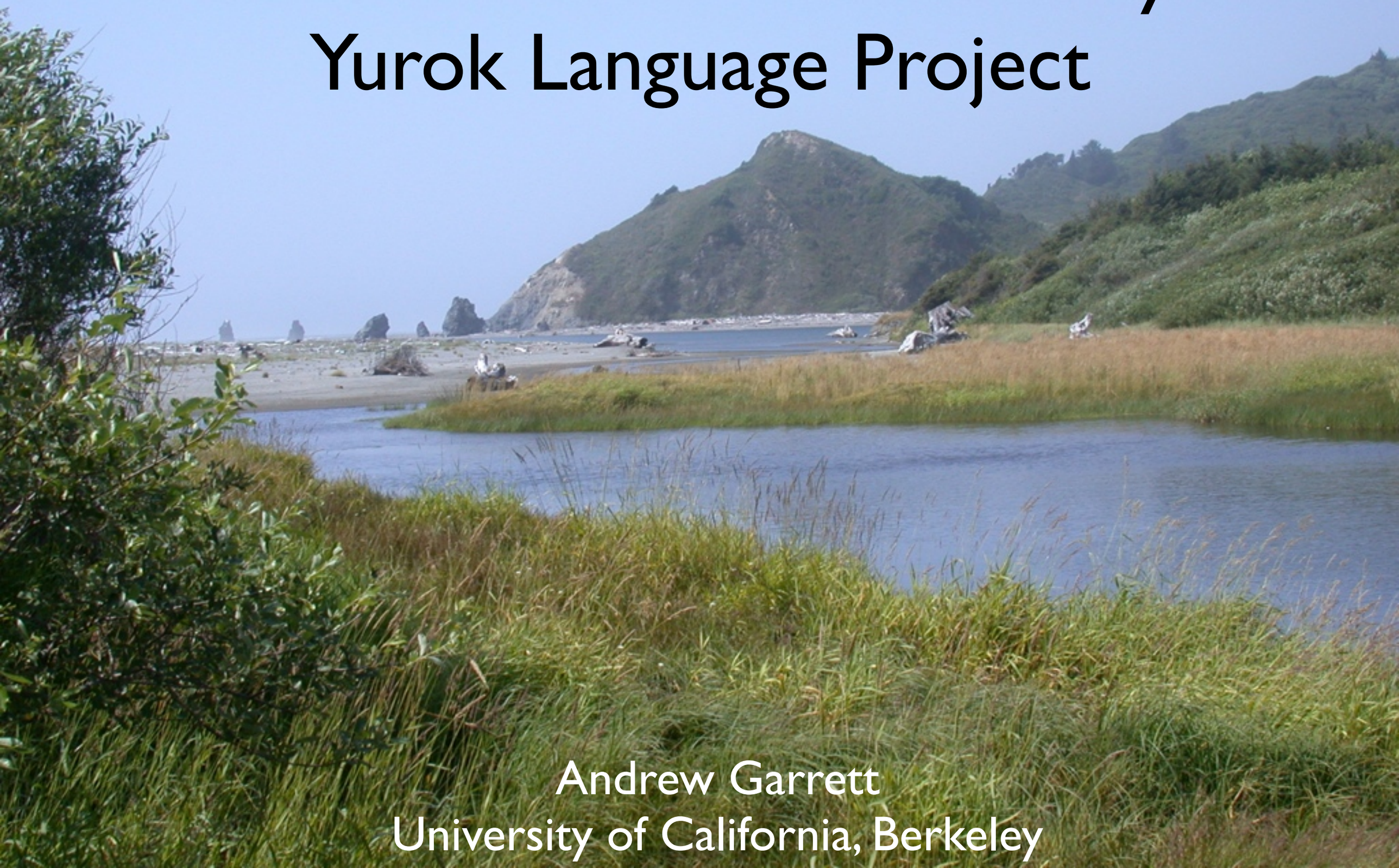
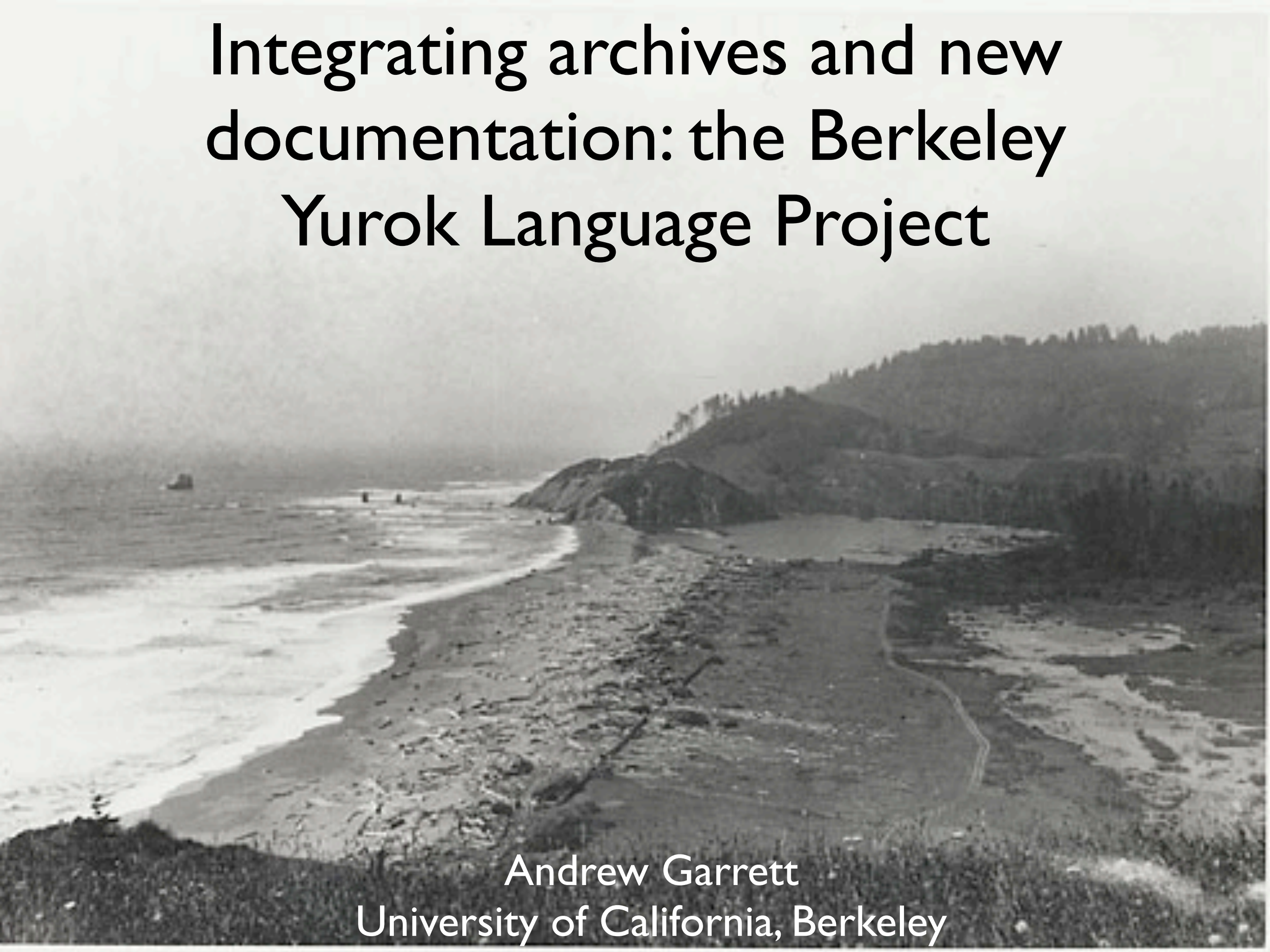
Integrating archives and new documentation: the Berkeley Yurok Language Project



Andrew Garrett
University of California, Berkeley

Integrating archives and new documentation: the Berkeley Yurok Language Project

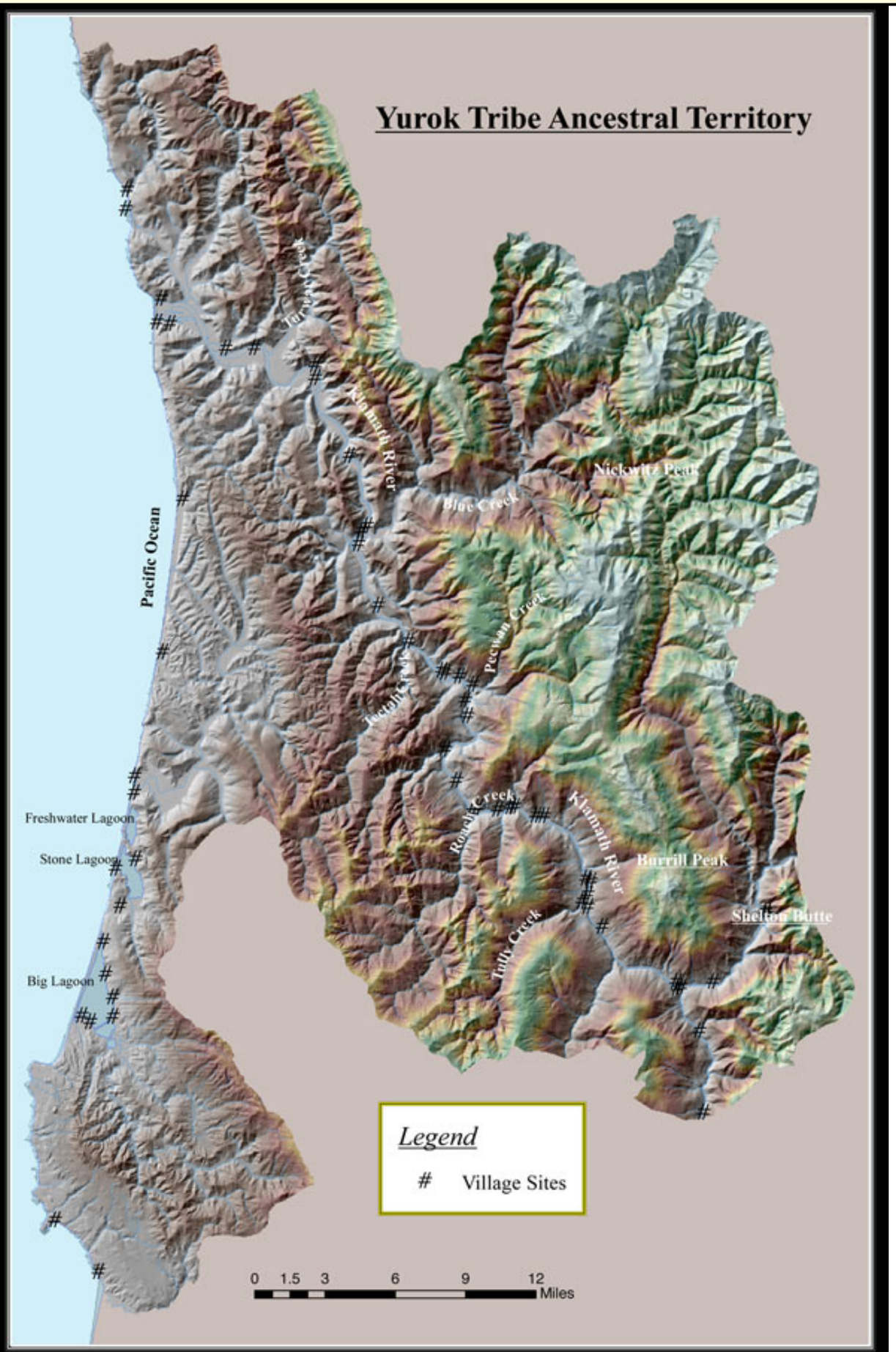


Andrew Garrett
University of California, Berkeley

Plan of this presentation

- Background
- Database structure
- Dynamic online tools

Background



Northwest California (yuroktribe.org)

Yurok (Algic, NW California)

- About 2000-3000 speakers before white contact (indigenous California was linguistically complex, with many relatively small languages)
- Very few (< 6) fluent speakers today, all elderly
- Generally open attitude toward sharing language
- Active language program with basic instruction in all schools; several years in one local high school
- Reasonably good computer infrastructure, with computers accessible in some language learning settings (but not in the following slide)

Yurok elder & language teacher Jimmie James (bottom right), at a summer language camp

Methodological postulates

- A documentary *corpus* lies at the heart of many kinds of project.
- Such a corpus is available for academic research, for community or tribal research, and for planning language pedagogy.
- Primary data should be accessible if possible.

  Primary data = data that are not derivative of other data, including some that Himmelmann 2009 would call "raw" (recordings), some that he would call "primary" (notes)

Database structure

Yurok elder & language teacher
'aawokw Glenn Moore Sr.

Elements of the database

- Audio recordings
- Texts
- Dictionary
- Lame ontology
- Photographs (not discussed today)
- Shoebox-reminiscent XML format

Audio recordings

- Primary recordings: field recordings
  - 87 texts recorded from 1902 through 2008: narratives ("myths"), ceremonies, procedures, usufruct, local history, anecdotes, conversations, etc.
  - Linguistic elicitation sessions from 1933, 1951, 1962, and from 1980 through 2008: over 200 hours total
- Secondary recordings: recordings of about 3800 words, selected from field recordings, together with metadata

Secondary recording metadata

- A fragment of audio.xml (lemmatized!)

```
<item>
    <url>http://linguistics.berkeley.edu/~yurok/Words/AileenFigueroa/nepuy.mp3</url>
    <word>nepuy</word>
    <tr>salmon</tr>
    <lx id="2140"/>
    <speaker id="AF">Aileen Figueroa</speaker>
</item>
```

- Though audio.xml contains about 3800 <item>s, thousands remain to be added.

Texts

- 130 *edited* texts
  - Some from audio, others recorded in field notes only
  - About 5500 sentences
  - About 26000 words in all: lemmatized!
  - "Texts" include self-contained narratives, etc., but an elicitation session is also classified as a "text"
- Metadata for 110 *unedited* texts
  - Some from audio, others recorded in field notes only

Metadata for an unedited text

```
<text>
   <metadata>
      <ref>X14y</ref>
      <spkr>X</spkr>
      <author>Captain <alph>Spott</alph></author>
      <title>"The Mouth of the River"</title>
      <year>1907</year>
      <collector>A. L. Kroeber</collector>
      <transcript>A. L. Kroeber, Yurok field notebook 81</transcript>
      <translation>A. L. Kroeber, <i>Yurok Myths</i> (1976), 430-433 (myth X14y)</
translation>
      <audio-yurok-ref>24-1029</audio-yurok-ref>
      <status>unedited</status>
      <genre>myth</genre>
   </metadata>
</text>
```

Text data: Transcribed elicitation

```
<s>
   <parsetx>
      <word id="485">Hikoch</word>
      <word id="448">hes</word>
      <word id="4300">'o</word>
      <word id="1983">myah</word>
      <word id="1124">ku</word>
      <word id="1365">'we-le'loyhl</word>
      <word id="2441">pa'aahl</word>?</parsetx>
   <tx>Hikoch hes 'o myah ku 'we-le'loyhl pa'aahl?</tx>
   <tr>Did the fire jump across the water?</tr>
   <audio>
            <path>PublicRecordings/MP3/LC/</path>
            <filename>LC-01-1_023.mp3</filename>
   </audio>
   <start-time>9:20</start-time>
</s>
```

Lexicon

- About 4500 lemmas, including this one:

```
<lxGroup id="1749">
   <lx>mewihl</lx>
   <ps>n</ps>
   <ge>elk</ge>
   <sci>Cervus elaphus</sci>
   <rf>WEM264<spkr>WG</spkr></rf>
   <rf>R222</rf>
   <rf>JE48</rf>
   <sd>45</sd>
   <photo>mewihl.jpg</photo>
</lxGroup>
```

- But “elk” is relatively simple.

```
<lxGroup id="1543">
    <lx>hloykok'</lx>
    <ps>vt oo-class</ps>
    <ge>I try</ge>
    <rf>R219</rf>
    <rf>LA138-011<spkr>FS</spkr></rf>
    <rf>JE139</rf>
    <pdGroup>
        <pd>3sg</pd>
        <pdf>hloyko'm</pdf>
        <rf>I4</rf>
    </pdGroup>
    <pdGroup>
        <pd>imperative sg</pd>
        <pdf>hloo'yk'os</pdf>
        <rf>R219</rf>
    </pdGroup>
    <mr>
        <m>38</m>
        <m variant="1">14</m>
        <m>4</m>
    </mr>
</lxGroup>
```

Dynamic online tools

Yurok elders and language teachers
Archie Thompson & *'aawokw* Aileen Figueroa

What happens

- You type search terms into our web pages
- Request goes to Berkeley Linguistics server, specifically to axkit

  axkit.org: "Apache AxKit is an XML Application Server for Apache. It provides on-the-fly conversion from XML to any format, such as HTML ...."

- axkit transmits request to an appropriate XSL document, which hunts around in the databases and serves up a web page as per its programming

Demo

— During the oral presentation these capabilities were shown online. —

- Looking up words in the dictionary
- Ontological and morphological travel
- Reading texts
- Searching in texts
- Listening to audio
- Random words (beta)

Wokhlew!

- Our website: linguistics.berkeley.edu/~yurok
  (or google "yurok language")

- This talk (or google "andrew garrett talks"):
  linguistics.berkeley.edu/~garrett/Hawaii2009.pdf