

Template for Multi-layered Language Learning Resources

4th ICLDC, February 2015

Edwin Ko¹, Elodie Paquette², Ethan Rimdzius², Elizabeth Grillo², Catherine O'Connor²
ek684@georgetown.edu, elodie.paquette@gmail.com, e.rimdzius@gmail.com, egrillo@bu.edu, mco@bu.edu,

¹Georgetown University ²Boston University

1 Background

Northern Pomo, one of seven distinct Pomo languages, was spoken in Northern California. The language is no longer fluently spoken. Recordings of two native speakers of Northern Pomo were made by Catherine O'Connor between 1979 and 2005.

In a project to archive and make available these resources to support revitalization efforts, there are often several options (i.e. place the materials into an archive, create a print dictionary, etc.). Ultimately, we decided to create mobile apps, and then a website with a variety of different entryways into the language.

2 Northern Pomo Language Tools

The **mobile apps** – Android and iOS (iPad only) – “jano sho:jin – Hearing Northern Pomo Words”.

- Contains 85 Northern Pomo words with pictures, grouped within 7 categories.
- Each word contains sound files by fluent speakers.
- Android version includes a quiz function.

The **website**, also optimized for viewing on mobiles and tablets, can be found at <http://northernpomolanguagetools.com>.

The website includes the following sections:

- **Sounds and Letters**
 - Includes a matrix of sounds in isolation, and videos training learners to hear the difference between linguistically significant sounds.
- **Talking Dictionary**
 - Contains sound files of individual words searchable in English and Northern Pomo.
- **Phrasicon**
 - Contains phrases and sentences in interlinear gloss format with sound files, searchable in English and Northern Pomo. Dictionary entries contain links to phrasicon entries, and vice versa.
- **Everyday Expressions**
 - Each expression is accompanied by a short explanatory video.
- **Stories and Texts**
 - Includes videos of each text featuring the voice of the speaker, and printable texts alongside each video, in both Northern Pomo and English.

3 Creating the Open Source Template

We decided to create templates, using the Northern Pomo website and Android app as models, and make them available broadly so others may use them for their own language work.

The web template consists of the following components: MAIN MENU, ABOUT THE MATERIALS, LINKS TO MOBILE APPS, SOUNDS AND LETTERS, EVERYDAY EXPRESSIONS, TALKING DICTIONARY, PHASICON, BASIC SENTENCE STRUCTURES*, STORIES AND TEXTS, and CONTACT.

The web template may be viewed on the following website: <http://eddersko.com/template/demo/>.

To test the templates with other languages, we implemented the following:

- A version of the Android app for Kashaya Pomo, a Pomo language of Northern California.
 - Contains 113 Kashaya Pomo words within 13 categories, and a quiz.
- A version of the website for Medumba, a Grassfields Bantoid language located at this site: <http://elodiepaquette.com/medumba/>.
 - This version was set up by a member of our team who has prior experience with Java programming, but who has no background in web programming.
 - The dictionary and phrasicon contain an additional annotation layer to encode tone.
 - This version uses only the talking dictionary and phrasicon.

Note: It is possible to add additional layers of annotation, or/and make use of only parts of the template.

4 Implementing the Templates

The templates can be downloaded, and used by people working on other revitalization projects.

4.1 Web App Template

Ideally, the team will include a person who is familiar with computer programming, or is willing to explore the code.

Once a server to host the website has been established, the template can be copied into the root directory.

Simply by changing a few lines of code, you will be able to transform the template into a fully functional language learning resource which you can then populate with sound files and materials of your own.

The documentation, and installation package can be found here: <http://eddersko.github.io/web-template/>.

4.2 Android App Template

In order to create an Android app using the template, it will require someone who is either familiar with Android development, or Java programming.

The documentation, and installation package can be found here: <http://eddersko.github.io/android-template/>.

5 Technical and Design Decisions

In our website described above, Northern Pomo is the “focal language” and English is the “user language.” However, any language could be the focal language and any language could be the language of the users building and using the website.

5.1 Talking Dictionary

5.1.1 Dictionary Database Format

```
<dictionary>
  <metadata>
    <!-- Open Language Archives Community Metadata -->
  </metadata>
  <entry id="266">
    <form>
      <orth>shaku:l</orth>
    </form>
    <sense>
      <cit type="translation" lang="en">
        <usg type="hyper">fish</usg>
        <quote>little grilled fish</quote>
      </cit>
    </sense>
    <gramGrp>
      <pos>noun</pos>
    </gramGrp>
    <note>name for small surf fish found at Pacific coast, grilled and served at
    feast</note>
    <medial mimeType="audio/wav" url="little_grilled_fish_ES.wav"/>
    <ref>ES</ref>
  </entry>
</dictionary>
```

Figure 1. XML representation of a Talking Dictionary entry.

The dictionary database adopts OLAC metadata (Bird and Simmons, 2001), and TEI consortium standards.

5.1.2 Organization

Semantically related words are grouped in the field `<usg type="hyper">fish</usg>` which is loosely based on the definition of hypernym (cf. superordinate).

In Figure 1, the hypernym of *little fish* is ‘fish’. The hypernym of *fish* would also be ‘fish’. The base form of the verbs *went* and *going* is ‘go’. (Here, we extend the term *hypernym* a bit to include lemmas of verbs.)

Entries that share the same hypernym appear collected on the “detail view page” that appears when users click on a word in the search result list.

5.1.3 Search Function

The dictionary is searchable in English by selecting a letter (A-Z) category, or searching in English and the focal language via typing a word into the search bar.

Querying a word in the search bar retrieves entries where the target word appears in the head word or hypernym field. Retrieved words are listed in columns linking to the word’s entry page.

When the query looks to see if the query contains more than three characters.

- If so, results may contain head words that start with the characters in the query.
Example: Searching for “head” may return “headache”.
- If the word contains three or less characters, results may contain headwords consisting of several words, where one of these words matches the query.
Searching for “man” may return “old man”, but not “manner”.

When no results are found for English, users are presented with an option to “Try an extended search”. Doing so will pass the original query through the Porter Stemmer algorithm (Porter 1980) written in PHP, and will try the search again.

Note: While the stemmer executes faster than a lemmatizer, the results are sometimes less accurate.

5.1.4 Linking to the Phrasicon

Dictionary entries may contain a link to the Phrasicon under the following condition:

- If the dictionary *focal language* head word appears as a morpheme, or in a phrase or sentence in the *focal language* Phrasicon entry, then provide a link to these Phrasicon records.

The linkage also checks for homophones and polysemy.

5.2 Phrasicon

5.2.1 Phrasicon Database Format

```
<phrasicon>
  <metadata>
    <!-- Open Language Archives Community Metadata -->
  </metadata>
  <phrase id="58">
    <ref>ES</ref>
    <source>Bo khe hayu na</source>
    <morpheme>
      <m id="58.1">Bo</m>
      <m id="58.2">khe</m>
      <m id="58.3">hayu</m>
      <m id="58.4">na</m>
    </morpheme>
    <gloss lang="en">
      <g id="58.1">Bo</g>
      <g id="58.2">my</g>
      <g id="58.3">dog</g>
      <g id="58.4">is</g>
    </gloss>
    <translation lang="en">Bo is my dog</translation>
    <media mimeType="wav" url="Bo_is_my_dog.wav"/>
  </phrase>
</phrasicon>
```

Figure 2. XML representation of a Phrasicon entry.

The dictionary database adopts OLAC metadata, and loosely follows TEI consortium standards.

The format is modeled after the Bow-Hughes-Bird (BHB) interlinear glossed text (IGT) model (Bow et al., 2003), and IGT-XML (Palmer and Erk, 2007). These formats were initially designed for texts.

The XML representation in Figure 2 (see below) was primarily designed for a corpus of phrases and sentences collected from fieldwork.

5.2.2 Search Function

The Phrasicon is searchable in English by selecting a letter (A-Z) category or via the search bar, and is also searchable in the focal language via the search bar.

Querying an English word will return entries where the word appears in the gloss line or in the translation. Querying a focal language word will return entries where the word appears as in the morpheme line or in a phrase in the focal language.

Searching with the “hyper search” is only compatible in English. The search first checks to see if the word exists in the dictionary.

- If so, retrieve Phrasicon entries with words that share the query’s hypernym.
- If not, perform a regular search.

5.2.3 Linking to the Dictionary

Morphemes in Phrasicon entries may link to the Talking Dictionary under the following condition:

- If the gloss appears as the English head word in a dictionary entry, then provide a link to that dictionary entry from the Phrasicon.

5.2.4 Data Entry

When the user is creating a new entry for the Phrasicon, they will start by typing in the focal language morphemes into the cells in the data entry form. The **auto-fill function** attempts to automatically fill in the English translation for each focal language morpheme. The auto-fill function uses a semi-supervised learning technique (cf. maximum entropy classifier) written in PHP. The algorithm predicts the morpheme’s gloss according to the morpheme itself, and its context within two morphemes, before and after (Palmer et al., 2009).

5.3 Web Design Framework

The website’s responsive design is built on Foundation 5, an open-source front-end framework.

5.4 Android App Quiz

The Android quiz uses the Java Random class to pull words ‘randomly’ from within the particular quiz category.

In designing the quiz, it was decided that it would be more motivating and encouraging to avoid the frequently encountered “error buzzers” or “red X” error indicators for incorrect choices in the quiz. Instead, clicking on any word within the quiz will grey out the incorrect answers, form a

yellow border around the correct one, and replay the sound file of the correct answer, helping the user form an association between the image of the word and the sound of the word.

6 Future Plans

Currently, the talking dictionary and phrasicon support only one user language and one focal language. Our aim is to modify the existing template to support at least two user languages. E.g. the website built around Medumba as a focal language could use both English and French as user languages, because there are many potential users who speak either or both English and French.

We are also looking forward to releasing the iPhone version of the app, and making available the templates for iOS versions.

Finally, we hope to create conversion tools from other dictionary/IGT formats.

For further information, contact Edwin Ko: ek684@gerogetown.edu.

7 References

- Bird, Steven and Gary Simons. 2001. The OLAC meta-data set and controlled vocabularies. In *Proceedings of ACL Workshop on Sharing Tools and Resources for Research and Education*, pages 7–18, Toulouse.
- Bird, Steven and Gary Simons. 2003. Seven dimensions of portability for language documentation and description. *Language*, 79(3):557–582.
- Bow, C., B. Hughes, and S. Bird. 2003. Towards a general model of interlinear text. In *Proceedings of EMELD Workshop 2003: Digitizing and Annotating Texts and Field Recordings*, LSA Institute: Lansing MI, USA.
- Palmer, A., T. Moon, J. Baldrige. Evaluating automation strategies in language documentation, In *Proceedings of the NAACL-HLT 2009 Workshop on Active Learning for Natural Language Processing*, Boulder, CO.
- Palmer, Alexis and Katrin Erk. 2007. IGT-XML: An XML format for interlinearized glossed text. In *Proceedings of the Linguistic Annotation Workshop (LAW-07), ACL07*, Prague.
- Porter, M. F. 1980, An algorithm for suffix stripping, *Program*, 14(3) pp 130–137.
- TEI Consortium. TEI P5 Guidelines Version 2.7.0 September 16, 2014. <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/>.
- Zurb Foundation. Foundation 5 Version 5.2.3 May 28, 2014. <http://foundation.zurb.com/docs/>.