Annotating and exploring Lushootseed morphosyntax

Deryle Lonsdale and Hitokazu Matsushita
Brigham Young University

In this paper we provide information on an initial prototype investigation, on a modest scale, into the morphosyntactic structure of Lushootseed, a Salish language. We begin by describing language resources and tools that were instrumental for the process: an XML-encoded dictionary, a morphological parser, and a syntactic parser. We then illustrate how the output can be stored in a relational database and queried to extract relationships and statistics about them. We also sketch ongoing work to integrate these tools and render them more accessible to users with modest technical skills.

1 Introduction

Lushootseed, like its related Coast Salish languages, exhibits rich morphological and syntactic structure. Though few native speakers of the language remain today, there is an active revitalization effort and a growing number of young learners of the language. The work reported in this paper is being carried out to help stakeholders in the language understand how to collect, analyze, and deploy resources that can serve in documenting the language and ideally developing materials for language learning. These resources should also help learners and teachers gain a fuller appreciation for the language's grammatical structures and their richness. In this paper we focus on the technical aspects of assembling, annotating, and publishing language data to that end, and the tools and resources that enable such effort.

2 Resources and tools

Crucial to the analysis of language data are language resources, which provide information about the basic components of a language and exemplify how they combine to account for the full range of possibile structures. This section sketches some language resources that have been accumulated for Lushootseed and adapted for various usages.

Foremost among language resources are those that provide lexical information: words, vocabulary, and phrases. A significant lexical resource for Lushootseed is the canonical dictionary for the language in its two instantiations: the Dictionary of Puget Salish (Hess, 1976) and the Lushootseed Dictionary (Bates et al., 1994). In prior work we described efforts to rehabilitate and update the legacy data that constituted the typesetting input to the publication process

for the latter form of the dictionary, and how we were able to update the content to current best-practice format following the Text Encoding Initiative (TEI) format[1] for XML encoding of print dictionary information (Bates and Lonsdale, 2010). For reasons that will be apparent in the next section, this format of the dictionary is crucial for use in other levels of linguistic analysis.

Another type of language resource that is valuable for language processing is the corpus: a principled collection of textual and/or spoken language documents. The most valuable type of corpora have been systematically assembled based on predefined specifications, requirements, and end goals. Corpus documents can come from a wide variety of sources such as archival repositories, publications, web content, and audio transcriptions. In the next section we describe a corpus that for other languages would be negligible but for Lushootseed is among the largest analyzed to date.

Annotated corpora are particularly valuable. They consist of linguistic information that is explicitly embedded in the text or noted in a standoff fashion in parallel documents. The specifics of the annotation may vary depending on the annotators' purposes, theoretical inclinations, or available resources. Typical types of corpus annotation include part of speech tags for words, syntactic structure in the form of parsed sentences (e.g. in treebanks), codes representing word senses for particular usage instances, or dialogue turns in a conversation. Performing linguistic annotation is almost always a costly proposition: the expertise, time, and effort required for annotation is substantial. Consequently much current research in corpus annotation focuses on ways of automating the process.

2.1 Morphological parsing

Lushootseed is a polysynthetic language, exhibiting a rich morphological structure. Its basic system of roots, however, is fairly simple, most being either monosyllabic or disyllabic. Derivation and inflection are pevasive, and most roots can take any inflection (e.g. aspect can occur on nouns and adjectives). Affixation, reduplication and incorporation are very frequent, though compounding is not. An elaborate system of bound lexical morphemes also adds to morphological complexity.

We required a tool for parsing and generating words in the language, and clearly ad-hoc "cut-and-paste" methods, such as the Porter stemming algorithm used for English web searches, are inadequate. For processing Lushootseed we therefore adopted a finite-state model—the two-level model—of computational morphology for this task (Koskenniemi, 1983). The two-level approach has been applied to a variety of languages, generally morphologically complex ones such as Finnish, Turkish, Arabic, and at least one native American language: Aymara.

The morphology engine we use is PC-Kimmo (Antworth, 1990)[2]. The system requires language-specific knowledge sources including lexicons, rule

[1] See `www.tei-c.org`

[2] See the website at `www.sil.org/pckimmo/`

files, and grammar files. The engine is is capable of two basic modes of operation: (i) recognition, in which a fully inflected word is processed by the system to arrive at a description of the word's morphological decomposition(s); and (ii) generation, in which a specification of underlying morphemes is processed by the system to produce the corresponding surface form(s) of the word. The system uses a collection of one or more lexicons to represent the basic morphemic inventory of a language. As a word is processed letter-by-letter, the lexicon subsystem is used as a basic device to control and license search through possible sequences of letters and morphemes for a word. More technical details on the PC-Kimmo implementation for Lushootseed are available elsewhere (Lonsdale, 2001; Lonsdale, 2003). Figure 1 shows examples of parsed words with their morphemic decomposition and corresponding English morphemic glosses.

```
PC-KIMMO>recognize gWEdsutudZildubut
gWE+d+s+?u+^tudZil+du+b+ut   Dub+my+Nomz+Perf+bend_over+OOC+Midd+Rfx

PC-KIMMO>recognize adsukWaxWdubs
ad+s+?u+^kWaxW+du+b+s     Your+Nomz+Perf+help+OOC+Midd+his/hers
```

Figure 1: Sample PC-Kimmo word parse output for Lushootseed.

As with many implementations, we use a separate lexicon for each of the possible positions where inflection and derivation can take place in Lushootseed words: our implementation uses over ten separate lexicons. All data in each lexicon is represented in Romanized ASCII transcription, as PC-Kimmo is not capable of handling UTF-8 data. Each entry in a lexicon consists of the usual information for a morpheme: the underlying (or lexical) form of the morpheme; its lexicon name; possible continuation classes for subsequent morphemes; its English gloss; and features that describe, constrain, or pertain to the morpheme in question.

All of this lexical information is indispensable for the morphological engine. Fortunately, though, we do not need to hand-code this information, though the first version of the dictionary lexicon entries was in fact hand-coded. Now that an XML version of the dictionary is available, we are able to take the TEI XML-encoded version of the Lushootseed Dicitonary and extract the relevant material, converting it as necessary for use as dictionaries in the PC-Kimmo system. In this way a centralized lexical resource serves as a tool in its own right, but secondarily as input to the PC-Kimmo engine. Most of the words in the language can be treated by the engine; only a few complex and infrequent morphological configurations cannot be handled.

Since the morphology engine is useful for a wide variety of annotation tasks, we have developed a web interface that supports relevant processing. A user can enter a word in Romanized form and have the engine parse it as described above. The result can then be fed into further processing, as described in the next subsection.

2.2 Syntactic parsing

Although Lushootseed morphological structure is compelling in its flexibility and complexity, clause and sentence structure also exhibit striking properties. Though some pedagogical grammars exist for the language (Hess, 1995; Hess, 1998; Hess, 2006), no systematic grammatical exploration of the language's syntax has yet been published.

Computerized syntactic parsers are particular types of engines that have been developed to take sentences from a language and analyze their syntactic structure, outputting representations of the constituents of the sentence and their relationships. Most are closely (or even inalienably) associated with a particular linguistic theory so that principled coverage of grammatical phenomena can be assured. More recent engines have been developed with statistically based approaches, but they are only helpful when trained on a sizable corpus of the language to be parsed.

Using traditional parsing approaches for Lushootseed is problematic. On the one hand—as mentioned above—the language has not been thoroughly described in any extant theory of syntax, so theory-dependent parsing is not yet possible. On the other hand, no collection of the language is sizable enough to train statistically-based parsers on the types of constructions to expect.

In our work on Lushootseed we have chosen a different kind of parser (Lonsdale, 2005). Called the Link Grammer parser (Sleator and Temperley, 1993), it was developed for efficient processing of dependency-like syntax (Grinberg et al., 1995). Freely available for research purposes, it is more robust than traditional parsers and has been widely used in such NLP applications as information retrieval, speech recognition, and machine translation[3]. Written in the C programming language, it is comparatively fast and efficient.

The Link Grammar parser does not seek to construct constituents in the traditional linguistic sense—instead, it calculates simple, explicit relations between pairs of words. A link is a targeted relationship between two words and has two parts: a left side and a right side. For example, links associate such word pairs as: subject + verb, verb + object, preposition + object, adjective + adverbial modifier, and auxiliary + main verb. Each link has a label that expresses the nature of the relationship mediating the two words. Potential links are specified by a set of technical rules that constrain and permit word-pair associations. In addition, it is possible to score individual linkages and to penalize unwanted links. Though the formalism for describing links and their participants is unfamiliar to linguists, the system is well documented and customized language-specific grammars can be developed. Figure 2 shows a small set of sample link declarations from the Lushootseed grammar.

We take output from the PC-Kimmo process as described in the previous subsection, so that morphemes can be separated and annotated with basic information (i.e. their status as a prefix, root, suffix, or bound lexical morpheme).

[3] For a bibliography see http://link.cs.cmu.edu/link/papers/

```
<pref-asp1>: {(PRF- or STV- or PRG-)};
<pref-asp2>: {HAB-} & {DUB-} & {AD-};
<predprefs>: {NZ-} & {<pref-asp1>} & {SX-} & {<pref-asp2>} & {(FUT- or PT-)};

<root-main>:  <predprefs> & {DT-} & {LX+} & {BNF+} & {TX+} & {TC+} & {ACH+} &
              {TC+} & {TX+} & {ASP+};

<pred1>: ((<root-main> & <main-args>) or (<root-middle> & <middle-args>) or
          (<root-ditrx> & <ditrx-args>)) & {Wd-};
```

Figure 2: Sample Link Grammar rules.

This output is then fed into the Link Grammar parser so that links can be set up not only between words in a sentence, but also between the morphemes and clitics and their associated roots, so that morphological relationships can also be annotated. Figure 3 shows three sample Lushootseed sentences with their associated Link Grammar parses. Though a detailed examination of each link is beyond the scope of this paper, the link types generally reflect morphological markers such as aspect, valency changes, possession, and reduplication as well as syntactic relations such as modification, complementation, and arguments.

```
linkparser> ?u+ da?a +d ?ElgWE? ?E kWi s+ gWistalb ti?E? SukWE?.

    +-----------------------------Xp----------------------------+
    |           +--------------------SOo--------------------+   |
    |           +------EX------+------P-----+             |   |
    +-----Wd----+---SOs--+     |  +----DT---+             |   |
    |      +-PRF+-TX+    |     |  |   +--NZ-+      +--DT--+   |
    |      |    |   |    |     |  |   |     |      |      |   |
LEFT-WALL ?u+ da?a +d ?ElgWE? ?E kWi s+ gWistalb ti?E? SukWE? .

linkparser> bE+ Lil +t +Eb +ExW ?ElgWE? ?E ti?E? bE+ ?Es+ istE?.

    +-----------------------------Xp----------------------------+
    |          +-----------EM----------+                       |
    |          +--------PA-------+     +---------P--------+   |
    |          +----ASP----+     |     |   +------DT------+   |
    +----Wd----+--MD--+    |     |     |   |    +----AD---+   |
    |      +-AD+-TX+  |    |     |     |   |    |    +-STV+   |
    |      |   |   |  |    |     |     |   |    |    |    |   |
LEFT-WALL bE+ Lil +t +Eb +ExW ?ElgWE? ?E ti?E? bE+ ?Es+ istE? .

linkparser> q'ili +t +Eb +ExW ?E ti?E? s+ ?il =aXad ?E ti?E? captain.

    +------------------------------Xp------------------------------+
    |       +-------EM-------+                                     |
    |       +-----ASP----+   +-----P-----+                         |
    |       +---MD--+    |   |   +---DT--+----MV---+-----P----+    |
    +---Wd--+-TX-+  |    |   |   |    +NZ+-LX-+    |   +--DT--+    |
    |       |    |  |    |   |   |    |  |    |    |   |      |    |
LEFT-WALL q'ili +t +Eb +ExW ?E ti?E? s+ ?il =aXad ?E ti?E? captain .
```

Figure 3: Sample Link Grammar sentence parses for three Lushootseed sentences.

The Link Grammar parser is well suited for Lushootseed for several reasons noted above: (i) it does not implement any particular linguistic theory;

(ii) it provides a low-level parse that targets morphosyntactic relationships, without any hierarchical constituency; and (iii) unlike most traditional parsers it presumes no traditional phrase structure rules or any type of movement.

The Link Grammar parser does require significant lexical resources. Again, though we initially created prototype dictionaries by hand, we are now able to automatically generate complete dictionaries from the TEI XML version of the canonical language dictionary. The encoded dictionary thus serves in yet another way as a repository for lexically-related information needed for downstream language processing tasks.

As with the morphological engine, we found the Link Grammar parser so useful that we created a web interface that allows users to access the parser via a browser, enter a morphologically parsed sentence, and retrieve the Link Grammar parse that is returned by the system.

3 A morphosyntactic database

In order to explore the usefulness of the annotations we have been describing, we resolved to apply morphological and syntactic processing methods to a sizable set of Lushootseed language data. We began by collecting a corpus for annotation from various sources:

- published stories told by Ruth Sehome Shelton (Hilbert and Hess, 1995)
- transcribed sentences from liturgical materials translated by a 19th century Oblate missionary, Father Chirouse (Lonsdale, 2011)
- sentences from the existing pedagogical grammars mentioned above
- sentences from Lushootseed Dictionary usage examples

Five hundred sentences of varying length and complexity were randomly selected from these sources. Each sentence was Romanized and sent through the PC-Kimmo morphological engine and the Link Grammar syntactic parser. The output was collected in linearized format (as opposed to the graphical link parse format illustrated in Figure 3 above. Figure 4 shows a small sampling of the linearized links produced from parsing one of the 500 sentences.

```
1 tu+   PT  1 <---PT----> 2 PT  LC.r
1 LC.r  SOs 2 <---SOs---> 5 SO  ship
1 LC.r  ACH 2 <---ACH---> 3 ACH +il
1 kWi   DT  4 <---DT----> 5 DT  ship
1 .     RW  6 <---RW----> 7 RW  RIGHT-WALL
```

Figure 4: Sample links for a parsed Lushootseed sentence in linearized format.

Through straightforward data manipulation, all of the links from the 500 sentences (2143 links in total) were uploaded into a relational database. This permitted queries via the SQL query language to retrieve facts about the

morphosyntactic composition of these sentences as determined by the morphological and syntactic parses. Given knowledge of how to structure SQL queries, the following are samples of linguistic constructions that can be retrieved from the annotations database:

- sentences that have a negative and an aspectual marker
- sentences that have two oblique complements
- questions with a past tense
- finding out which is more commonly used: perfective or non-perfective verbs
- finding verbs with irrealis prefixes and/or out-of-control suffixes

Figure 5 gives two examples of the results from such queries: the top one shows the sentence with the longest link (in this case attaching a predicate with its subject); the bottom example shows the most morphologically complex predicate from all of the sentences.

```
Longest link:

linkparser> huy lEk'W +t +Eb +ExW ?E ti?E? dxWsT'alb ti?E? lEpEskWi.

    +-----------------------------------Xp-----------------------------------+
    |              +-------------------------PA-------------------------+    |
    |              +--------EM-------+                                   |    |
    |              +-----ASP-----+   |                                   |    |
    |              +---MD---+    |   +-------P------+                  |    |
    +---Wg--+--Wd--+--TX-+  |    |   |    +----DT---+         +---DT---+    |
    |       |      |     |  |    |   |    |         |         |        |    |
LEFT-WALL huy.a lEk'W.r +t +Eb +ExW ?E ti?E?.d dxWsT'alb.r ti?E?.d lEpEskWi .

Most complex predicate:

    T?u+ tu+ s+ takW +yi +t +Eb +s
```

Figure 5: Results from two queries about Lushootseed sentence structure.

Of course, basic statistics are also computable from the information uploaded into the database. Consider, for example, these overall statistics from the parsed corpus:

```
Sample statistics (tokens)
* # sentences: 500
* # words: 1625
* # morphemes: 2954
* # suffixes: 623
* # prefixes: 607
* # S's with only monomorphemic words: 43
```

```
 10 =bixW                  777 punctuation
  3 =igWEd                 763 determiner
  2 =ELdat                 629 PP
  2 =a?kW                  618 subject
  2 =aCi?                  528 PP-object
  2 =aXad                  380 aspectual
  2 =ali                   323 transitive
  1 =abac                  228 middle
  1 =al?txW                205 stative
  1 =alikW                 193 nominalizer
  1 =alus                  188 past
  1 =aq                    122 adverbial (sentential)
  1 =gWas                   99 achievement
  1 =gWiL                   97 perfective
  1 =gWil                   85 possessive
  1 =i                      82 future
  1 =iC                     73 lexical suffix
  1 =qid                    66 oblique
  1 =ucid                   61 habitual
                            53 subordinating
                            51 adverbial (predicate)
                            45 passive
                            39 dubitative
                            32 benefactive
                            29 progressive
                            29 causative
                            15 object
                            10 adjective
                             8 determiner (feminine)
                             7 partitive
                             2 reflexive
```

Figure 6: Statistics for structural properties as derived from links in the 500-sentence corpus.

Similarly, Figure 6 displays more detailed statistics on structural information from the 500-sentence corpus; lexical suffixes and their frequencies are listed on the left, and general linkage type statistics are displayed on the right.

4 Conclusions and future work

In this paper we have illustrated how various language resources and tools can be pipelined together to provide annotation and analysis capabilities for Lushootseed sentences. An XML-annotated dictionary can serve as a fundamental lexical resource for direct access by users, and it can also provide lexical and morphological information to morphological and syntactic parsers. A finite-state engine can parse and generate morphological structure for Lushootseed words. The Link Grammar parser can analyze Lushootseed sentences and label the morphosyntactic relationships inside them. The results can be uploaded into a relational database and queried using standard means to retrieve characterizations, statistical and otherwise, of the annotated sentences.

Although this work constitutes what would appear to be the most extensive analysis of Lushootseed structure to date, much progress remains.

For example, for this prototype implementation the 500-sentence sampling in this work was piecemeal and random; scaling up the approach would require using a more systematic corpus for annotating.

The approach also requires some knowledge of SQL scripting to query the database for annotation facts; this is difficult for non-technical users. We are experimenting now with a more user-friendly database management system and web interface that would allow for specification of queries in a more lingustically-grounded manner.

Finally, we intend to integrate these tools into an eventual web portal that will assemble the language resources and tools discussed—and indeed others not described here–so that outside users can use them to explore their own Lushootseed words, sentences, passages, and texts.

References

Antworth, E. (1990). *PC-KIMMO: a two-level proecssor for morphological analysis*. Number 16 in Occasional Publications in Academic Computing. Summer Institute of Linguistics, Dallas, TX.

Bates, D., Hess, T., and Hilbert., V. (1994). *Lushootseed Dictionary*. University of Washington Press.

Bates, D. and Lonsdale, D. (2010). Recovering and updating legacy dictionary data. In *Proceedings of the 44th Annual International Conference on Salish and Neighboring Languages (ICSNL)*, volume 27 of *University of British Columbia Working Papers in Linguistics*, pages 1–12.

Grinberg, D., Lafferty, J., and Sleator, D. (1995). A robust parsing algorithm for Link Grammars. Technical Report CMU-CS-95-125, School of Computer Science.

Hess, T. (1976). *Dictionary of Puget Salish*. University of Washington Press.

Hess, T., editor (1995). *Lushootseed reader with introductory grammar*, volume 1 of *University of Montana Occasional Papers in Linguistics No. 10*. University of Montana, Missoula, MT.

Hess, T., editor (1998). *Lushootseed reader with introductory grammar*, volume 2 of *University of Montana Occasional Papers in Linguistics No. 14*. University of Montana, Missoula, MT.

Hess, T. (2006). *Lushootseed Reader with English Translations. Volume III: Four More Stories from Martha Lamont*, volume 19 of *University of Montana Occasional Papers in Linguistics*. University of Montana.

Hilbert, V. and Hess, T. M. (1995). *siastEnu: 'Gram' Ruth Sehome Shelton*. Lushootseed Press, Seattle, WA.

Koskenniemi, K. (1983). Two-level model for morphological analysis. In *Proceedings of the 8th International Joint Conference on Artificial Intelligence*, pages 683–685.

Lonsdale, D. (2001). A two-level morphology engine for Lushootseed. *Proceedings of the 36th International Conference on Salishan and*

Neighbouring Languages, 6:203–214.
Lonsdale, D. (2003). Two-level engines for salish morphology. In *Proceedings of the Workshop on Finite-State Methods in Natural Language Processing, 10th Conference of the European Chapter of the Association for Computational Linguistics*, Budapest. Association for Computational Linguistics (ACL).
Lonsdale, D. (2005). A link grammar parser for Lushootseed. In *40th International Conference on Salish and Neighboring Languages*, Musqueam Nation; Vancouver, BC. (Presentation).
Lonsdale, D. (2011). Transcribing a Salish catechism into modern orthography. In *Conference on Endangered Languages and Cultures of Native America*, University of Utah, Salt Lake City, UT. (Presentation).
Sleator, D. and Temperley, D. (1993). Parsing English with a link grammar. In *Proceedings of the Third International Workshop on Parsing Technologies*.