A Web-Accessible Dictionary of Southeastern Pomo

Yao Yao, Charles B. Chang, Shira Katseff,
Russell Lee-Goldman, and Marta Piqueras-Brunet[*]

University of California, Berkeley
1203 Dwinelle Hall, Berkeley, CA 94720-2650, USA
yaoyao@berkeley.edu, cbchang@berkeley.edu, skatseff@berkeley.edu,
rleegold@berkeley.edu, marta.piqs@gmail.com

Abstract

In recent years, building web-accessible dictionaries has become a new way of organizing and publishing data from linguistic fieldwork on indigenous languages. This paper provides an overview of one such online dictionary. The language we worked on was Southeastern Pomo, an endangered native American language which was historically spoken in the area of Clear Lake, California. The structure of the dictionary and the files that are needed to construct it are introduced in detail. We also demonstrate three types of searches that are made available on the dictionary website.

Keywords: online dictionary, language documentation, native American languages.

1. Background

In the past decade, as resources for linguistic analysis have entered cyberspace, so have the products of linguistic documentation. Online dictionaries have been created for a number of languages native to the Americas, including Yurok, Hupa, Washo, and Northern Paiute (cf. Babel et al. 2006, Dick and Haynes 2006). These online dictionaries make the information gathered through fieldwork accessible to researchers, tribe members, and in particular, young learners. In this paper, we present our work on an online dictionary for another native American language, Southeastern Pomo, with the goal of providing an overview for the linguistic fieldwork community of the technical details involved in building an online dictionary website.

[*] For their help, feedback, and support throughout the course of this project, we are grateful to Jocelyn Ahlers, Zhenya Antić, Catherine Callaghan, Thera Crane, Donna Fenton, Andrew Garrett, Robert Geary, Hannah Haynie, Leanne Hinton, Jisup Hong, Loretta Kelsey, Julius Moshinsky, Lindsey Newbold, Ronald Sprouse, Maziar Toosarvandani, Corey Yoquelet, the audience at CIL 18, and the Department of Linguistics at the University of California, Berkeley. Any errors remain our own.

1.1. The Southeastern Pomo dictionary project

Southeastern Pomo (Northern Hokan, Pomoan) is an acutely endangered language historically spoken in the area around Clear Lake, California (Moshinsky 1974, Gordon 2005, cf. Figure 1).

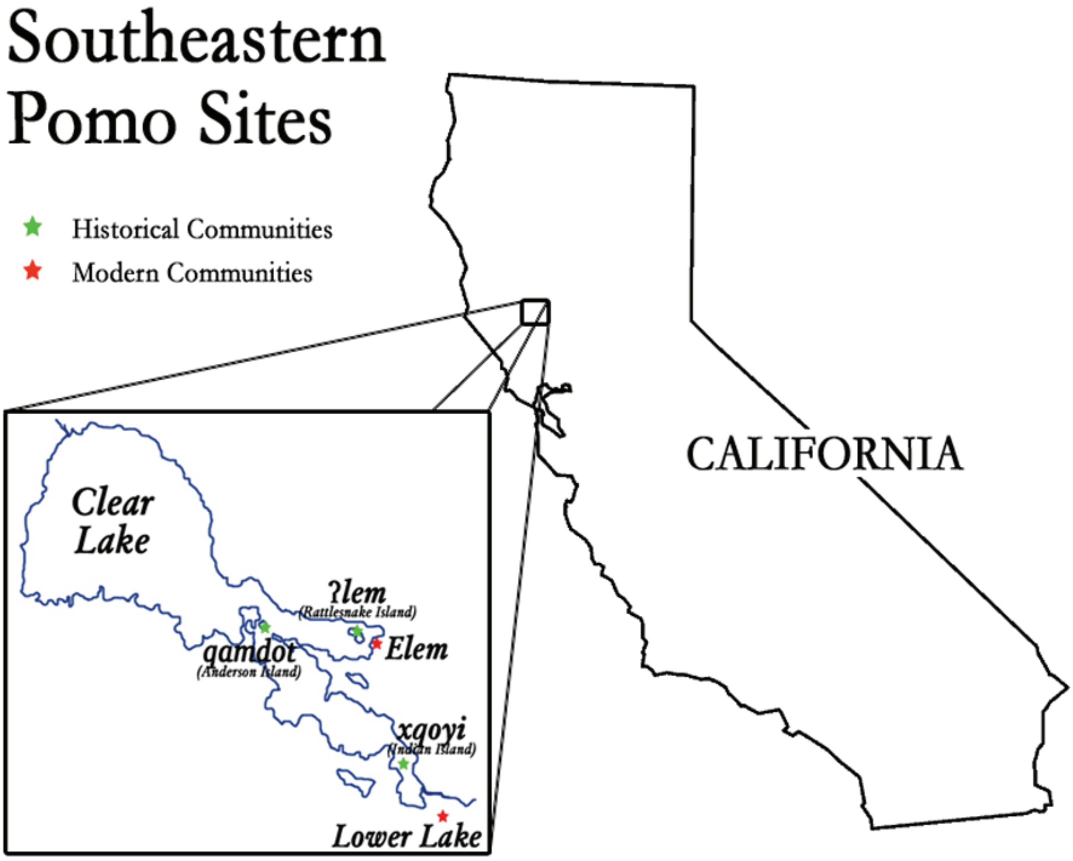


Figure 1. Locations of the Southeastern Pomo speech communities (from Haynie 2007)

Speakers and learners are mostly affiliated with the Elem Pomo Tribe. Today only two fluent speakers remain. However, revitalization efforts are underway, led by Loretta Kelsey, one of the two remaining speakers, and Robert Geary (cf. Shavelson 2006, Fagan 2007). Through collaboration with linguists, they have developed a community orthography, teaching materials, and a print dictionary with pictures of local flora and fauna. Language camps have also been organized for people to learn the language.

The present online dictionary is another component of this revitalization work. We started working with Kelsey and Geary in the fall of 2006, and have so far elicited hundreds of hours of data comprising word lists, sentences, and texts. In order to organize and publish the data in the most efficient and user-friendly way, we constructed a web-accessible dictionary. The dictionary has three main functions: searching for words, searching for audio files, and searching for sentences/texts.

1.2. Transcribing the sounds of Southeastern Pomo

Southeastern Pomo (henceforth, SEP) has a rich inventory of consonants covering a wide range of places of articulation, including labial, dental, alveolar, velar, post-velar, and glottal. In addition to a voiced/voiceless laryngeal distinction, the language also has a series of ejectives (both in stops and affricates). Table 1 below shows the inventory of consonant phones, as well as our practical ASCII orthography (which differs in some respects from the community orthography) in parentheses underneath the IPA forms.

	LABIAL	DENTAL	ALVEOLAR	PALATAL	VELAR	POST-VELAR	GLOTTAL
STOP	p p’ b (p p’ b)	t̪ t̪’ (th th’)	t t’ d (t t’ d)		k k’ (k k’)	q q’ (q q’)	ʔ (7)
AFFRICATE			ts ts’ (ts ts’)	tʃ tʃ’ (ch ch’)			
FRICATIVE	f (f)		s (s)	ʃ (sh)	x (x)	χ (X)	h (h)
NASAL	m (m)		n (n)		ŋ (ng)		
FLAP			ɾ (r)				
APPROXIMANT	w (w)		l (l)	j (y)			

Table 1. Consonant inventory of Southeastern Pomo

The vowel inventory of SEP is much simpler. The language basically has a five-vowel system, with an additional schwa that appears predictably in epenthetic positions (and therefore is not spelled out in our orthography). Table 2 shows the vowel system together with the orthography in parentheses beside the IPA forms.

	FRONT	CENTRAL	BACK
HIGH	ɪ (i)		ʊ (u)
MID	ɛ (e)	ə	o (o)
LOW		a (a)	

Table 2. Vowel inventory of Southeastern Pomo

2. Dictionary Structure

The online dictionary is composed of three subparts: a lexicon, an audio dictionary, and a text archive.

2.1. Lexicon

The lexicon contains entries for individual words, affixes, and fixed expressions. Each entry displays the following information about an item: (1) its SEP transcription, (2) its part of speech, (3) an English gloss, and (4) links to sound clips in the audio dictionary (if available). Different lexicon entries correspond to different forms, but not necessarily different lexemes (e.g. different morphological forms of a verb get separate, though linked, entries).

2.2. Audio dictionary

In addition to the lexicon, there is a separate audio dictionary. Each entry in the audio dictionary corresponds to an audio file clipped out of master field recordings to correspond to a lexicon entry. An entry in the audio dictionary displays the following information about an item: (1) its SEP transcription, (2) an English gloss, (3) the name of the speaker who produced the item, (4) a link to the corresponding lexicon entry, and (5) a reference to its source, i.e. the filename of the master field recording together with the time point within it at which the relevant audio begins (for developers).

The audio dictionary is independent of the lexicon, but corresponding lexical entries and sound clippings are linked to each other from both sides. The two dictionaries are kept apart for the purposes of facilitating two different types of searches – word searches and audio searches (see §4).

2.3. Text archive

The text archive contains entries for elicited sentences, narratives, and other discourse above the sentence level. Each entry displays the following information about an item: (1) the name of the speaker who produced the text, (2) the genre of the text, (3) transcriptions of individual sentences, (4) free translations, and (5) interlinear glosses.

3. Building the Website

In this section, we briefly review the construction of the website, in terms of data preparation and storage, query processing, and results display.

In the preparation stage, fieldwork recordings are manually processed and clippings of individual words, sentences, and texts are saved individually. Meanwhile, information about these data is entered into three separate databases corresponding to the subparts of the dictionary. Lexicon entries are first entered into a mySQL database and then converted into a grand XML (eXtensible Markup Language) file, while information about audio clippings and texts is entered into separate XML files directly.

The mySQL database of lexicon entries has the following fields:

- SEP transcription
- transcriptions of variants (if any)
- community orthography (if available)
- part of speech
- free gloss & interlinear gloss
- semantic domain
- source file & start time
- links to other morphologically related entries
- notes

The mySQL database has a number of desirable features. For one, it automatically generates an unique ID number for each new entry, and is fully sortable and searchable. It also allows several researchers to make and edit entries at the same time without overwriting each other's work. In addition, the database can be easily converted into XML format, which is essentially a text file with all the above information in a fixed format. Figure 2 below shows a sample entry in the lexicon XML file. As shown below, the SEP word *kachuchu* (ID = 101) is transcribed as *kuchechoo* in the community orthography. It is a noun; means 'cap' in English; belongs to the semantic domain of clothes; and is a headword. Furthermore, it was first elicited in the recording named 21sep06_LK1b, at time 18:39.

The information in the audio dictionary and the text archive is stored in XML files directly, in a similar format as the lexicon XML file. Thus, all the information contained in the dictionary is stored in three separate XML files: one for the lexicon, one for the audio files, and one for the texts. Such a text-based file format allows the data to be easily manipulated into other formats as the technology of documentation changes over time.

```
<lemma>
    <id>101</id>
    <lx>kachuchu</lx>
    <community_orthography>kuchechoo</community_orthography>
    <ps>n</ps>
    <ge>cap</ge>
    <short-gloss>cap</short-gloss>
    <ref>21sep06_LK1b</ref>
    <time>18:39</time>
    <sd>clothes</sd>
    <is-headword>yes</is-headword>
</lemma>
```

Figure 2. Sample XML for a lexicon entry

Separate XSL (eXtensible Stylesheet Language) files are written to process search queries and control the display of the results. The function of the XSL files is essentially to decide what information from the XML files needs to be retrieved in response to a certain query and how the retrieved information should be formatted on screen. It should be noted that by having separate XML and XSL files, we are able to separate data processing from data storage, which prevents accidental data overwriting and also allows more flexibility in data processing and display.

To sum up, all relevant information in the dictionary is saved in XML files, and the processing is taken care of by XSL stylesheets. Figure 3 illustrates the data flow in the processing of an incoming query.

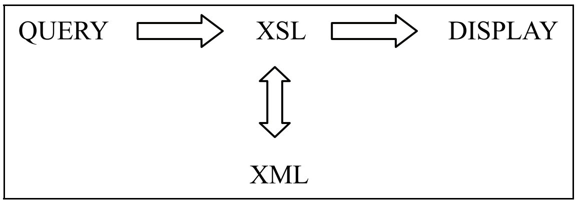


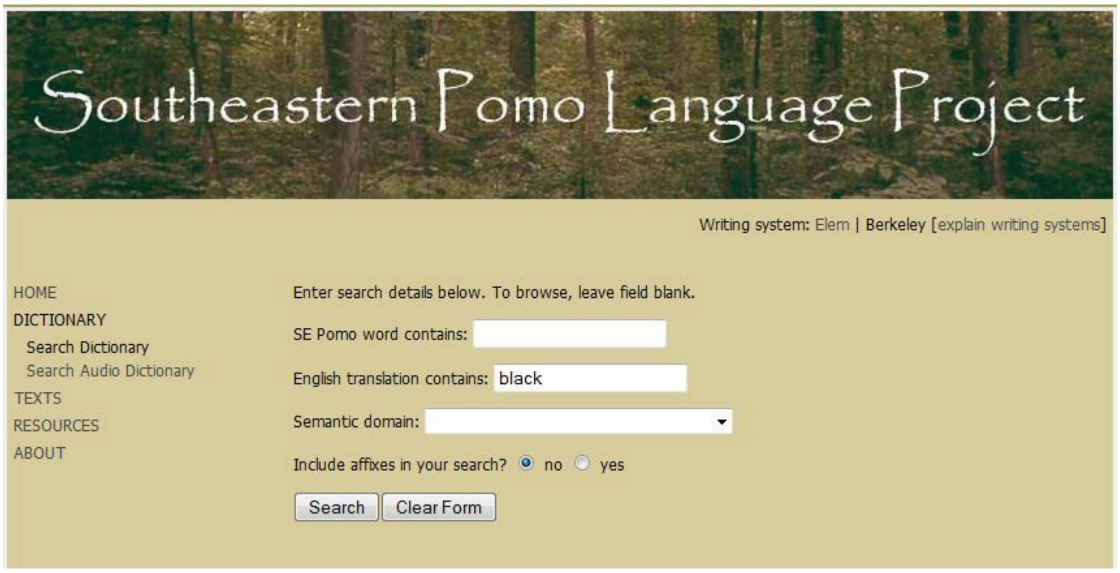
Figure 3. Illustration of query processing

4. Dictionary Searches

In this section, we demonstrate three types of searches handled on the dictionary website: word search, audio search, and text search. As mentioned above, the three sub-dictionaries are stored separately, but cross-linking between any two is possible.

4.1. Word search

Lexicon entries can be searched using up to three search criteria: SEP transcription, English gloss, and semantic domain. Users can also choose whether affixes are to be included in the search. Figure 4 demonstrates a search for lexicon entries for which the English glosses contain the word *black*.



Southeastern Pomo Language Project

Writing system: Elem | Berkeley [explain writing systems]

HOME
DICTIONARY
Search Dictionary
Search Audio Dictionary
TEXTS
RESOURCES
ABOUT

Enter search details below. To browse, leave field blank.

SE Pomo word contains:

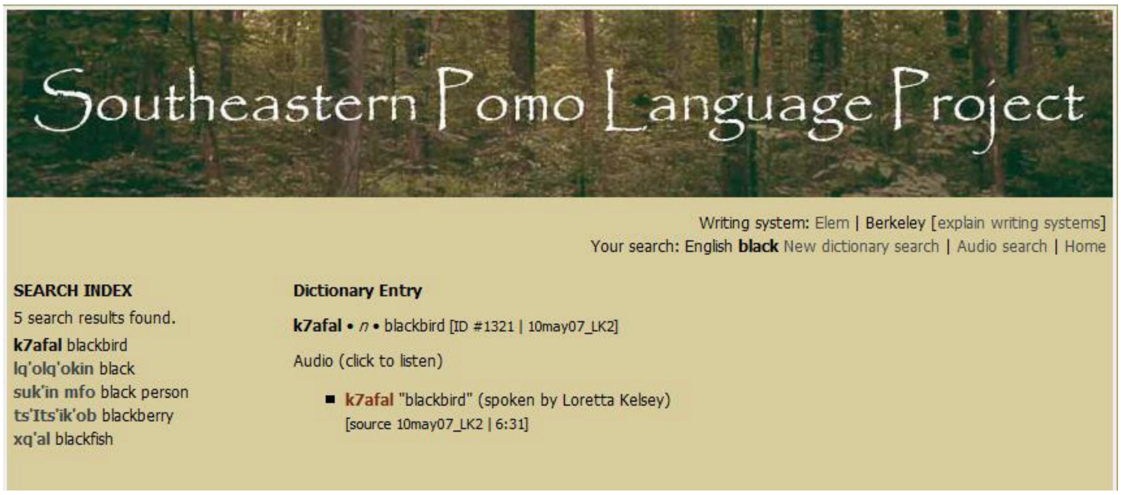
English translation contains: black

Semantic domain:

Include affixes in your search? ● no ○ yes

Search | Clear Form

Figure 4. Web interface for a sample lexicon search



Southeastern Pomo Language Project

Writing system: Elem | Berkeley [explain writing systems]
Your search: English **black** New dictionary search | Audio search | Home

SEARCH INDEX
5 search results found.
k7afal blackbird
lq'olq'okin black
suk'in mfo black person
ts'Its'ik'ob blackberry
xq'al blackfish

Dictionary Entry

k7afal • *n* • blackbird [ID #1321 | 10may07_LK2]

Audio (click to listen)

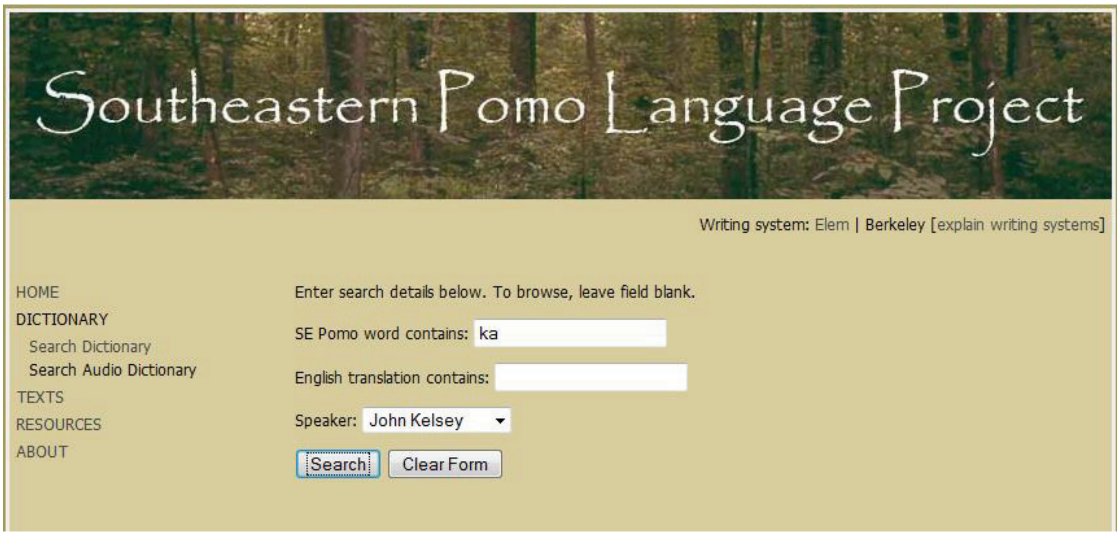
- k7afal "blackbird" (spoken by Loretta Kelsey)
[source 10may07_LK2 | 6:31]

Figure 5. Results of a search for words with English glosses containing *black* (entry *k7afal* selected)

Five lexicon entries are returned by the query, and their SEP transcriptions and English glosses are shown in the left pane of the results page (cf. Figure 5 above). Upon clicking on an individual entry in this list, say *k7afal* ‘blackbird’, the full entry for the word and a link to the corresponding audio file will appear in the main pane on the right.

4.2. Audio search

Audio clippings can be searched by SEP transcription, English gloss, speaker, or any combination of the three.


Southeastern Pomo Language Project
Writing system: Elem | Berkeley [explain writing systems]
HOME
DICTIONARY
Search Dictionary
Search Audio Dictionary
TEXTS
RESOURCES
ABOUT
Search
Clear Form

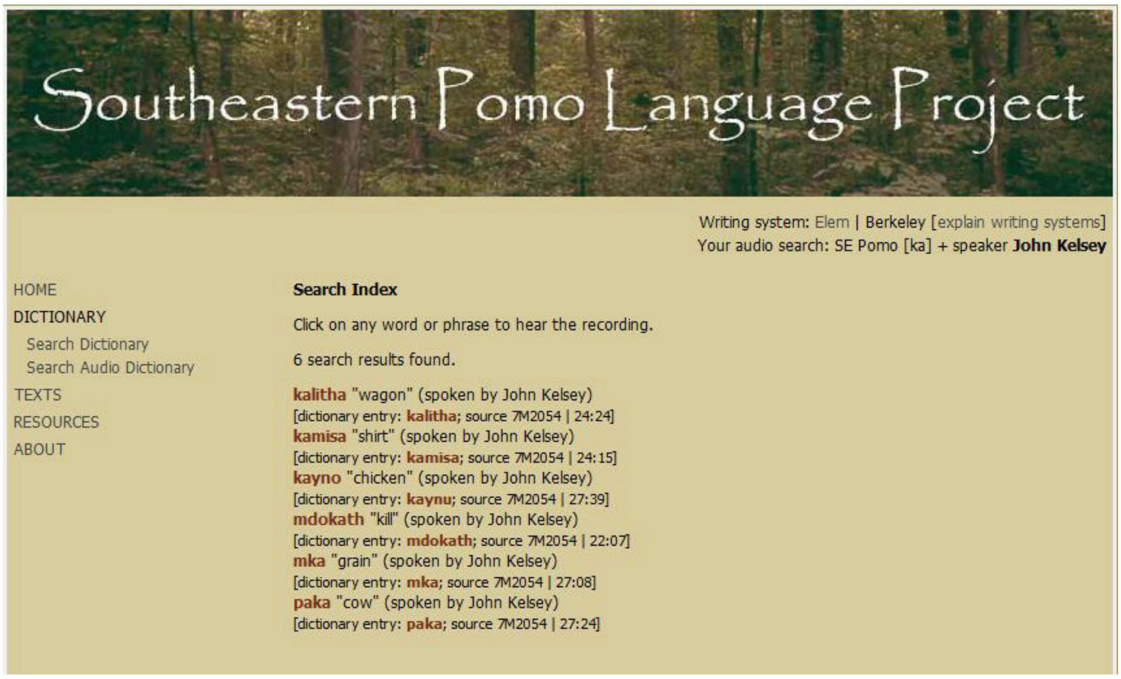
Enter search details below. To browse, leave field blank.

SE Pomo word contains: ka

English translation contains:

Speaker: John Kelsey

Figure 6. Web interface for a sample audio search


Southeastern Pomo Language Project
Writing system: Elem | Berkeley [explain writing systems]
HOME
DICTIONARY
Search Dictionary
Search Audio Dictionary
TEXTS
RESOURCES
ABOUT

Your audio search: SE Pomo [ka] + speaker **John Kelsey**

Search Index

Click on any word or phrase to hear the recording.

6 search results found.

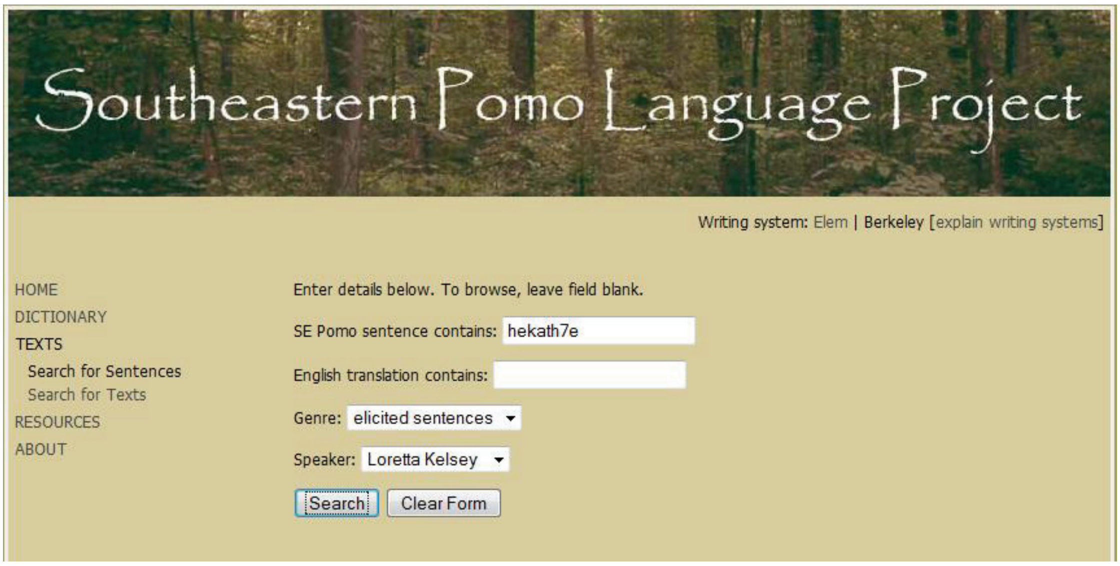
kalitha "wagon" (spoken by John Kelsey)
[dictionary entry: **kalitha**; source 7M2054 | 24:24]
kamisa "shirt" (spoken by John Kelsey)
[dictionary entry: **kamisa**; source 7M2054 | 24:15]
kayno "chicken" (spoken by John Kelsey)
[dictionary entry: **kaynu**; source 7M2054 | 27:39]
mdokath "kill" (spoken by John Kelsey)
[dictionary entry: **mdokath**; source 7M2054 | 22:07]
mka "grain" (spoken by John Kelsey)
[dictionary entry: **mka**; source 7M2054 | 27:08]
paka "cow" (spoken by John Kelsey)
[dictionary entry: **paka**; source 7M2054 | 27:24]

Figure 7. Results of a search for words with the segment sequence -*ka*- spoken by speaker John Kelsey

As a demonstration, if the user searches for all sound clippings of words with *-ka-* in their transcription that are spoken by John Kelsey (cf. Figure 6), six search results will be found and displayed (cf. Figure 7). Each search result can be clicked to play the audio, and a link is provided below the audio entry to the corresponding lexicon entry.

4.3. Text search

Finally, users can also search for texts using multiple search criteria.



Writing system: Elem | Berkeley [explain writing systems]

Enter details below. To browse, leave field blank.

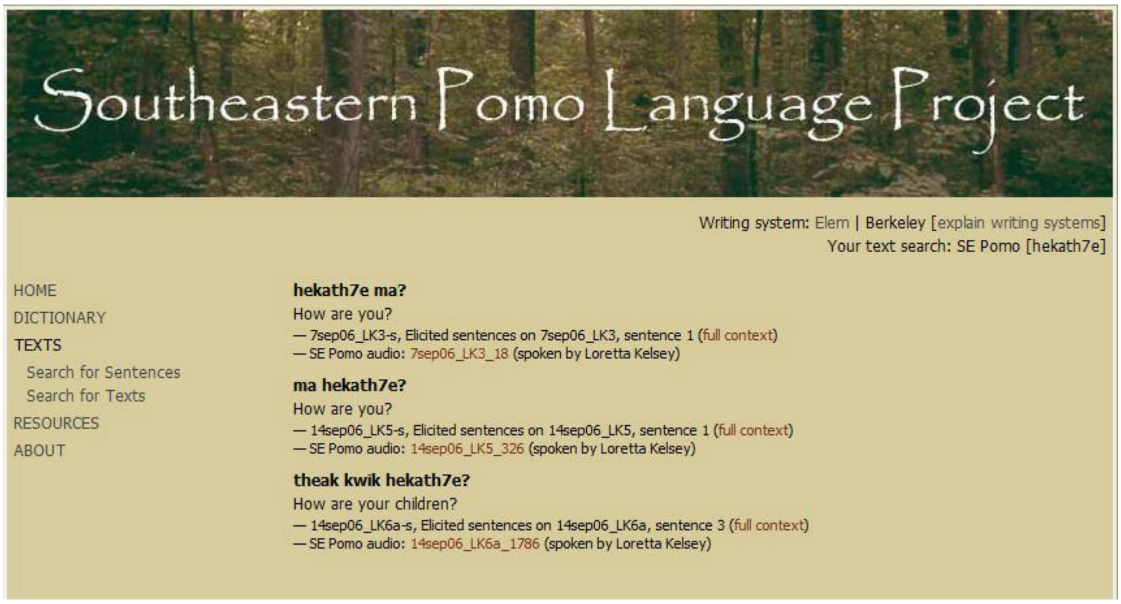
SE Pomo sentence contains: hekath7e

English translation contains:

Genre: elicited sentences

Speaker: Loretta Kelsey

Figure 8. Web interface for a sample text search



Writing system: Elem | Berkeley [explain writing systems]
Your text search: SE Pomo [hekath7e]

hekath7e ma?
How are you?
— 7sep06_LK3-s, Elicited sentences on 7sep06_LK3, sentence 1 (full context)
— SE Pomo audio: 7sep06_LK3_18 (spoken by Loretta Kelsey)

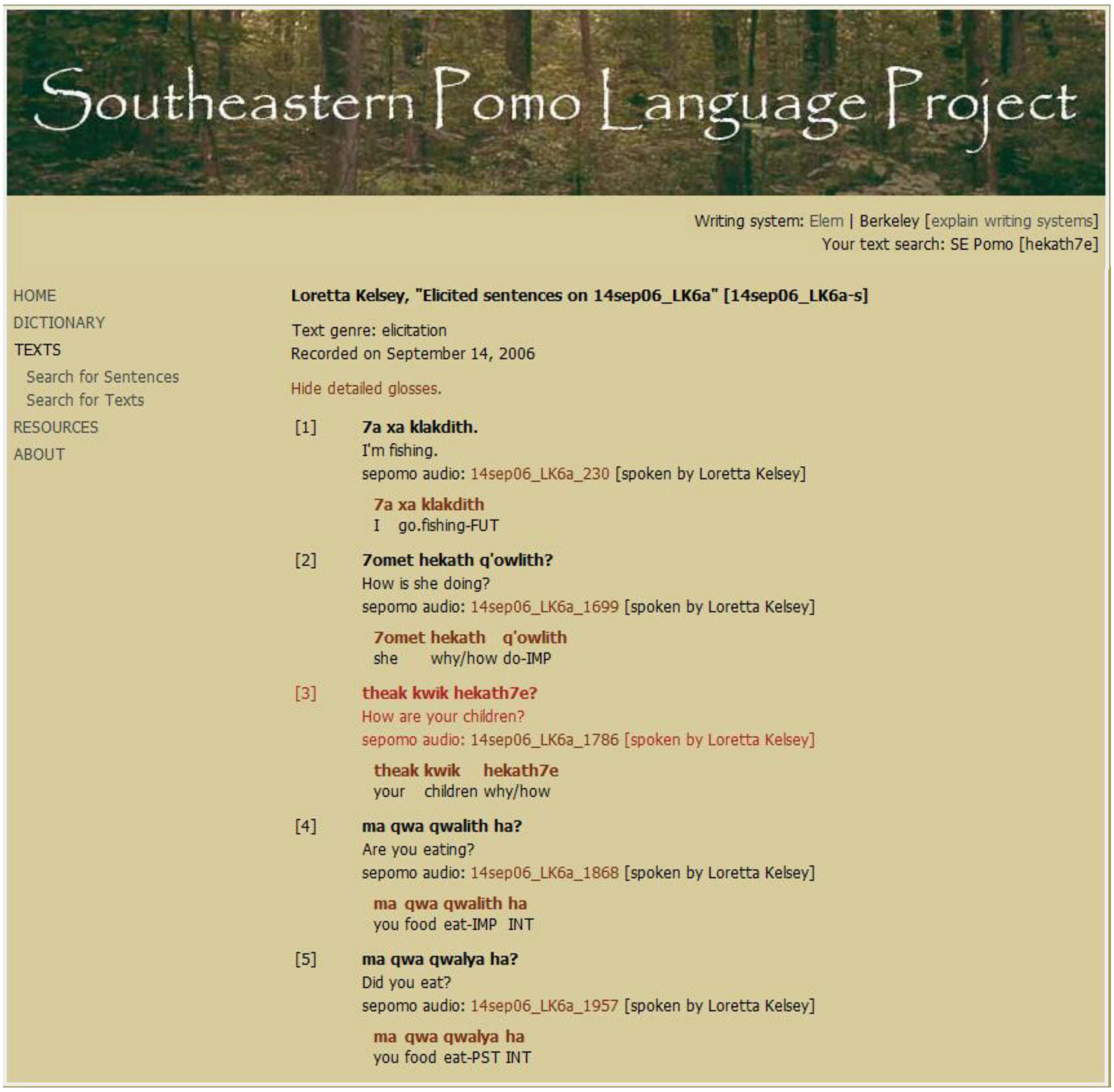
ma hekath7e?
How are you?
— 14sep06_LK5-s, Elicited sentences on 14sep06_LK5, sentence 1 (full context)
— SE Pomo audio: 14sep06_LK5_326 (spoken by Loretta Kelsey)

theak kwik hekath7e?
How are your children?
— 14sep06_LK6a-s, Elicited sentences on 14sep06_LK6a, sentence 3 (full context)
— SE Pomo audio: 14sep06_LK6a_1786 (spoken by Loretta Kelsey)

Figure 9. Results of a search for elicited sentences containing *hekath7e* from speaker Loretta Kelsey

On the text search page, users can search by the SEP/English words contained in the text, text genre (dialogue, elicited sentences, procedural text, remarks, and stories), or speaker. We demonstrate a sample search for elicited sentences that contain the SEP word *hekath7e* 'how' and are spoken by Loretta Kelsey (cf. Figure 8 above).

Three instances are found and displayed on the results page (cf. Figure 9 above). Each sentence is listed with the English translation and references to the original elicitation file and sound clipping. A 'full context' link can be clicked to view the whole text with detailed interlinear glosses. Figure 10 below shows the full text page for the third search result *theak kwik hekath7e?* 'How are your children?', with detailed glosses (the original search result sentence appears in red).



Writing system: Elem | Berkeley [explain writing systems]
Your text search: SE Pomo [hekath7e]

Loretta Kelsey, "Elicited sentences on 14sep06_LK6a" [14sep06_LK6a-s]

Text genre: elicitation
Recorded on September 14, 2006

Hide detailed glosses.

[1] **7a xa klakdith.**
I'm fishing.
sepomo audio: 14sep06_LK6a_230 [spoken by Loretta Kelsey]

7a	**xa klakdith**
I	go.fishing-FUT

[2] **7omet hekath q'owlith?**
How is she doing?
sepomo audio: 14sep06_LK6a_1699 [spoken by Loretta Kelsey]

7omet	**hekath**	**q'owlith**
she	why/how	do-IMP

[3] **theak kwik hekath7e?**
How are your children?
sepomo audio: 14sep06_LK6a_1786 [spoken by Loretta Kelsey]

theak	**kwik**	**hekath7e**
your	children	why/how

[4] **ma qwa qwalith ha?**
Are you eating?
sepomo audio: 14sep06_LK6a_1868 [spoken by Loretta Kelsey]

ma	**qwa**	**qwalith**	**ha**
you	food	eat-IMP	INT

[5] **ma qwa qwalya ha?**
Did you eat?
sepomo audio: 14sep06_LK6a_1957 [spoken by Loretta Kelsey]

ma	**qwa**	**qwalya**	**ha**
you	food	eat-PST	INT

Figure 10. Full text page for the selected sentence *theak kwik hekath7e* 'How are your children?' with detailed glosses. The selected sentence is in red.

5. Concluding Remarks

In this paper, we presented our work on an online dictionary for Southeastern Pomo. Specifically, we introduced the structure of the dictionary and demonstrated how the data are stored and processed on the website. We also showed the search interface of the website and demonstrated sample queries. However, this only represents the first step of the dictionary project. In the near future, we plan to develop and improve the dictionary website in the following ways. First, the data structure of the lexicon data file needs to be more elaborated (e.g. with nested structure) in order to accommodate polysemy more neatly. Second, the data of the print dictionary will be merged with that of the online dictionary, and all entries will be updated with their spelling in the Elem orthography. Third, different types of multimedia (e.g. photos of local flora and fauna, videos of the actions described by verbs of motion and placement) will be added to make the dictionary more informative and easier to use. Our ultimate goal, however, is to make the dictionary website widely accessible to teachers and learners so that they can make regular use of it as a CALL (Computer-Aided Language Learning) tool.

References

Babel, Molly, Andrew Garrett, Erin Haynes, Michael Houser, Reiko Kataoka, Fanny Liu, Nicole Marcus, Ruth Rouvier, Ronald Sprouse, Ange Strom-Weber, and Maziar Toosarvandani. 2006. A web-accessible Mono Lake Paiute dictionary and text archive. Paper presented at the Friends of Uto-Aztecan Conference. Salt Lake City, UT, University of Utah, August 24.

Dick, Grace, and Erin Haynes. 2006. A web-accessible Mono Lake Paiute dictionary and text archive. Paper presented at the Great Basin Language Conference. Bishop, CA, October 21.

Fagan, Kevin. 2007. Only living Elem Pomo speaker teaches so she won't be the last. *San Francisco Chronicle*, September 30. http://www.sfgate.com/cgi-bin/article.cgi?file= /c/a/2007/09/30/MNAISEMAH.DTL. Retrieved 1 July 2008.

Gordon, Raymond G., Jr., ed. 2005. *Ethnologue: Languages of the World*, 15th edition. Dallas, TX: SIL International. Online version: http://www.ethnologue.com.

Haynie, Hannah. 2007. Southeastern Pomo. http://hjhaynie.berkeley.edu/southeasternpomo. Retrieved 5 November 2007.

Moshinsky, Julius. 1974. *A Grammar of Southeastern Pomo* (University of California Publications in Linguistics 72). Berkeley, CA: University of California Press.

Shavelson, Lonny. 2006. California tribe tries to save its language. *Voice of America News*, March 30. http://www.voanews.com/english/archive/2006-03/2006-03-30-voa46. cfm?CFID=88126261&CFTOKEN=81958375. Retrieved 1 October 2006.