Recovering and updating legacy dictionary data

Dawn Bates / Deryle Lonsdale
Arizona State University / Brigham Young University

Ongoing efforts to annotate and web-enable Lushootseed language resources involve displaying complex dictionary information in ways suitable for diverse users. In this paper we discuss how we have converted dictionary data from its legacy format into a best-practice, state-of-the-art XML format. We also describe how we have been able to further leverage this XML data by making it compatible with Kirrkirr, a new dictionary browser designed to display dictionary information for Australian aboriginal languages. We sketch the process in adapting the dictionary browser to make it a very workable visualization tool for Lushootseed data as well. We also explain and demonstrate how Kirrkirr displays the various types of Lushootseed data, and we document some of the issues and difficulties inherent in using Kirrkirr as a visualizer for the Lushootseed data.

1 Background

The Lushootseed (Puget Salish or dx^{w}ləšucid) language (SIL ISO/DIS 639-3 language code LUT) was at one time spoken by more than forty tribes and bands living along the eastern, southern and southwestern shores of Washington state's Puget Sound (Smith 1941), (Suttles 1990) and on islands in the Sound and to the north (Sampson 1972), (Roberts 1975), (Collins 1974). Each tribe is spoken of as having had its own dialect; today, we speak of a main division between Northern and Southern Lushootseed for the fourteen dialects that remain (Hess 1977) (Czaykowska-Higgins 1997).

Though the number of first-language speakers of Lushootseed continues to dwindle, language revitalization programs teach the language to adults and children of Salish heritage and others in Native communities and schools around Puget Sound. These language programs have benefited enormously from the efforts of one particularly energetic first-language speaker, teacher, and researcher, the late Violet taq^{w}šəblu Hilbert, an elder of the Upper

Skagit Tribe, in the Northern Lushootseed area. Since the early 1970s, Hilbert dedicated her life to the documentation, preservation, and revival of Lushootseed (Hilbert 1982) by her vast collection of lexical, textual, audio and video resources. Along with collaborators her works cover a wide spectrum of publications including dictionaries. This paper discusses work carried out to bring the data from one work in particular—a dictionary—into the twenty-first century in a way that hopefully will provide a valuable tool for Lushootseed-speaking communities.

The data discussed here originated over several years of work by Thom Hess with Louise George and Mr. and Mrs. Lamont. During this time he kept a file box recording lexical items and associated data. During his first years teaching at the University of Victoria, he contracted with the staff there in the Department of Linguistics to type up the cards, perhaps after promising Mrs. George a published dictionary. Derived directly from that effort, the University of Washington Press published the Dictionary of Puget Salish (DPS), (Hess 1976). This was the first major dictionary of the Lushootseed language; it was published before the term Lushootseed caught on in English. The format of DPS follows fairly closely the format of Hess's file cards. The dictionary is artistically bound with silver gilt-blocked cloth and is oversize and rather weighty, with heavy paper and ample whitespace throughout its 770 pages. Researchers and tribal elders alike held Hess's dictionary in great regard, but because of its uneconomical format it was prohibitively expensive. Still, it went out of print rather quickly.

2 The Lushootseed Dictionary

Hess and Hilbert began collaborating around the time the DPS was published. She started doing field work on her own and kept her own lexical file cards. She extended the lexical inventory by perhaps 20% for a revised and expanded dictionary and hoped to see this additional material published. Accordingly she and her colleagues proposed a grant to publish an updated Lushootseed dictionary.

The overall goal of the new Lushootseed Dictionary project was to use computer technology to provide a less expensive, more portable dictionary that included all of the DPS material as well as Hilbert's and Hess's work since its publication. Part of the effort included using LEXWARE, a custom-developed computer system to help in encoding dictionary data and formatting it for print (Hsu 1989). The system was being widely used elsewhere in preparing dictionaries for Salish and other languages (Czaykowska-Higgins 1997, 63) (Mithun 1999). The system was greatly appreciated by lexicographers and

greatly improved dictionary presentation formats while decreasing production time.

Use of the system required the lexicographer to define a hierarchical set of bands, which consist of text fields identified by tags (Poser n.d.). Though the tags don't appear in the output, they are interpreted to create the presentation format for the dictionary data. Preparation of the LD thus required fitting the DPS data and subsequently collected material into the LEXWARE format. Then custom computer software rendered the data, interpreting the bands, into the desired print publication format.

The resulting 381-page Lushootseed Dictionary (Bates, Hess and Hilbert 1994) is a bilingual (bidirectional) dictionary with data tagged for morphological information, cultural commentary, and dialect information; its main feature is a Lushootseed-English dictionary although it also has an English-to-Lushootseed glossary. LD also contains references to particular native-speaking consultants and the introduction has biographical information about each of them. Each subentry in the Dictionary provides multiple potential points of intratextual linkage suggested by extensive cross-referencing to previously collected sound recordings and text corpora.

LD is more affordable, more convenient to use, and more efficient in its print presentation than DPS was. It has sold well and received enthusiastic reviews (Lonsdale 1996) (Galloway 1995). The publisher has been able to keep LD in press, and has also made copies available for Native language programs at a reduced price.

3 Recovering LD data

The computerized data used to produce the dictionary fell into disuse almost immediately because of the success of the print version and the lack of need to refer back to the original data files. For several years it languished in electronic mothballs as computers, operating systems, and application programs evolved, rendering the data increasingly out-of-date and inaccessible.

The authors of this paper began collaborating about ten years ago and decided that one of their first projects would be to rehabilitate as much as possible of the computerized LD data. The files were copied onto 31/2-inch floppy disks for transfer to another machine. The first goal was to extract as much of the recoverable data as possible and convert it to raw ASCII text. Programs were written in Perl that included regular expression matching and bit-level conversion to map the Lushootseed letters to an unambiguous but idiosyncratic Romanized transliteration. The LEXWARE codes were preserved in-place for downstream processing. Flat files were produced that were only usable by users capable of running macros, searches, or regular expression

matches across the data to locate items of interest. Even then, the idiosyncratic Romanized text rendered the task even more opaque. Still, the text had been successfully rescued from the brink of digital abandonment.

4 Adopting best practices

For several years now work in language resource development and archiving has led to the identification of best practices that should be followed to assure longevity and forward-compatibility of language data (see, for example, http://emeld.org/school/). Recommendations include following computing industry standards for character encoding, data markup, and file management. Up to this point in the LD recovery process time was of the essence, so recovering the data was done with shorter-term objectives in view. It was always intended, though, that once a stable foundation was established, best practices would be adhered to in rehabilitating and redeploying the LD data.

To overcome these difficulties the next step was to convert the raw ASCII text to HTML code. Again, Perl scripting was used to reformat the transliterated Lushootseed text to HTML entities. The characters themselves (more precisely their Romanized representations) were converted to UTF-8 (i.e. Unicode) entities. This was occasionally problematic since the Unicode standard at the time did not provide entities for all of the Lushootseed characters, and font display was similarly adversely affected.

During the conversion process many of the LEXWARE fields were mapped to HTML tags, particularly when they reflected some of the more visual aspects of the paper dictionary format, for example italic text. During this process several errors in the original data files came to light and were corrected at the time; we are currently working on a webpage that lists errata for the dictionary. With some hand-cleaning we were able to produce a research prototype website with all of the entries from the Lushootseed-English portion of LD, with all words starting with a given letter on their own web page. Another page was created that contained all of the Lushootseed entries to provide dictionary-wide searching online.

Creation of this website allowed us for the first time to browse or search the contents of the dictionary in a browser (Internet Explorer or Firefox), which has proven very useful for research purposes but needs further enhancement to support end-users. To date no keyboard entry method has been integrated with the browsers, so for entering “exotic” characters into a search field cut-and-paste is currently required.

One problem with the version of the dictionary marked up in HTML was that it was still primarily display-oriented, which meant that the encoded

information was more aimed at form than at content. One way around this problem was to coerce the data to a representation that focuses on tagging the fields more for their semantics and their functional role in the entry. Similar issues had arisen with other types of data, and XML (the eXtensible Markup Language) was designed to address this challenge. XML allows users to create their own tags for content-based markup, as opposed to the format-based tags that HTML provides. As long as a proper schema for XML markup is defined and followed meticulously, manipulating XML-structured data in a wide variety of applications becomes feasible. For example, applications had been developed to convert XML dictionary data automatically into a print dictionary format. The issue then became how to define a workable XML tagset for the LD data.

Fortunately, the Text Encoding Initiative was introduced by scholars interested in developing a unified and authoritative XML standard for several types of linguistic markup including dictionary entries. The TEI XML standards that have emerged from their deliberations have been widely adopted by researchers interested in best-practices encoding and preservation of linguistic data. An early dictionary application using TEI-encoded data displayable on the web was one for the Slovene dialect of Resia (see http://www.tei-c.org/Activities/Projects/re02.xml). Using that project as a model we decided to encode the LD data into the TEI XML format.

Version P4 of the TEI XML standard includes a provision specifically designed for print dictionaries (see http://www.tei-c.org/P4X/DI.html). This includes a schema for marking up dictionary sections, individual entries, and all possible types of data that entries typically contain. We converted the HTML version of the dictionary into TEI P4-compliant structure using a combination of Perl scripts and macros in raw text editors. The result is the complete set of about 3600 fully tagged Lushootseed entries.

Recently the P4 version of TEI was upgraded to a new version, TEI P5. This new standard, like the old one, has a format for dictionaries with some minor modifications. Converting the P4 dictionary data to P5 format was relatively straightforward, this time using a commercial XML editor to run macros over the data. The use of this type of editor was very convenient since it was possible to associate the LD XML data file with pre-existing TEI schema definitions, allowing the editor to continually monitor the LD data and signal immediately any erroneous tagging that didn't adhere to the TEI guidelines.

One lingering and important issue confronted us, though, in using the TEI P5 version of the dictionary. The data was all encoded in raw XML, which is not usable by anybody except the most enthusiastic corpus markup experts. Until recently, though, there was no publicly available tool for visualizing the dictionary data in a browser-like environment to access the various fields. In

principle it was possible to develop such a tool, but we were not in a position to invest the time necessary to accomplish this task. We did expect, however, that since TEI XML is a widely implemented markup scheme, that someone would eventually develop an open source dictionary browser.

5 XML for a dictionary browser

Meanwhile a group working with the Warlpiri language was developing a browser for displaying XML-based Australian Aboriginal language dictionary data (Manning, Jansz and Indurkhya 2001). Called Kirrkirr, the system was developed in a Java framework and is highly portable to various computer platforms. Because of its modular design it's an extremely flexible program that allows the user access to a wide range of data in various formats. The system's executables were released to the public, and since its design was adaptable and fairly well documented, in theory other language dictionaries could be substituted for the Warlpiri dictionary with minimal effort, provided the data is in XML format. In fact, small sample dictionaries for other languages are distributed with the system as a pattern to follow for the implementation of other languages. Another advantage of Kirrkirr is that it also supports links to multimedia resources like images and audio samples. The interface is highly configurable by the user so that only items of interest are visible at any given time. Finally, the user can record notes for any entry and maintain personalized word lists.

We determined to see whether browsing LD entries would be viable in Kirrkirr. In order to make the LD data displayable via Kirrkirr, several steps were required. Carrying them out was relatively straightforward following the system documentation provided at the Kirrkirr website, as well as via occasional emails with the developers.

One not altogether necessary step was to convert the LD data from its TEI P5 XML format to a more neutral XML representation. This was deemed necessary since the TEI markup was structured differently enough from Kirrkirr's required format that the conversion in one step seemed too daunting (though in theory it would be possible). Instead, the LD markup was converted to a new set of tags that more closely matched those given in the language samples for Kirrkirr dictionary integration. For example, the sample TEI snippet <form type="lemma"><orth>cícu? </orth> </form> was converted to the item <FORM><HW>cícu?</HW></FORM>. Another step, this time necessary, was to specify how the data in the entries should be displayed in the browser. It displays individual entries using HTML format, so it was necessary to define an XML-to-HTML mapping to assure entry display. For example, colors were

assigned to certain semantic fields, while italics and boldface features were added to others.

For both of these XML conversion steps we developed transforms in the XSLT programming/scripting language, especially designed to facilitate mapping between different XML representations. Our XML editor allowed us to associate a transformation engine (in this case Saxon-B) with the two conversions (i.e. to HTML for entry display and to XML for the Kirrkirr program).

In addition, some adaptation of the interface had to take place. For example, Lushootseed data is displayed using a variety of open-source and proprietary fonts that the browser needed to use. Integrating this new font into Kirrkirr required some interaction with the developers; ultimately the specification of which font(s) to use is specified in a parameter file that the browser consults on startup. The browser also allows for icons to represent the source and target languages for bilingual dictionaries. We simply chose the image of an orca as the icon for the Lushootseed language and replaced the Australian flag (the system's default English icon) with an image of the U.S. flag.

6 Visualizing the LD data

Kirrkirr supports many methods for accessing data, and hence is a very flexible environment for visualizing the LD lexicon. In this section we briefly describe some of the ways a user can browse LD entries; fuller description of Kirrkirr's capabilities are documented elsewhere (McElvenny 2008).

Of course, one way to access entries is by headwords. A window in the browser lists in alphabetical order all of the headwords in the dictionary. Selecting a headword brings up its entry. Each LD entry has several possible fields including phonological structure (e.g. CV tier or syllable information), morphological information (e.g. morpheme type), semantic information (e.g. lexical meaning and semantic domain), dialect notes, usage examples, and a cross-reference to the corresponding entry in DPS as well as to related LD entries. The display information, as described above, has been rendered into HTML format, originating from XML tags directly mapped from LEXWARE band information. When entries have subentries associated with them, the subwords also show up indented in the headword list. Figure 1 shows the browser displaying a portion of the headword list and a typical entry.

One useful feature of the browser is the ability to show a network of related words based on various lexical relations (e.g. antonyms, synonyms, compared or contrasted forms) that are encoded in the dictionary data but not advantageously presented in the published format of the LD. For example, the

network display of the entry for the Lushootseed word for “black bear” also shows related words such as dialectal variants, derived words such as the words for “bear cub”, “bear cubs”, and “bears”; additionally, the words for “grizzly bear” also appear in the network. Figure 2 shows the semantic network for entries associated with these concepts.

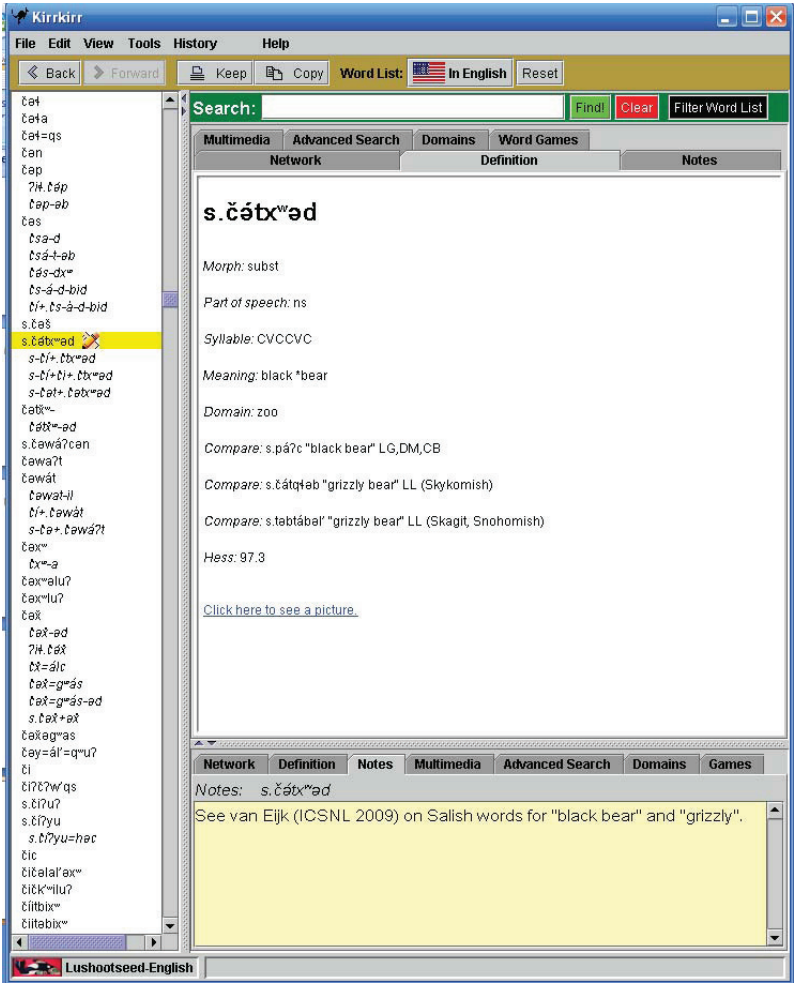

Figure 1: Sample LD entry with headword list (left) and user notepad (bottom right).

Kirrkirr also groups entries together by semantic domain, which allows the user to see all words associated in this manner. Some LEXWARE bands included semantic domain information, so rendering these into XML tags usable by Kirrkirr was straightforward. Thus it is possible to see, for example, in one place all of the words that are subsumed under the BOTANY domain, or any of the thirty or more domains.

Entries can also be searched by custom searches including regular expression matching and fuzzy matching on specific fields or across the whole entry. Finding entries is thus extremely convenient. The system even supports a reverse index function, which allows browsing and accessing Lushootseed entries by English keywords that are flagged in the entries.

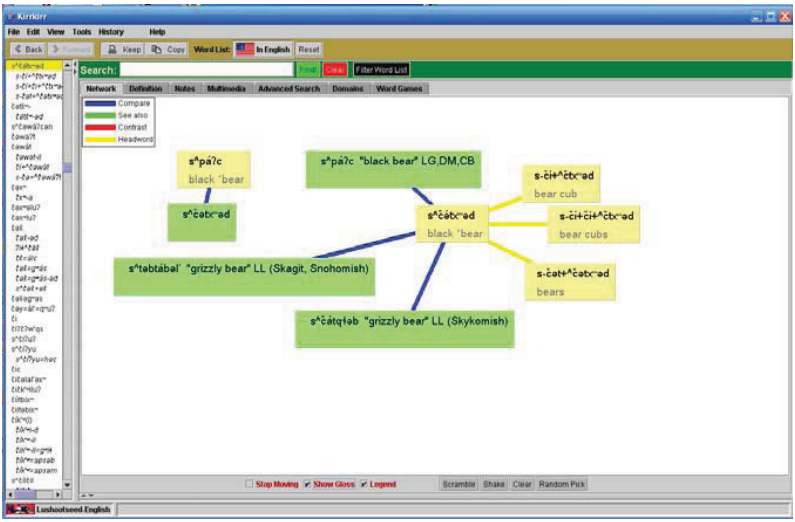

Figure 2: Browsing the semantic network for associated entries.

Finally, dictionary entries can be linked to multimedia files that illustrate the concept associated with the entry, or give sound samples of the pronunciation of the headword or derived words. We incorporated into the LD entries several examples of sound files and images we had previously harvested from various websites, and these were therefore available to the user. Figure 3 shows a sample entry with an image and associated audio clips.

7 Future work and vision

Though a prototype version of the LD dictionary data has been implemented in Kirrkirr, several items of work are still ongoing to perfect the integration of the data with the visualizer.

One area requiring more work concerns the polysemy-homonymy spectrum. Some LD entries exhibit polysemy (i.e. several related meanings grouped under the same entry), and others exhibit homonymy (i.e. unrelated meanings with the same spelling split out as separate entries). Kirrkirr has some flexibility for handling this distinction, but it doesn't completely agree with how LD entries were set up. The two approaches can be reconciled during the XML mapping process, but so far this has not been attempted.

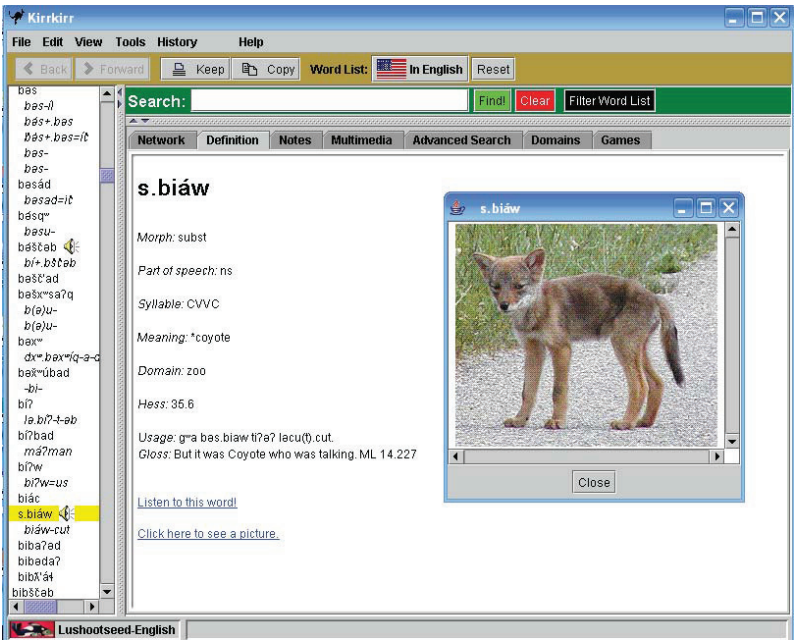

Figure 3: An entry with multimedia content (image and sound file).

In working with the various version of the XML encodings of LD data, the need for cleaning up various types of data has become evident. Accordingly, a more concerted effort is needed to focus on better tagging of some items that are not yet completely marked, such as the consultant codes used in many usage examples.

Since the publication of LD, corpus research and data collection (including archival work) have identified many more Lushootseed lexical items than currently represented in the LD data. These additional items could and should be integrated into the XML version to assure more complete lexical coverage of the language.

There are some lingering difficulties with using Kirrkirr that would ideally be resolved, but that would require developer involvement or access to the source code at the very least. Two characters used extensively in LD data do not display properly everywhere in the browser: the raised dot, which signifies vowel lengthening, and the square root sign, which indicates the root. While alternative characters have been used in the prototype LD implementation as a workaround, they are not optimal. Kirrkirr has functionality to generate games from the entry data provided, such as vocabulary quizzes. Unfortunately, games do not work well for "exotic" script languages, so Lushootseed games cannot be generated. The Kirrkirr program was designed to be installed and run on individual machines, but it would be nice to have a similar interface deployable on the web. Finally, we expect to be able to allow Kirrkirr access to external computational tools, for example the morphology engine described elsewhere (Lonsdale 2003). This would associate with the browser the ability to parse out words for their morphological structure on demand.

Our motivation for undertaking this effort is to lay the groundwork for a comprehensive online lexicon of Lushootseed words that will serve language learners and researchers alike. Several considerations favor computerized deployment of LD in addition to the hardcopy version. First, the flexible, customizable browser presentation format can be tailored to more readily meet the needs of curriculum designers and students; entries in the hardcopy dictionary are arranged alphabetically by root, while most students and teachers would prefer to look up material by full word or semantic domain. The improved access methods and information content will provide end users new functionality that will empower them to create customized solutions for their particular needs. A further benefit from computerized access to the Dictionary is that the XML-encoded materials will be dynamic, continually updated and more easily integrated with hypermedia formats (images, sound, websites, etc.). This will enable users to effectively access a wide range of language resources appropriate to individual and pedagogical needs.

References

Bates, Dawn, Thom Hess, and Vi Hilbert. *Lushootseed Dictionary.* Seattle and London: University of Washington Press, 1994.

Czaykowska-Higgins, Ewa and M. Dale Kinkade, ed. *Salish Languages and Linguistics: Theoretical and Descriptive Perspectives.* Vol. Volume 107 of Trends in Linguistics: Series and Monographs. Berlin: Mouton de Gruyter, 1997.

Galloway, Brent. "Review of Lushootseed Dictionary." *American Indian Culture and Research Journal* 19 (1995): 293–296.

Hess, Thom. *Dictionary of Puget Salish.* Seattle and London: University of Washington Press, 1976.

Hess, Thom. "Lushootseed Dialects." *Anthropological Linguistics* 19 (1977): 403-419.

Hilbert, Vi (taqwšEblu), and Thom Hess. "The Lushootseed Language Project." In *Language renewal among American Indian tribes*, edited by Robert St. Clair and William Leap, 71-89. Rosslyn, VA: Nat'l Clearinghouse for Bilingual Education, 1982.

Hsu, Robert. *Lexware Manual.* Vol. Second Edition. Honolulu, HI: Linguistics Dept., University of Hawaii, 1989.

Lonsdale, Deryle. "Book Review: Lushootseed Dictionary." *Language* (Linguistic Society of America) 72, no. 3 (1996): 644-645.

Lonsdale, Deryle. "Two-level Engines for Salish Morphology." *Proceedings of the Workshop on Finite-State Methods in Natural Language Processing.* European Association for Computational Linguistics, 2003. 35-42.

Manning, Christopher D., Kevin Jansz, and Nitin Indurkhya. "Kirrkirr: Software for Browsing and Visual Exploration of a Structured Warlpiri Dictionary." *Literary and Linguistic Computing* 16, no. 2 (2001).

McElvenny, James. "Review: Kirrkirr." *Language Documentation & Conservation* 2, no. 1 (June 2008): 160-165.

Mithun, Marianne. *The Languages of Native North America.* Cambridge: Cambridge University Press, 1999.

Poser, Bill. *Parsing Lexical Database Files.* http://www.billposer.org/Linguistics/Computation/LectureNotes/ParsingLexica.html.

Roberts, Natalie. *A History of the Swinomish Tribal community.* University of Washington, 1975.

Sampson, Martin J. *Indians of Skagit County.* Mount Vernon, WA: Skagit County Historical Society, 1972.

Smith, Miarian W. "The Coast Salish of Puget Sound." *American Anthropologist* 43 (1941): 197-211.

Suttles, Wayne, ed. *Handbook of North American Indians.* Vol. 7. The Northwest Coast. Washington, DC: Smithsonian Institution, 1990.

Deryle Lonsdale / Dawn Bates
lonz@byu.edu / dawn.bates@asu.edu