LANGUAGE DOCUMENTATION & CONSERVATION

Vol. 8 (2014), pp. 1-37
http://nflrc.hawaii.edu/ldc
http://hdl.handle.net/10125/4604

Using TEI for an Endangered Language Lexical Resource: The Nxaʔamxcín Database-Dictionary Project

Ewa Czaykowska-Higgins, Martin D. Holmes, Sarah M. Kell
University of Victoria

This paper describes the evolution of a lexical resource project for Nxaʔamxcín, an endangered Salish language, from the project's inception in the 1990s, based on legacy materials recorded in the 1960s and 1970s, to its current form as an online database that is transformable into various print and web-based formats for varying uses. We illustrate how we are using TEI P5 for data-encoding and archiving and show that TEI is a mature, reliable, flexible standard which is a valuable tool for lexical and morphological markup and for the production of lexical resources. Lexical resource creation, as is the case with language documentation and description more generally, benefits from portability and thus from conformance to standards (Bird and Simons 2003, Thieberger 2011). This paper therefore also discusses standards-harmonization, focusing on our attempt to achieve interoperability in format and terminology between our database and standards proposed for LMF, RELISH and GOLD. We show that, while it is possible to achieve interoperability, ultimately it is difficult to do so convincingly, thus raising questions about what conformance to standards means in practice.

> If digital language documentation and description should transcend time, they should also be reusable in other respects: across different software and hardware platforms, across different scholarly communities... and across different purposes. (Bird & Simons 2003: 558)

1. INTRODUCTION. Since Hale et al.'s (1992) collection on the seriousness of language loss appeared in *Language* there has been increased discussion in linguistics about the need for endangered language[1] documentation to create language resources, such as lexica, that are both enduring and reusable in various formats and for various purposes, including for scientific linguistic research and for language maintenance and revitalization. As many linguists would agree, and as the quote from Bird and Simons (2003) above suggests, the creation of reusable lexical resources, "...benefits from conformance to established standards" (Thieberger 2011: 463). Not surprisingly, therefore, the expanding use of technology in endangered language documentation has increased interest in and discussion about the development of resources and, correspondingly, about the development of established digital standards for lexical information. In this paper, we contribute to these discussions by describing one lexicon project from its beginnings in 1991 as a database, constructed using Lexware (Hsu 1985), DOS and WordPerfect 5, to a web-based database encoded using TEI (Text Encoding Initiative) XML markup.

[1] Since the term 'endangered language' has gained wide currency in linguistics over the last twenty years we use it here. See Hill (2002), Errington (2003), Perley (2012) for critical responses to the rhetorics of language endangerment.

E-ISSN 1934-5275

Digital lexical resources developed by linguists working on endangered languages have been reported to make use of various kinds of lexicon-building tools, including *Lexique Pro*, SIL's *Toolbox* or *FLEx*, and, more recently, *LEXUS-ViCoS*. Few endangered language lexical resources are reported to use TEI markup language (but see Bates and Lonsdale 2010; Thieberger 2013), even though TEI is an established Digital Humanities standard with a dictionary module (TEI Guidelines, ch. 9; Romary and Wegstein 2012). In this paper, by outlining and exemplifying our use of TEI in the construction of a lexical resource for Nxaʔamxcín, an Interior Salish language spoken historically in Washington State, USA, we take a step towards filling the gap in reporting.

The primary goal of this paper is to illustrate that TEI is a valuable tool for creating lexical resources which are both enduring and reusable. We show that as an open, mature standard, TEI is a useful encoding strategy for lexical material, providing a reliable archival format for Nxaʔamxcín words and phrases, as well as flexibility for encoding the complex morphology and morphological relationships found in a Salish language.

Nevertheless, even a project using TEI faces significant challenges when attempts are made to conform to recent standard-harmonization or interoperability initiatives in linguistics such as E-MELD, GOLD, LEGO, or RELISH. A second goal of this paper, therefore, is to exemplify the kinds of challenges that standard-harmonization leads to. We distinguish two types of challenges: those involving what we refer to as format interoperability and those involving terminological interoperability.[2] In this paper, we approach the issue of format interoperability through consideration of convertibility between TEI P5 XML and Lexical Markup Framework (LMF). We discuss terminological interoperability through consideration of the challenges faced in attempting to align concepts and terminology used in our project with those used in the General Ontology for Linguistic Description (GOLD), as expressed in the ISOcat Data Category Registry. While other researchers address these issues more abstractly (Broeder et al. 2011, Romary (forthcoming)), our paper presents a case study involving a substantial language database.

Our paper is organized as follows. In §2 we outline the scope and history of the Nxaʔamxcín lexical resource project. In §3 we discuss the structure of entries in a TEI XML database, and provide evidence of the suitability of TEI for a project of this kind. In §4 we turn to the challenges of conformance with digital standards for interoperability. It is here that we distinguish and discuss two types of interoperability, and show that while it is possible in principle to comply with digital standards, in practice this is a more difficult goal. Ultimately, we conclude that true conformance to digital standards and interoperability is an important long-term goal in producing a legacy lexical resource, but in practice it can be sufficiently time-consuming that it may not be the most pressing priority when trying to produce a resource that serves the urgent needs of a language community working to combat language loss.

2. THE PROJECT HISTORY: FROM FILECARDS TO LEXWARE, FROM LEXWARE TO TEI. The late M. Dale Kinkade worked with 22 Nxaʔamxcín language speakers, some of whom were in their 80s and 90s at the time. Most of this work was completed over a period of 10 years from the late 1960s to the late 1970s, with some additional recording done in the early 1990s. Kinkade's record of Nxaʔamxcín thus contains the most extensive attested

[2] Bird & Simons (2003: 563) discuss the need for portability in format and terminology. See §4 below.

material of the language recorded by a linguist. As was typical of the time, Kinkade not only recorded sessions with language speakers on reel-to-reel tape and in field notebooks, but as his understanding and linguistic analysis of Nxaʔamxcín progressed, he also systematically categorized the Nxaʔamxcín materials and recorded this categorization on thousands of filecards. The filecards reflect examples of directly elicited words and sentences (see below for description of filecard headwords). Kinkade's direct elicitation included systematic elicitation of paradigms and derivational morphological categories, listing of personal and place names, and elicitation of vocabulary from different semantic domains, as well as recording of texts and stories; some words and sentences on the filecards come from the stories, but Kinkade seems not to have made filecards based on all the texts he recorded.

In this section we briefly discuss the kinds of lexical resources that have been developed based on these filecards. The history of attempts to create lexical resources is not unique to Nxaʔamxcín, but in fact parallels histories of similar projects for other indigenous languages studied by linguists. We present this history here in order to provide a context for the issues we discuss in §3 and §4.

2.1 LEXICAL RESOURCES. In 1981, Kinkade and the Colville Confederated Tribes published a dictionary in the form of a word-list of Nxaʔamxcín (Kinkade 1981). This dictionary is based on a subset of common wordforms selected by Kinkade from the filecards and is organized alphabetically. Kinkade always had the intention of publishing in print a larger, comprehensive morphologically-focused dictionary, similar to his dictionary of Upper Chehalis (Kinkade 1991) which, like various other dictionaries of Salish languages, is organized such that morphological roots and unanalyzable stems serve as headwords. In 1991, Kinkade and Ewa Czaykowska-Higgins decided to work together to compile this more comprehensive dictionary of Nxaʔamxcín based on Kinkade's filecards, with Kinkade serving as consultant and Czaykowska-Higgins leading the project. This work was supported by Nxaʔamxcín-speaking community members that Kinkade and Czaykowska-Higgins were working with in 1991.

In 1991, many Salish scholars were using Lexware, a print dictionary-making format developed by Robert Hsu of the University of Hawai'i (Hsu 1985). Kinkade's filecard database was thus input into a computer database using a system involving a combination of Lexware and WordPerfect, running on DOS. In addition to inputting the language materials, Kinkade and Czaykowska-Higgins aimed to enrich the data by providing detailed morphological breakdowns for all lexical items as they were being entered into the Lexware database (Figs. 3 and 4 below illustrate the morphological complexity of lexical items).

By the end of the 1990s, most of the data[3] had been entered into the computer database, but the project slowed down. The database was left stored on an old desktop computer, but eventually this computer refused to boot, and then rescuing the database became urgent. Consequently, in 2003, Czaykowska-Higgins began working with several programmers at the University of Victoria, including Greg Newton and Martin Holmes,

[3] Given the technical nature of this paper, we use terms like 'data' and 'harvest' when discussing the database and language examples. The language materials are a record of the linguistic heritage of Colville Tribes, however, and from that perspective are much more than 'data.'

to rescue the Lexware database from the old hardware, and the WordPerfect files were retrieved from the computer hard drive. In order to prevent the need to rescue the data a second time, we decided to avoid dependency on specific hardware or software programs, and therefore to adopt an established widely-supported standard, which in our case meant converting the Lexware database into TEI XML. (See below for more discussion of the reasons for this choice.) An additional consideration in the development of the TEI database, independent of the standards issues, was that we wanted the database to be flexible enough to allow it to be transformed into different kinds of formats (web-based or print) for different kinds of users. A final consideration has been that Kinkade passed away in 2004. Czaykowska-Higgins has thus been particularly concerned to ensure that the database not only accurately reflects and honors the knowledge and understanding of the speakers who worked with Kinkade, but also that it reflects Kinkade's interpretation and understanding of Nxaʔamxcín as evidenced on his filecards and in his later writing on the language.

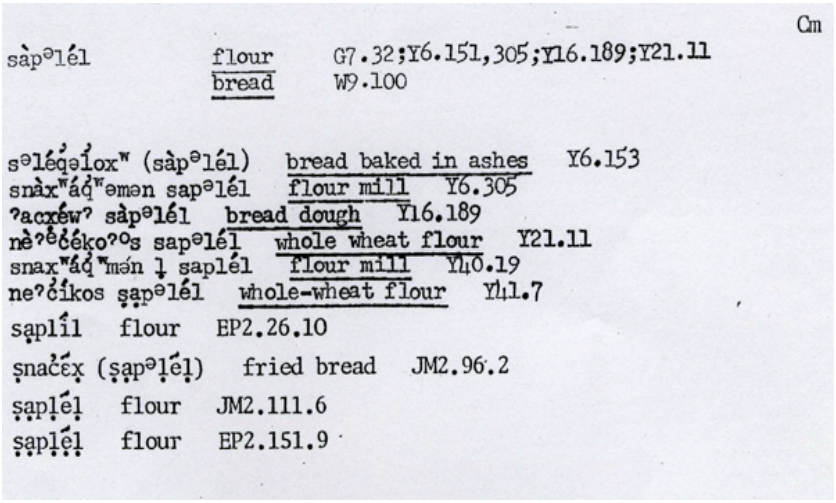
Between 2004 and 2010, the team worked on the project only periodically, as time permitted. In early 2010, Sarah Kell joined the project part-time as co-editor and database tester. By late 2010 we had produced a searchable, though not fully edited, online version of a Nxaʔamxcín-English, English-Nxaʔamxcín database in TEI containing over 10,000 entries (including postulated root morphemes), sentence examples, and rich morphological information.[4] Since then, we have been working on completing the editing of the language examples and on refining the coding. In addition, since 2010 we have been working with members of the Nxaʔamxcín Language Program and History and Archives of The Confederated Tribes of the Colville Reservation to attempt to ensure that the final shape of the database and the various projected outputs produced from it are guided by the needs of learners, teachers and speakers of Nxaʔamxcín, and not only by expectations and assumptions of the linguists and programmers involved in the project. One additional consideration governing the project has been that the Nxaʔamxcín Language Program, collaborating with linguist Nancy Mattina, has also worked on a dictionary in print format of Nxaʔamxcín, using the 1981 Kinkade-Colville Tribes word list mentioned above as a foundation. We have thus attempted to ensure that the TEI database we are producing complements the work of the Language Program as much as possible. Finally, the completion of this project has been made particularly urgent by the loss since 2011 of several key fluent speakers. When we first began writing this paper in the fall of 2012, there were two fluent elders left who were able to be active in language work. Now (spring 2013) there is only one.

2.2 FROM FILECARDS TO THE COMPUTER. In this section and in §2.3 we illustrate the stages of the project, and explain in more detail why we chose to use TEI when the Lexware database was rescued from obsolete hardware. The initial phase of the project, putting Kinkade's filecards into a Lexware database, was a huge undertaking since there are thousands of filecards. Figures 1 and 2 provide examples of two of the original filecards from Kinkade's file boxes. Figure 1 illustrates a monomorphemic stem based on a loanword from French; Figure 2 illustrates an example of a card for a root morpheme and polymorphemic forms derived from that root. As the two figures show, the headwords on the filecards are provided in the top left-hand corner. Usually, they are single morphemes,

[4] The online database is currently password protected; the XML data are stored in a Subversion version-control system which preserves their entire editing history.

as illustrated in the two figures given below, but they can also be polymorphemic words. Single morpheme headwords include root morphemes that can stand alone as independent words, root morphemes that are inferred from a set of words which contain them, particles or affixes. Filecards often contain lists of one or more words or phrases illustrating the headword. While there is (usually) only one filecard (or set of cards) for each morpheme and each simple or polymorphemic word, the same polymorphemic word can be found on more than one filecard since it can be listed under the root, and any of the various affixes it contains. In this sense, then, there are duplicate examples in the data, but not duplicate headwords.

In Figure 1 we see cases of collocations involving the headword /*ṣaṗl̓il*/. We also see different levels of transcription, from the highly detailed, more impressionistic narrow transcriptions closer to the top of the card, to broader transcriptions reflecting Kinkade's shifting interpretation of the phonemicization of the form, as well as details of pronunciation and variant pronunciations. In addition, we see different glosses and translations provided by different speakers. Kinkade and Czaykowska-Higgins wanted the database to be structured in such a way as to allow it to reflect adequately these different levels of interpretation, and the different types of information that Kinkade had recorded and that the fluent speakers of Nxaʔamxcín had provided to him.


Cm

sàpᵊlél flour G7.32;Y6.151,305;Y16.189;Y21.11
bread W9.100

sᵊlé́qə́lox̣ʷ (sàpᵊlél) bread baked in ashes Y6.153
snàxʷáq̓ʷəmən sapᵊlél flour mill Y6.305
ʔacx̣éwʔ sàpᵊlél bread dough Y16.189
nèʔᵊčéḳoʔᵒs sapᵊlél whole wheat flour Y21.11
snaxʷáq̓ʷmán ɬ saplél flour mill Yl0.19
neʔčíkos ṣaṗᵊlél whole-wheat flour Yl1.7
ṣaṗlíl flour EP2.26.10
ṣnačéx̣ (ṣaṗᵊl̓é) fried bread JM2.96.2
ṣaṗl̓é flour JM2.111.6
ṣaṗl̓é flour EP2.151.9

FIGURE 1. Snapshot of a Kinkade filecard illustrating a monomorphemic stem

Figure 2 illustrates a case where the headword is a root morpheme, *√sən,* that occurs in various polymorphemic words. Thus we see, in addition, the type of morphological complexity that is found in Nxaʔamxcín, and that Salish languages more generally are known for. Morphosyntactic categories which can be marked on words in Nxaʔamxcín include number, aspect, transitivity (voice and valence), control, and person marking (possessors, objects and subjects). In addition, although the lexical content of words is situated primarily in root morphemes, as one would expect, Nxaʔamxcín also has a large class of approximately 90 lexical suffixes, whose meanings refer to concrete nominal concepts such

as body parts, or to more abstract notions. Within Nxaʔamxcín words, morphemes from particular morphosyntactic categories occupy particular positions with respect to roots and to each other. Finally, nonconcatenative morphology is common, including five types of reduplication as well as infixation.[5]

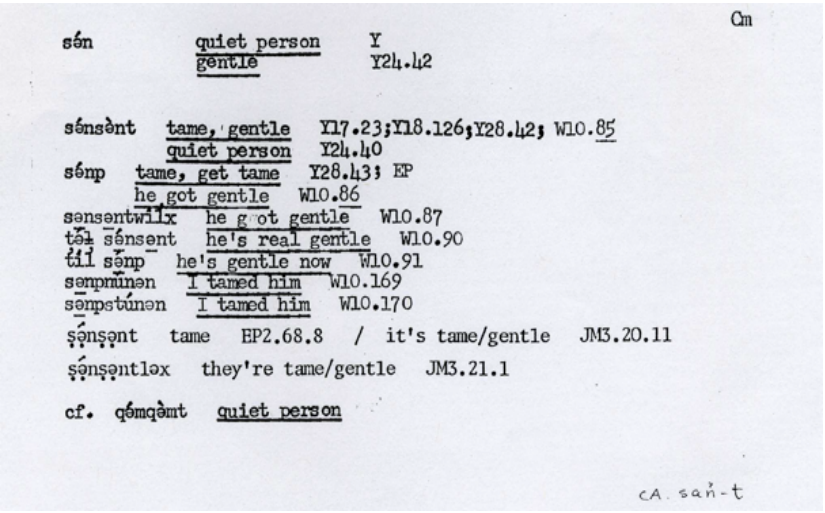


FIGURE 2. Snapshot of a Kinkade filecard for a root and forms derived from the root

In the example filecard in Figure 2, therefore, we can see, in addition to the different levels of interpretation seen in Figure 1, words containing reduplication (see Figure 3), various types of suffixation with cases of multiple suffixes in one word (see Figure 4), a cross referenced form based on a different root, and, finally, a note that Kinkade penciled in indicating a Coeur d'Alene (Interior Salish) cognate of the Nxaʔamxcín form.

ṣénṣənt tame EP2.68.8 / it's tame/gentle JM3.20.11

FIGURE 3. Example containing a reduplicative 'characteristic' suffix and the suffix *-t* 'stative'

senpstúnen I tamed him W10.170

FIGURE 4. Example containing *-p* 'inchoative', *-stu* 'causative' and *-n* 'control.transitive-transitivizer-3Object-1sgSubject'

The different types of information regarding speakers, multiple meanings, multiple transcriptions, multiple levels of interpretation, cognates, and cross references, combined with the complex morphological structure of Nxaʔamxcín words meant that the construc-

[5] See Czaykowska-Higgins and Kinkade 1998 on Salish; Czaykowska-Higgins 1998, Willett 2003 and references therein on Nxaʔamxcín morphology.

tion of the initial Lexware database itself was not only time-consuming, but also highly complex.

2.3 THE LEXWARE FORMAT. The Lexware database used a BAND format, including bandnames, headwords, numbering, and spaces to provide a hierarchical system of embedding sub-entries within entries (see Hsu 1985). The structure of the database itself therefore required Kinkade, Czaykowska-Higgins and the research assistants working with them to make interpretive decisions about morphological relationships and breakdowns and to encode those interpretive decisions through the structure of each entry. Figure 5 illustrates the structure of the Lexware format used to encode the morphological properties of the word *sənpnúnən* 'I tamed him' whose constituent structure includes a root morpheme *√sən,* as well as the suffixes *-p-nún-ən*.[6] As Figure 5 shows, the word is analyzed as a sub-sub-entry, the inflected form of the sub-entry stem *sənp* 'tame, get tame, he got gentle' which is itself derived from the entry headword *sən* 'quiet person, gentle'.

Headword = Entry

```
.rt   √sən
g   *quiet person, *gentle
k   Y; Y24.42
```

Derived Form = Sub-entry

```
..in   √sə́n-p
g   *tame, get tame
k   Y28.43; EP
var   √sə́n-p
g   he got gentle
k   W10.86
```

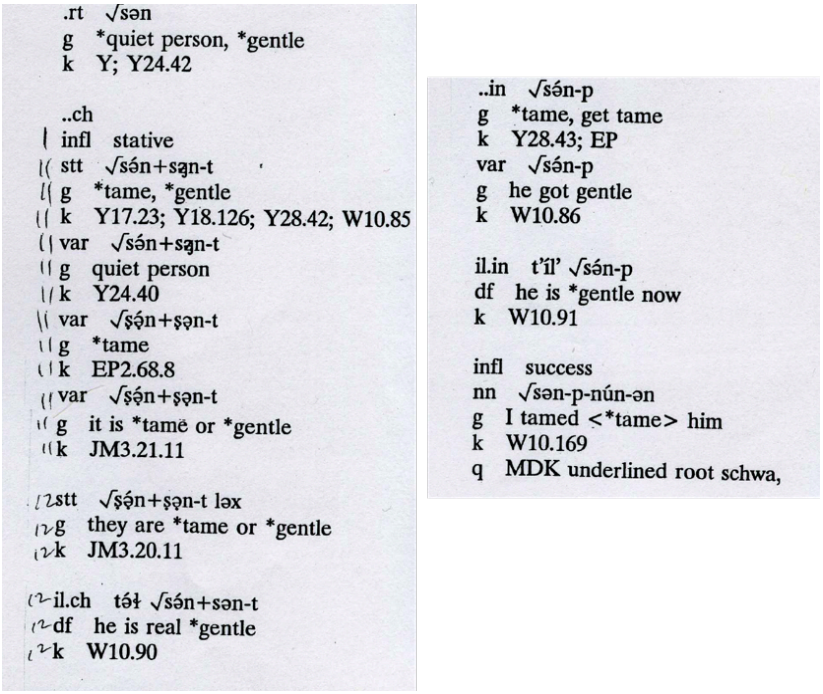
Inflected Derived Form = Sub-sub-entry

```
infl   success
nn   √sən-p-nún-ən
g   I tamed <*tame> him
k   W10.169
q   MDK underlined root schwa,
```

FIGURE 5. The structure of a Lexware entry

Figure 6 illustrates the complete entry for the root *√sən*, including all the different sub- and sub-sub-entries of which it is composed.

[6] The constituent -ən is the surface form of the string -n-t-Ø-n 'control.transitive-transitivizer-3Object-1sgSubject' (see Kinkade 1982).



.rt √sən
g *quiet person, *gentle
k Y; Y24.42

..ch
infl stative
stt √sə́n+sa̱n-t
g *tame, *gentle
k Y17.23; Y18.126; Y28.42; W10.85
var √sə́n+sa̱n-t
g quiet person
k Y24.40
var √sə́n+sən-t
g *tame
k EP2.68.8
var √sə́n+sən-t
g it is *tamë or *gentle
k JM3.21.11

stt √sə́n+sən-t ləx
g they are *tame or *gentle
k JM3.20.11

il.ch tə́ɬ √sə́n+sən-t
df he is real *gentle
k W10.90

..in √sə́n-p
g *tame, get tame
k Y28.43; EP
var √sə́n-p
g he got gentle
k W10.86

il.in t'ə́l' √sə́n-p
df he is *gentle now
k W10.91

infl success
nn √sən-p-nún-ən
g I tamed <*tame> him
k W10.169
q MDK underlined root schwa,

FIGURE 6. The entry for the root *√sən* and polymorphemic words based on this root

We see that the bandnames for each line of the entry provide the actual mark-up of the entry line: thus, for example, *g* is used to refer to the gloss line, *k* to information about the speaker that provided the relevant example word or phrase, *var* to a variant pronunciation or gloss, and *q* to a question provided by the coder; abbreviations like *nn* and *in* provide the morphological information. The numbering indicates inflectional subtypes within the entry.

To complete the example, Figure 7 illustrates the entire entry for the stem */ṣaplil/*. In this case, although the stem does not have a complex morphological structure, it does have many variant transcriptions, and several examples of collocations in which it occurs:

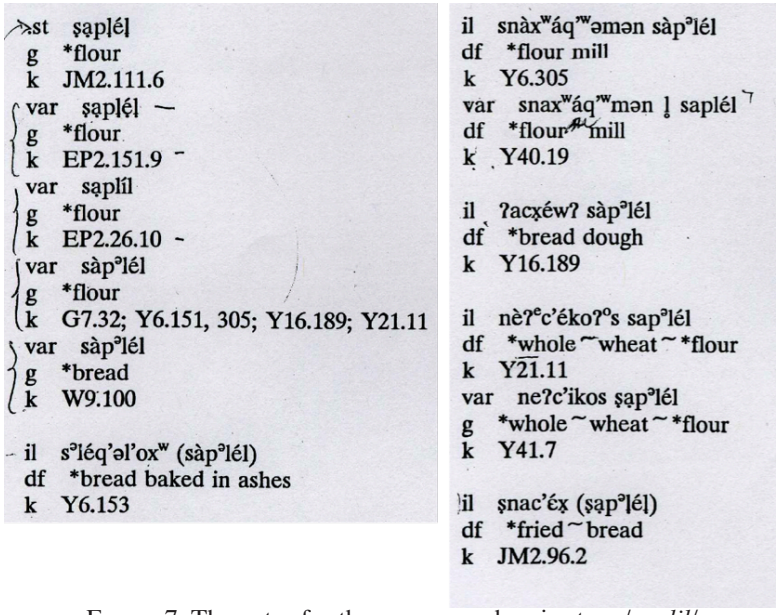


FIGURE 7. The entry for the monomorphemic stem */ṣaplḭl/*

The exemplification of the lexical materials and properties encoded in the Lexware database and the structure of the entries illustrate two important points. First, the BAND format of the Lexware software itself uses a 'field-oriented markup' where the content of each field is explicitly defined and in which the relationships between the different fields are explicitly encoded following a standard format, which makes use of numbering, spaces between lines, band names, sequencing of bands, and punctuation like one or more periods (amongst other devices) to encode entries, sub-entries, and the relations between them. The creation of such a consistently structured and standardized database[7] (which Thieberger (2011: 464) points out resembles the kind of markup used by later lexicon-making software) has meant that any subsequent work involving the database, including its conversion, has been rendered much more straightforward than it might otherwise have been.

Second, although the structure imposed by the software was consistent and explicitly defined, the materials from Kinkade's filecards that the database encoded are not themselves consistent. In other words, because the database is constructed from fieldnotes, its contents are determined by those forms which speakers provided to Kinkade and which he recorded in his fieldnotes. The database is thus constructed from *attested* forms and necessarily includes gaps of various kinds: for example, while Kinkade generally recorded transitive verbs in their 1st person singular subject – 3rd person object forms, some transitive verbs are only attested in other subject-object combinations from the paradigms. Consequently, the format of the database not only needed to be consistent and structured to allow for reusability and conversion to various formats, it also needed to be flexible since lexical entries were not necessarily always complete in consistent ways. This tension between

[7] Lexware allows for variability in how one chooses to structure it. Our use of it was systematic.

consistency in structure and incomplete consistency in the recorded materials has played a role in the subsequent conversion of the Lexware format into TEI, as well as in determining the kinds of lexical resource that can ultimately be produced from the recorded materials.

2.4 ONCE BITTEN, TWICE SHY. It is not unusual for slow-moving lexicon projects to be overtaken by substantial technological shifts, leaving their data stranded in an unsupported format. The original Nxaʔamxcín print-dictionary project, begun using Lexware (Hsu 1985), WordPerfect, and DOS, was carefully thought out and worked very well, as illustrated above, but it was dependent on specific resources (customized character-sets, printer fonts, macros, and a Hercules graphics card) which, although excellent at the time, nevertheless made the data difficult to use by the early 2000s.[8] The Nxaʔamxcín text consisted of incomprehensible sequences including non-printing characters, as shown in Figure 8:

```
.rt?    ÀEOTBELÀÀ1SOHÀkÀSOEOTÀ

1infl    transitive
11tr    ÀEOTBELÀÀ1SOHÀÀUSSOHÀkÀSOEOTÀ©À9SOHÀn
11g    *pull
q    should this be under a separate root?
```

FIGURE 8. Garbled data from Lexware/WordPerfect

Holmes and Newton (2008) describe in detail the work involved in programming of a custom Windows software application (*Transformer*) to transform the core data in binary WordPerfect files into a modern Unicode text representation.[9] Once this was done, transformation of the Lexware format to XML using Band2xml[10] provided a simple XML representation that could then be converted into something more future-proof. Bates and Lonsdale (2010) describe a similar process using Perl to rescue language data—coincidentally from a related language, Lushootseed, and also encoded using Lexware. Although the two projects were unaware of each other's work at the time, both faced the same crucial decision once the data had been retrieved: How should the rescued data be encoded to avoid the necessity for another such time-consuming and complicated operation in the future? Again coincidentally, both projects made the same decision: the data would be encoded in XML according to the guidelines of the Text Encoding Initiative. The Lushootseed data was encoded according to the TEI P4 DTD, which was current at the time, and was later

[8] Bob Hsu played a key role when Czaykowska-Higgins and Kinkade set up the initial Lexware Database for this project in the early 1990s. The Word Perfect fonts were developed and supported by Tim Montler, who also developed the program that we used to transform Lexware to XML. Our project is indebted to Bob Hsu and Tim Montler and to their earlier work.

[9] The transformation was complicated by the fact that there were various combining characters in the font sets which had been used inconsistently when the data were being input into Lexware. The resulting open-source application Transformer has been useful in numerous projects since. http://www.tapor.uvic.ca/~mholmes/transformer/.

[10] http://www.ling.unt.edu/~montler/convert/Band2xml.htm

converted to the newer P5; in the case of Nxaʔamxcín, we adopted the nascent P5 schema immediately, despite the fact that it was not yet finalized, and then made some adjustments to the encoding in later years as the guidelines evolved.

We approached the decision on encoding not only informed by common sense recommendations on best practices for interoperability, portability and open standards, as set forth, for instance, by Bird and Simons' seminal paper (2003), but also in the light of our recent experience in dealing with the consequences arising from earlier choices. We had been severely bitten, and were determined not to be so again. The adoption of Unicode was obvious, as was the choice of XML; both are governed by international standards, and widely used, not only throughout academia but also in the commercial world. Although the initial Lexware project's purpose was to produce a print dictionary, we used the rescue of the materials to create a web-based database, since this would allow for the largest number of possible uses of the lexical materials: they could become a searchable web-resource, or could be transformed into print formats of various kinds.

Choice of an XML encoding schema was harder than adoption of Unicode. Other possibilities such as the formats used by *Toolbox* and similar tools seemed to us at the time under-documented and somewhat less flexible than TEI P5. In fact, P5 was the only option that looked anything like a formal standard, and, partly because it is not discipline-specific, promised the flexibility and power that we anticipated needing as our project developed. Although TEI is a *de facto* standard—a community-sustained set of recommendations—rather than a *de jure* standard endorsed by an organization such as the ISO (International Organization for Standardization),[11] it has very broad support and acceptance, is widely used for born-digital documents, and provides a wide range of tags for dictionaries, linguistic analysis and corpus linguistics (*TEI Guidelines* chs. 15-18). At the time of our data-rescue effort, no other schema or set of encoding practices appeared to have the same level of broad support coupled with encoding power.

After we started our rescue project, Lexical Markup Framework, the ISO standard for NLP (natural language processing) and MRD (machine-readable dictionary) lexicons, emerged in 2008 as ISO 24613.[12] The goals of LMF "are to provide a common model for the creation and use of lexical resources, to manage the exchange of data between and among these resources...".[13] LMF could, in principle, have been a reasonable option for our project, had it been available to us at the time. However, as we show in the following sections, LMF would not in practice have served our project as well as TEI does. Indeed, it is notable that even since the emergence of LMF as an ISO standard, some large-scale dictionary projects have still opted for TEI; Budin, Majewski and Mörth (2012), for instance, describe how the Institute for Corpus Linguistics and Text Technology (ICLTT) of the Austrian Academy of Sciences has chosen TEI over LMF for its dictionary-encoding work: "Given the still small amount of available data using LMF and ongoing discussions, the decision was made to move towards TEI and keep an eye on the LMF specification as it develops" (Budin et al. 2012).

[11] See Stührenberg 2012.

[12] See http://www.lexicalmarkupframework.org/

[13] Cited from Wikipedia article on LMF; linked to from LMF website 2013/02/18.

3. THE STRUCTURE OF DICTIONARY ENTRIES IN TEI. In this section, we illustrate the structure of dictionary entries in TEI, and provide further justification for the usefulness of TEI as an encoding strategy for lexical material.

The TEI recommendations for encoding of dictionary entries are very straightforward, and in the Nxaʔamxcín database we follow them more or less exactly, although there are some slightly unusual aspects to our encoding at the lower levels. In our TEI encoding, an entry breaks down as shown in Figure 9 (explanations for each level are given below).

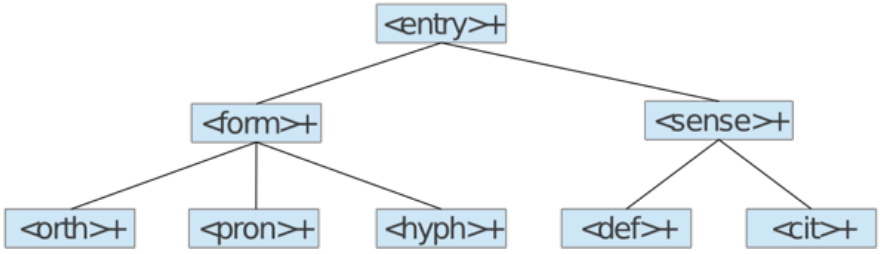


FIGURE 9. Basic TEI encoding

The plus signs indicate that all of these elements may appear more than once at the designated position. This basic structure is very similar to the abstract model of the Lexical Markup Framework, as we show in §4 below.

3.1 ENCODING DETAILS. Here we examine the components of our entry encoding in detail. The following sample encodings are all taken from the entry for *cahcímn*, 'I encouraged someone'. The complete entry appears in Figure 10.

```
<entry xml:id="cahcimn">
  <form>
    <orth>čahčímn</orth>
    <pron>
      <seg type="p" subtype="i">cahcímn</seg>
      <bibl corresp="psn:ECH">ECH</bibl>
      <seg type="n">cahcíməәn</seg>
      <bibl corresp="psn:W">W11.58</bibl>
    </pron>
    <hyph>√<m corresp="m:cah">cah</m>=<m corresp="m:cin">cí</m>-<m
    corresp="m:min m:t-TR m:Ø-OBJ m:n-SUBJ">mn</m>
    </hyph>
  </form>
<sense>
  <def>
  <seg>I <gloss>encourage</gloss>d someone</seg>
  <bibl corresp="psn:W">W11.58</bibl>
</def>
```

```
</sense>
<xr>See <ref target="m:ɕaḥcimn">ɕaḥcímn</ref>, and <ref
 target="m:n?iʕWqn">n?íʕʷqn</ref>.</xr>
</entry>
```

FIGURE 10. The entry for *cahcímn*

The <orth> element contains a straightforward representation of the form in the orthography used by the Nxaʔamxcín Language Program. The next component is the pronunciation element (<pron>):

```
<pron>
   <seg type="p" subtype="i">cahcímn</seg>
   <bibl corresp="psn:ECH">ECH</bibl>
   <seg type="n">cahcímən</seg>
   <bibl corresp="psn:W">W11.58</bibl>
</pron>
```

FIGURE 11. The pronunciation element

Two representations are included in this example, each encoded using the general-purpose TEI <seg> element. The first has @type="p", meaning (broadly) phonemic, while the second has @type="n", meaning narrow, or phonetic, transcription. The phonemic transcription also has @subtype="i", meaning 'inferred'; this transcription was not recorded on the original filecard, and has been inferred from the narrow transcription. The responsibility for constructing this representation is recorded in the following <bibl> element; 'ECH' is Ewa Czaykowska-Higgins in her role as chief editor, and the @corresp attribute on <bibl> points to the @xml:id of an entry in our personography. We will say more about how this type of linking works below. Similarly, the responsibility for the pronunciation represented in the narrow transcription is recorded in the <bibl> following it, and points to the language speaker identified as 'W.' The @corresp attribute again points to an entry in the personography, in which the full name and other details about the speaker are included, while the content of the <bibl> element provides more detail about the precise occasion on which this information was recorded, using Kinkade's original notation. Explanations of the values of @type and @subtype attributes are included in the TEI ODD file from which the project schema is generated, and are thus automatically included in the schema and in the project documentation. We discuss this feature of the TEI infrastructure below.

Note that the current organization of the contents of <pron> is not ideal; the two pronunciation representations are siblings, and are linked to their respective <bibl> elements only by contiguity and sequence. We would prefer to assign attribution through the use of @resp on the <seg> elements rather than through the use of <bibl>, or to make the structure more robust through the use of nested <pron> elements (as we do with <cit>, shown below), but at present neither of these approaches is TEI-conformant. A feature request relating to the former is currently under consideration by the TEI Council. When a better solution is available, it will be trivial to convert our encoding using XSLT.

The third component in our TEI <form> element is the <hyph>, or hyphenated form:

```
<hyph>√<m corresp="m:cah">cah</m>=<m corresp="m:cin">cí</m>-<m
corresp="m:min m:t-TR m:Ø-OBJ m:n-SUBJ">mn</m>
</hyph>
```

FIGURE 12. The hyphenated form element

Here the analytical approach we have taken to the data becomes apparent. The focus of encoding is on providing information about the morphological breakdown of the words.[14] The textual content of the <hyph> element is:

√cah=cí-mn

FIGURE 13. The textual content of a <hyph> element

This looks like a conventional morphological breakdown of the entry. Our encoding explicitly delimits the components using three inline <m> (morpheme) elements. Each of these morphemes is itself linked, through its @corresp attribute, to one or more other entries in the dictionary. For example, the first morpheme links to the <entry> with the @ xml:id "cah." This illustrates two key features of our approach:

- All attested morphemes in the language, whether bound or free, appear as individual entries in the database.
- All morphemes which are components of polymorphemic forms are linked to one or more monomorphemic entries through their @corresp attributes.

In this case, each of the first two morphemes in *cahcímn* links to a single base morpheme, the first to the root *cah* and the second to the lexical suffix *cin*. However, the final segment in *cahcímn* is actually comprised of multiple morphemes, realized together on a single phonological sequence that cannot be linearly decomposed. Here the @corresp attribute contains pointers to four source morphemes, 'min', 't-TR', 'Ø-OBJ', and 'n-SUBJ'. The definition of the @corresp attribute in TEI is "one or more URIs, separated by whitespace," so multiple links can be included in the single attribute. (The m: prefix is a private URI scheme used for convenient linking within the project dataset.)

These links provide great flexibility in our dictionary interface. From within any given entry, we can provide links to the entries of each individual base morpheme comprising it, and for each of those component morphemes, we can retrieve every entry which includes that morpheme. This makes it very simple for the user of the interface to find out how the constituents of a word or phrase work together, what each individual morpheme means, and how those morphemes may be used in other related forms. The existence of the <hyph> element, and the flexibility in how it can be used, has made it possible to construct a lexical resource that encodes a complex morphology in a straightforward manner. As a result, the database can serve the needs of several different kinds of users, including begin-

[14] As mentioned, this is a facet of many dictionaries of Salish languages produced by linguists in the 1980s and 1990s, and since our database was begun in the 1990s, it is not surprising that this focus is found in our database as well. The morphological breakdown of words reflects the analytical interests of the linguists who were working with speakers of Salish languages in the 1960s-1990s.

ning learners, advanced learners and their teachers, and linguists interested in questions of morphological structure. As is true of other types of lexical databases, TEI also allows for different 'views' of the database. Thus, even though the morphological information is easily encoded through the use of the <hyph> element, this information does not have to be present in every output that the database can produce. Most likely, beginning learners, for instance, would not need to access the most complex morphological information. In the 'Learner view' of the database, a single link to 'Related words' retrieves all other entries which share a root or stem morpheme with the current entry being viewed.

The next component in our entry is <sense>, which in our schema comprises definitions (<def>) and optional example phrases or sentences (<cit>). (Our example entry *cahcímn* in fact has no <cit>, but we will examine <cit>s below.)

```
<sense>
 <def>
   <seg>I <gloss>encourage</gloss>d someone</seg>
   <bibl corresp="psn:W">W11.58</bibl>
 </def>
</sense>
```

FIGURE 14. The sense element

The TEI <def> element included in <sense> allows a huge range of different types of content, including plain text and nearly 200 other elements. In our database, at the present time we use only two child elements, <seg>, which contains the word or phrase comprising the definition, and <bibl>, which provides the attestation for the definition.

You will also notice the embedded <gloss> tag.[15] One of the output targets we need to generate from the data is a simple English-Nxaʔamxcín word list, and <gloss> is used to tag any English form under which we would like the containing entry to appear. We can then harvest all <gloss> tags automatically to create a crude but workable English-to-Nxaʔamxcín index such as the following, in which all entries which contain <gloss>encourage</gloss> are grouped together.

[15] Note that the English 'gloss' of the Nxaʔamxcín headword, in the sense used by linguists and as used in §2.2 and §2.3 above, is contained in the <seg> within the <def> element.

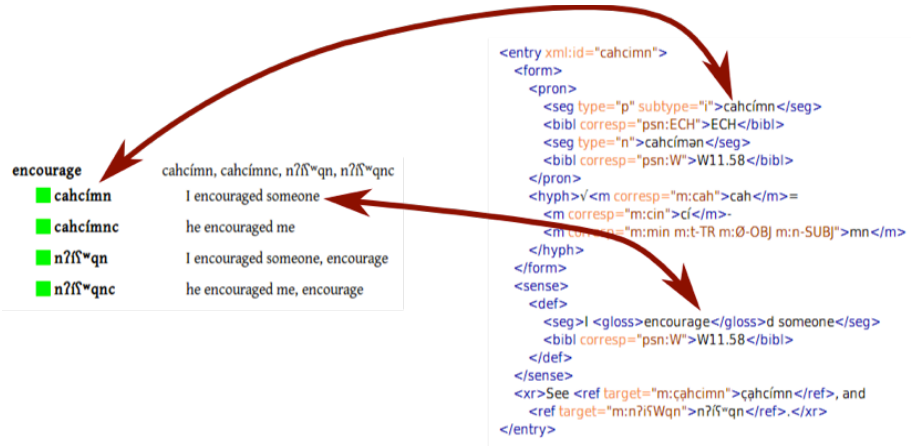


FIGURE 15. Part of an English- Nxaʔamxcín wordlist generated from <gloss> tags.

Our usage of the <gloss> tag illustrates one of the strengths of TEI: many of its inline elements, such as this one, can appear in a huge range of contexts, and we can therefore freely use <gloss> wherever we encounter such a useful English form anywhere in an entry, whether it be in a <def> or a <quote>. (See discussion of <quote> below.)

If the English wordform attested by Kinkade's consultants differs from the target headword to be tagged as a <gloss>, we add <gloss subtype="i">, meaning inferred, as shown in Figure 16. The <bibl> attributes this inference to the editor.

```
<def>
   <seg>he ran</seg>
   <bibl corresp="psn:S">S2a.130</bibl>
   <seg><gloss subtype="i">run</gloss></seg>
   <bibl corresp="psn:ECH">ECH</bibl>
</def>
```

FIGURE 16. An inferred gloss.

The second component of <sense> is <cit>, used to provide an example usage, and here we will move to discussion of a different entry, since *cahcímn* currently lacks examples. This is the <sense> element from the entry for *scáslqs* 'mosquito':

```
<sense>
 <def>
   <seg><gloss>mosquito</gloss></seg>
   <bibl corresp="psn:JM psn:AM psn:EP">Y1.145; Y14.219-222;
    Y25.15,16;EP2.46.11</bibl>
   </def>
   <cit>
      <cit type="example">
```

```
      <cit>
         <quote xml:lang="col" type="p" subtype="i"> s-√cə́s = lqs k’ʷá?ncás </quote>
         <bibl corresp="psn:ECH">ECH</bibl>
      </cit>
      <cit>
        <quote type="n" xml:lang="col">s-√cə́s = əlqs k’ʷá?əncás</quote>
        <bibl corresp="psn:JM psn:AM">Y14.219,220</bibl>
      </cit>
    </cit>
    <cit type="translation">
       <quote xml:lang="en">a mosquito bit me</quote>
      <bibl corresp="psn:JM psn:AM">Y14.219,220</bibl>
    </cit>
  </cit>
</sense>
```

FIGURE 17. The sense element: <cit> tags

A <cit> tag here consists of a <quote> element and a <bibl> providing the reference for it, but the actual breakdown of the structure is more complex than that. The outer <cit> encloses a pair of child <cit>s, one with @type="example" and one with @type="translation", the latter being an English translation of the former. The <cit type="example"> is further broken into two nested <cit>s; the first contains a <quote> which is @type="p" (phonemic) and @subtype="i" (inferred), meaning that the phonemic transcription is not directly attested on Kinkade's filecards, but has been inferred by ECH (attributed through the <bibl> element). The second <cit> contains a <quote> with @type="n", meaning that this is an original narrow transcription originating from the source filecard, and attested to by the language speakers who worked with Kinkade. The nested structure here follows an example in Romary and Wegstein (2012).

The two final <bibl> tags demonstrate another useful feature of our TEI encoding. The original filecards created by Kinkade attributed these items to 'Y.' We later determined that Y denotes two people, JM and AM, acting together (both also appear separately on other occasions). Through the @corresp attribute, we are able to point to the two individuals separately, while we can still preserve the attribution from the original source data in the text content of the <bibl>. In this way, the encoding allows us to be consistent and precise in the way we link to our personography, while at the same time remaining faithful to the data on the filecards, which themselves are significant historical records benefitting from archiving. We will discuss this aspect of our source data again below.

There are two further components of our encoding, both of which are siblings of <form> and <sense>. The simplest is <xr>, which encodes a cross-reference. This is the example from the entry for *cahcímn*:

```
<xr>See <ref target="m:ca̱hcimn">ca̱hcímn</ref>, and <ref
target="m:n?iʕWqn">n?íʕʷqn</ref>.</xr>
```

FIGURE 18. The cross-reference element

Here a prose explanation appears, in which the TEI <ref> element is used to link to other entries in the dictionary.

The second sibling of <form>and <sense> is <fs> or 'feature structure.' Within TEI,

> [a] feature structure is a general purpose data structure which identifies and groups together individual features, each of which associates a name with one or more values. Because of the generality of feature structures, they can be used to represent many different kinds of information, but they are of particular usefulness in the representation of linguistic analyses … Feature structures represent the interrelations among various pieces of information, and their instantiation in markup **provides a metalanguage for the generic representation of analyses and interpretations.** (*TEI Guidelines*, Chapter 18; emphasis added).

In our database, all monomorphemic entries are provided with a feature structure, using the TEI feature structure system, which is the basis of ISO Standard 24610-1 *Language Resource Management — Feature Structures — Part One: Feature Structure Representation*. Given the morphological nature of the database, the feature structures are primarily being used to represent the morphological aspects of entries, and in this sense they instantiate an encoding of a particular analysis of the morphology of Nxaʔamxcín.[16] In the process of trying to align our use of terminology in the feature structures with the GOLD ontology (see §4 below), we have reworked our use of feature structures multiple times, but are now close to settling on a final configuration. The example in Figure 19 illustrates the types of morphological information that are necessary to encode all and only the different morpheme categories found in Nxaʔamxcín, including position type (i.e., suffix, infix, prefix), function (i.e. derivational or inflectional, etc), and meaning.

```
<fs>
  <f name="baseType">
     <symbol value="affix"/></f>
 <f name="positionType">
    <vAlt>
      <symbol value="infix"  n="1"/>
      <symbol value="suffix" n="2"/>
    </vAlt>
 </f>
<f name="affixType">
```

[16] Given that all polymorphemic entries are linked to the entries for the morphemes that compose them, it should be possible to generate composite feature structures for polymorphemic entries through union of the feature structures of the component morphemes. We will be interested to see to what extent such composite feature structures turn out to be accurate representations of the composite forms, and to what extent it will be necessary to take into account the relative ordering of affixes in polymorphemic forms. As a reviewer correctly points out, in cases of complex derivation, relative ordering of affixes may be meaningful in many languages; to what extent derivational ordering information needs to be encoded in the database in addition to feature structure information is an empirical question.

```
      <symbol value="derivational"/></f>
    <f name="derivationalType">
      <symbol value="secondaryAspectual"/></f>
   <f name="secondaryAspectualType">
     <symbol value="inchoative"/></f>
</fs>
```

FIGURE 19. A feature-structure for ʔ/p 'inchoative', a morpheme with a suffix and infix allomorph

The <vAlt> structure indicates that this morpheme can appear both as an infix and as a suffix. The following features specify that 'inchoative' is a specific type of aspect, which Kinkade referred to as 'secondary aspectual' (Kinkade 1989, Willett 2003).

3.2 BEYOND LEXICAL ITEMS: OTHER ADVANTAGES OF TEI ENCODING. The project data consists of much more than the list of lexical items in our developing dictionary. Outside of the <entry>-level encoding discussed above, there are some significant advantages arising from the choice of TEI. The TEI guidelines have evolved over many years to support the needs of a very broad community of Humanities scholars. There are over 500 elements and 400 attributes in the full schema, including modules for the encoding of literary texts, primary sources, manuscript descriptions, critical apparatuses, graphs, networks and trees, and, of course, dictionaries and language corpora. The encoding of our dictionary entries naturally requires only a very small subset of TEI elements and attributes, described in detail above, but the availability of other features has been very useful. For instance, attested forms in the database are tied to specific speakers; we have been able to build a personography in TEI, using <listPerson>, <personGrp>, <person>, <persName>, <surname>, <forename> and related elements, and link people in the personography to the attestation instances in the entries. The personography is covered by the same schema as the entries themselves.

We have also been identifying and tagging instances of names in the entries themselves, including the Nxaʔamxcín names of individuals and groups,[17] as well as geographical features, flora, and fauna. The TEI includes support for this type of tagging, including the <persName>, <placeName> and <orgName> elements, and the ability to customize the <name> tag with the @type attribute to distinguish other types of names, as shown in Figure 20.

```
<def>
        <seg>a ridge (that flattens out) east of <seg> <placeName
        xml:lang="en">Badger Mountain</placeName>- was a campground for
        digging
       <name xml:lang="en" type="flora">camas</name></seg>
        <bibl corresp="psn:JM">JM3.150.3</bibl>
</def>
```

FIGURE 20. Encoding of placenames and flora

[17] Exactly how and to whom personal names will be accessible once the database is completed is still to be worked out. Issues of privacy obviously pertain here.

This tagging allows us to generate indices containing entries such as the one for *camas* in Figure 21.

Plant Name	**Entry**
camas	**c'ɤ̓ʷlʔúsaʔ** - camas (white)
	kàtp'áƛ'kn - a ridge (that flattens out) east of Badger Mountain - was a campground for digging camas
	ʔítɤ̓ʷaʔ - camas, black camas (cooked) (like white)

FIGURE 21. Excerpt from plant names index

The plant names index is generated by searching for all instances of <name type="flora"> in the database, creating a list not only of plant names, but also of other entries in which plants are mentioned. This provides valuable geographical and ethnographic information about plants used in the Nxaʔamxcín community.

The @xml:lang attribute encodes the language of the item enclosed in the tags. Within the <pron> element in a name entry, the Nxaʔamxcín name is tagged with <name xml:lang="col">,[18] allowing us to generate a similar Nxaʔamxcín-English index of plant names. The TEI also includes elements for GIS coordinates, which we hope to use to map Nxaʔamxcín placenames.

Our dictionary web application includes introductory and explanatory material, and this is easily encoded in TEI, which is regularly used for born-digital content. We are beginning to access audio data from Kinkade's fieldwork too, and this can be transcribed and linked using available TEI elements and attributes. In fact, whenever a new encoding requirement has arisen in the project, we have been able to support it through the existing TEI framework.

Of course, no project would want to use all of the TEI's elements and attributes; although a complete TEI schema is available (known as 'tei_all'), it is rarely used in real projects. Instead, the TEI provides a set of tools for customizing and constraining schemas, based around the idea of an 'ODD' ("One Document Does it all") file. An ODD file is itself a TEI XML file, in which the user can specify the modules, elements and attributes that will be used in their project, and those which will be excluded. At the same time, it is possible to constrain the use of existing elements and attributes more tightly than in the default TEI schema, so (for instance), where the @type attribute can have any enumerated string values in general TEI, in our project's ODD file, we constrain and document the use of @type on <seg> and <phr> so that only the values 'p' and 'n' are allowed, and those values are explained. A TEI tool called Roma is then used to generate both schema and documentation from the ODD file, so the schema for our project is highly constrained and customized.

[MosesDictionary.odd.xml]

At the same time, because our customizations are all constraints, our TEI files are still valid against the complete tei_all schema, and therefore are interoperable with other TEI projects and tools, and we can use the same schema for all of our different encoding tasks

[18] 'col' is the ISO language code for 'Columbia-Wenatchi.' Nxaʔamxcín is also known as Columbian (Salish) or Moses-Columbia Salish.

(personography, website pages, audio transcriptions etc.). The Oxygen XML editor we use for encoding is fully schema-aware, and can use the schema to prompt encoders with helpful information as they work, as in the screenshot in Figure 22.

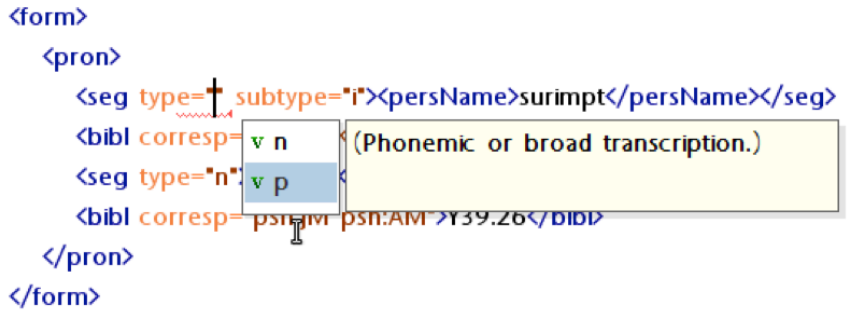


FIGURE 22. Oxygen prompts the encoder with attribute value definitions from the schema.

3.3 SUMMARY: ADVANTAGES AND DISADVANTAGES OF TEI. Our use of TEI has provided us with relatively straightforward, human-readable encoding structures which provide great flexibility in linking among lexical items and their components. We can encode what we need to encode, while at the same time we are not required by the schema to include components which are not appropriate, useful, or available.[19] The relative freedom which TEI allows is convenient too; we can tag English words as potential glosses anywhere they happen to occur in the data, rather than having to collect them into a special field, and we can similarly tag names and other identifiers anywhere in the structure.[20] Finally, as our project expands to include new encoding requirements, such as parts of speech or semantic domains, we can expect to find TEI elements and attributes available to support them.

On the other hand, this freedom comes at a cost. The *TEI Guidelines* are lengthy and complex, and often provide several different methods of encoding the same phenomenon. For example, our encoding uses the <entry> element for lexical items, but TEI also includes another candidate for the same task, <entryFree>. The latter is intended for encoding dictionary entries in an unstructured manner; this might be suitable for a project digitizing a historical dictionary, where it is essential to retain the original textual content, tagging components such as <def> where they happen to appear in the text sequence rather than organizing them into a structured hierarchy as we do in <entry>. If we had chosen to transcribe and encode Kinkade's original filecards as if they were historical texts, we might have used <entryFree> instead. Similarly, our approach to tagging broad and narrow

[19] Where, for instance, the RELISH schema requires the presence of a <Lemma> element, TEI has no such requirement; this suits us because we are still thinking about exactly how we plan to organize the print dictionary so we do not, at present, have lemma versions of our lexical items (see §4 for more discussion).

[20] When exporting the data into another format, or for insertion into a traditional relational database using XSLT, it is straightforward to harvest these elements, copy them to a specific field, and strip the inline tagging.

transcriptions using <seg>, and associating source information with them using <bibl>, is specific to our project; it is perfectly legitimate and conformant, but other TEI encoders might well choose to do things in a different way. This means that there is significant diversity in encoding practices across the TEI community. For this reason, it is especially important to make use of the TEI's ODD file and documentation features to make these decisions explicit. When we spell out our practices in the ODD file, this information then appears in the project documentation generated from the ODD file, as well as in the schema itself, and thus allows others to interpret and understand our markup decisions.

4. THE QUESTION OF INTEROPERABILITY. Bird and Simons (2003) claim that for language documentation and description to transcend time, it is essential that they be portable—that is, reusable across platforms, communities and purposes. Although our choice of TEI was made many years ago, as we have been getting closer to completion of the data-encoding phase of the Nxaʔamxcín project, we have been working on trying to ensure that the database is indeed portable. One aspect of this task has focused on ensuring interoperability of the database. Amongst other things, true interoperability, which can be defined as "the ease of moving between systems and platforms" requires standardization of format and terminology (two of the many portability issues discussed in Bird and Simons 2003).[21] Conformance to format and terminological standards makes it easier to ensure that data can be archived successfully, and can be made available in ways that might allow other projects to make use of it, other tools to read it, and other lexical databases to import it. We knew from the beginning of our work on the database that, by choosing TEI XML, we were making transformation to some other encodings and formats relatively practical, since TEI provides quite strong data typing[22] and validation features, and there is a wealth of mature tools such as XSLT and XQuery designed specifically for the purposes of querying and converting XML data. What was not clear initially, however, was the extent to which TEI is easily convertible to other standards being used in language documentation and, more specifically, for creation of lexical resources. In this section, therefore, we address the question of how easily convertible to other standards TEI is by describing standardization issues in terminology and format that have been relevant in our own work.

The first issue for interoperability of any kind is to determine the target format. We would like to be able to convert the data into a format which is some kind of *industry standard* in the field of linguistics and lexicography. There is a wide choice of possible candidates; as Budin et al. (2012) point out,

[21] In their article, Bird and Simons (2003) divide the issue of how to attain reusability into seven kinds of portability problems. One of these portability problems involves content, or "the information content of the resource"; they include the terminology used in a linguistic description as one facet of the content portability issue. A second portability problem involves format or what they term the "manner in which the information is represented electronically." The format portability problem includes the concepts of openness of format, "encoding of characters within textual information," "markup of structure," and "rendering of information in human-readable displays."

[22] Data typing in XML is the ability to constrain the value of an element or attribute (usually the latter) so that it conforms with a specific data type, such as a date, a floating-point number, an integer or a single word.

> A great number of divergent formats have coexisted: MULTILEX and GENELEX (GENEric LEXicon) are systems that are associated with the Expert Advisory Group on Language Engineering Standards (EAGLES). Other formats used in digital dictionary projects are OLIF (Open Lexicon Interchange Format), MILE (Multilingual ISLE Lexical Entry), LIFT (Lexicon Interchange Format), OWL (Web Ontology Language) and DICT (Dictionary Server Protocol).

Budin et al. (2012) also mention ISO 1951 (LEXml), and finally ISO 24613:2008, the Lexical Markup Framework (LMF), and we can also add MDF (Multi-Dictionary Formatter) to this list. However, LMF, which is a recent standard and has the ISO stamp of approval, appears on the face of it to be the best candidate for testing the extent to which our database is convertible to an industry standard. Of the other options, only LIFT appears to have broad support; the software tools *WeSay*, *FLEx (Fieldworks Language Explorer)* and *Lexique Pro* can all make use of it. However, the LIFT standard is currently at version 0.13, and there have been no new releases for over a year, and no commits to the codebase since 2011. In considering interoperability in this section, therefore, we focus on conversion of TEI to LMF.

In the course of considering TEI to LMF conversion, we distinguish *format interoperability* and *terminological interoperability,* beginning with *format interoperability,* or convertibility between file formats, encodings or schemas. We expected this to be relatively straightforward, since LMF can be serialized as (written out as) XML; XML-to-XML conversion with XSLT, the XML-based language used for the transformation of XML documents, is relatively trivial. As we show below, however, attempting to achieve format interoperability does involve two challenges for our project, one associated with the difference in the use of feature structures in LMF and TEI, the other centered on the notion of lemma. *Terminological interoperability* also faces several significant, though not insurmountable, challenges which are associated with the level at which detailed linguistic analysis is expressed, namely the level of feature structures. Thus, feature structures appear to play a crucial role in mediating between TEI and LMF, and must therefore be taken into consideration in any attempt to reach standard-harmonization.

4.1 FORMAT INTEROPERABILITY. Lexical Markup Framework is not a file format in the normal sense; you cannot 'save a file' in LMF. LMF is an abstract data model—the term used in the standard itself[23] is 'metamodel,' and the standard takes the form of a set of classes and Unified Modeling Language[24] diagrams representing the relationships between them. As such, it does not constitute a practical target for conversion from TEI; what is required is an authoritative XML 'serialization' of LMF into which our TEI XML can be converted. A serialization in this context means a format that can be stored on disk, such as an XML document. In what follows, we examine three possible target serializations of LMF, and discuss the practicality of converting our TEI data into them.

[23] ISO/TC 37/SC 4 N453 (N330 Rev.16)

[24] Unified Modeling Language is "a standardized, general-purpose modeling language in the field of software engineering. The Unified Modeling Language includes a set of graphic notation techniques to create visual models of object-oriented software-intensive systems." (Wikipedia; accessed 18 April 2013).

The first candidate serialization, actually, might be TEI itself. In fact, Laurent Romary presented a workshop on "Using the TEI framework as a possible serialization for LMF" in 2010 (Romary 2010), and Romary and Wegstein (2012) investigate this in detail, discussing the mapping of TEI encoding constructs to components of the LMF abstract model. However, their worked example consists of an encoding of a historical print dictionary (Johnson's 1755 dictionary of English), so much of the encoding discussed is not directly relevant to our project. Nevertheless, the core aspects of their proposed encoding are quite similar. The TEI <entry> element corresponds to the LMF LexicalEntry class, and it is divided into two subcomponents, <form> (LMF Form) and <sense> (LMF Sense). Within Form/<form>, various representations (orthographical, phonological etc.) are available. Within Sense/<sense>, Definition/<def> and SenseExample/<cit> are available.

Is it therefore reasonable to consider whether we could simply tweak our TEI encoding to make it map even more closely to LMF, and claim that it is LMF-compliant, and therefore that our interoperability requirements have been satisfied? There are a number of reasons why this is not an ideal solution.

First of all, TEI is a much larger and baggier standard than is required. The TEI <def> element, for instance, can contain nearly 200 other elements from the TEI schema. While undoubtedly empowering, this makes TEI over-specified for the job of lexical encoding. As Romary and Wegstein point out,

> ...the TEI has been seminal in offering a reference XML vocabulary for the representation of dictionaries, which actually offers a good compliance with LMF principles. However, the variety of constructions that the TEI actually allows for the representation of the same lexical phenomenon may be seen as a hindrance to the achievement of deep interoperability across heterogeneous lexical resources. (Romary & Wegstein 2012)

A TEI customization, substantially reducing the number of elements and attributes available, would be required to avoid the situation discussed in §3, where different projects would adopt different encoding strategies for the same phenomenon. For example, in our project, we capture the difference between narrow (phonetic) and broad (phonemic) transcriptions using @type="n" and @type="p" respectively on the <seg> element inside <pron>; another project might choose to create separate <pron> elements for each type of transcription, or choose to use different values for the @type attribute. All of these would be valid TEI, but interoperability would be compromised by the availability of this flexibility in encoding choices. A possible solution would be a TEI customization.

As we have mentioned, almost all TEI projects employ such customizations, and part of the TEI infrastructure is devoted to a toolset with the precise purpose of creating customized schemas for specific purposes. Mostly, such customizations consist of reductions of the main tagset resulting in much tighter and more constrained schemas (see 'Getting Started with P5 ODDs' for an introduction to this system).[25] Budin et al. (2012) also describe how they "tighten the many combinatorial options of TEI P5" for their dictionary

[25] LMF itself is customizable in that it is divided into core and optional modules, and users may choose which modules to include in their schema, but TEI's customization infrastructure is much more advanced and fine-tunable through its ODD system.

encoding. But schema reduction is not the only requirement. Romary and Wegstein point out that "Some LMF packages, such as the description of subcategorization frames, do not yet have any equivalence in the TEI vocabulary, but the TEI extension mechanisms do facilitate the description of such extensions." In other words, although Romary and Wegstein do advance a convincing set of proposals for aligning TEI and LMF, a practical TEI serialization for LMF would require not only constraint but also extension of TEI, and this is obviously some distance away. Even if it were to appear, it would also have to be broadly accepted by the community to the extent that lexicon toolbuilders would provide support for it. If this does happen, it will take time. We can claim notional compliance of our TEI with most aspects of the LMF metamodel, but interoperability obviously requires more.

A second candidate serialization is an example serialization in the form of an XML Document Type Definition (DTD), which is included as an annex to the LMF standard. However, there are several reasons why this is not an ideal format. It appears to be rarely used—Budin et al. (2012) point out "the still small amount of available data using LMF" (para 15)—and Romary (forthcoming) identifies four problems with it, including the fact that it *is* a DTD (rather than being defined in a more powerful modern schema language such as RelaxNG), it is 'carved in stone' and is not therefore being 'properly maintained,' it has no mechanism for customization, and it integrates poorly with other XML standards and vocabularies.

Similar considerations have led to the creation of the RELISH ('Rendering Endangered Language Lexicons Interoperable Through Standards Harmonization') project. RELISH, the third candidate serialization we considered, aims to create a "candidate for the official LMF serialization" which is based on the LMF abstract model and on the LIFT (Lexicon Interchange Format) XML schema developed by the Summer Institute of Linguistics (SIL). RELISH has a number of attractive features; it is expressed in the form of RelaxNG schemas, it is modular (users can combine the modules they need to create a custom schema), and it allows the use of TEI feature structures, in addition to the less flexible feature system available in the LMF DTD. The result is intended to be a "pivot format for lexica" (Aristar-Dry et al. 2012), and there are plans to add support for it into LEXUS (a web-based lexicon development tool created by The Language Archive at the Max Planck Institute for Psycholinguistics) and LEGO (Lexicon Enhancement via the GOLD Ontology, a project that aims to "facilitate the sharing and interoperation of lexical data" through the use of a restricted variant of LIFT called LL-LIFT).

This appears, in fact, to be exactly what we need for serialization purposes, so we started writing an XSLT conversion to generate a version of the data compliant with the RELISH schema, beginning with a comparison of our TEI structure with the equivalent RELISH encoding.

Figure 23 shows the structure of a lexical entry as represented in the RELISH schema, while Figure 24 shows the parallel encoding using TEI P5.

```
<LexicalEntry>
  <Lemma type="Form">
    <FormRepresentation type="Representation">
      <feat att="abc" val="xyz"/>
    </FormRepresentation>
```

```
  </Lemma>
  <Sense>
    <Definition>
      <TextRepresentation type="Representation">
        <feat att="abc" val="xyz"/>
      </TextRepresentation>
    </Definition>
    <SenseExample>
      <feat att="abc" val="xyz"/>
    </SenseExample>
  </Sense>
</LexicalEntry>
```

FIGURE 23. A minimal <LexicalEntry> element in LMF/RELISH, based on the example encoding in Aristar-Dry et al. 2012.

```
<entry>
  <form type="lemma">
    [...]
  </form>
  <sense>
    <def>
    [...]
    </def>
    <cit>
      <quote>
        [...]
      </quote>
      <bibl>
        [...]
      </bibl>
    </cit>
  </sense>
</entry>
```

FIGURE 24. A minimal <entry> element in TEI.

Note one key distinction between the LMF/RELISH serialization and the TEI equivalent: in RELISH, at all levels below <FormRepresentation>, <TextRepresentation>, and <SenseExample>, the only elements that can appear are feature structures. In other words, as soon as it reaches the level at which textual representations may appear, the encoding devolves to the use of feature structures. By contrast, in the TEI model, wherever '[...]' appears in Figure 24, the encoder can choose to provide plain text, choose from a wide variety of other specialist elements (as shown in example encodings above), or use a mixture of the two.

Both of these approaches provide a great deal of flexibility, but the RELISH approach is at once more verbose than TEI (including text content requires the inclusion of a feature

structure whose content is specified as text), less structured (since the range of attributes and values encodable in feature structures is unrestricted), and less helpful to the encoder (since user-friendly element names and documentation are not available as prompts from the schema when using a schema-aware XML editor such as Oxygen). This, in our view, makes the RELISH serialization much less effective than TEI for researchers working on data encoding. As Budin et al. (2012) note, "the concise and yet expressive set of [TEI] elements is definitely more easily readable to human lexicographers working on the XML source than...the LMF serialization", and Romary (forthcoming) agrees: "the generic character of feature structures, which comes with some degree of verbosity, makes it more difficult to maintain by human lexicographers."

More important, perhaps, is the fact that the use of feature structures in RELISH in itself presents a barrier to interoperability. To clarify this, consider the following two methods of encoding a part of speech. The first comes from an example in the *TEI Guidelines* and uses the TEI <pos> element; the second comes from the Aristar-Dry at al. (2012) paper, and uses a feature structure (in this case, a TEI feature structure[26]).

TEI

```
<tei:pos>verb</tei:pos>
```

RELISH, using TEI feature structures (based on Aristar-Dry et al. 2012)

```
<tei:f name="partOfSpeech">
  tei:string>Verb</tei:string>
    </tei:f>
```

FIGURE 25. Two contrasting encodings of part of speech data.

In the first case, the element name 'pos,' meaning 'part of speech', is clearly defined in the *TEI Guidelines* and schema: "<pos> (part of speech) indicates the part of speech assigned to a dictionary headword such as noun, verb, or adjective."[27] If we use a TEI <pos> element, we mean the same by it as every other user of TEI does, and this is a sound basis for interoperability.

In the RELISH case, there is the same clarity about the meanings of the generic TEI <f> and <string> elements. However, there is no such locus for agreement about the value 'partOfSpeech' for the @name attribute. This attribute is defined as "a single word which follows the rules defining a legal XML name..., providing a name for the feature".[28] An XML name is an alphanumeric token which complies with the rules for names defined in the XML specification by the W3C. Virtually any value might be used; 'partOfSpeech,' 'pos,' or 'pSpch' are all legitimate XML names. We can no longer rely on the schema and accompanying guidelines to ensure that we use the same values as other lexicon-builders; we must, instead, all agree on an independently determined set of feature categories,

[26] RELISH allows the use of two types of feature structure encoding, the ISO/TEI encoding and the native feature structure elements defined in the LMF DTD. Aristar-Dry et al. choose to use the TEI encoding in this example.

[27] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-pos.html

[28] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-f.html

names, and the relationships between them. We return to this issue below in our discussion of ontologies.

Despite these reservations, a mechanical conversion between our TEI and RELISH LMF would appear to be relatively straightforward, except for one barrier which is clearly significant. While TEI is agnostic as to the nature of a <form>—it may be @type="lemma", "compound", "derivative" etc.—LMF *requires* the presence of a <Lemma> element, while allowing optional <WordForm> elements to follow and complement it. This requirement has proved to be problematic for the Nxaʔamxcín dictionary data, because, focused as we have been on encoding the field data, we have not yet explicitly and exhaustively addressed the issue of how to lemmatize our entries. When this issue was raised on the LMF discussion list,[29] one of the responses pointed out that LMF is intended to be a format for machine-readable dictionaries, and all such dictionaries are founded on the notion of the lemma (which often has as its primary definition 'dictionary headword'[30]); without lemmas, our database is arguably a basis for a wordform inventory rather than for a full linguistic dictionary. If this is the case, then perhaps the Nxaʔamxcín database might not be ready for the level of interoperability for which LMF was designed.

We could, of course, produce a RELISH serialization of the data which has empty <Lemma> elements. But in fact there are two issues related to lemmatization that make the situation interesting and which we are still thinking about. First, as mentioned in §2.2, the Nxaʔamxcín database is constructed from attested forms; as a result, many lexemes have been recorded in only a subset of the inflected forms that they could appear in. In practice, what this means is that many lexemes are not attested in the database in consistent forms that could be chosen to serve as consistent headwords and thus as potential lemmas. Second, though, is the question of what a lemma should actually be in a language which has the kind of complex morphology that Nxaʔamxcín has.[31] As pointed out above, in many published dictionaries of Salish languages (e.g., Kinkade 1991, Thompson and Thompson 1996) headwords are often morphemes rather than words. In Thompson and Thompson (1996), for instance, main entry or headword types are specified as falling into three categories: stems, which are usually attested in words only as bound roots, and some of which can, but usually do not, appear as independent words; unanalyzable stems; and particles. In these kinds of situations, then, one could say that lemmas are essentially root morphemes rather than citation-type forms of words. In the Nxaʔamxcín database, roots, affixes, particles, unanalyzable stems and polymorphemic words all have entries and xml:ids. In the future, therefore, we will be able to choose what kind of form we wish to designate as a lemma-type in order to organize entries for a print dictionary. At present, however, we have not made this decision in practice because in the database itself, even without lemmas, there are no difficulties with navigation; all forms sharing any particular morpheme are linked, and all components of the XML structure are searchable, so finding entries and their related forms is very easy in the browser-based interface. In our system, then, morphemes perform something of the same role that lemmas do in traditional dictionaries; they serve

[29] Thanks to the LMF discussion list participants who responded to our posts.

[30] See Crystal 2011.

[31] An interesting discussion of the notion of lemma in corpus linguistics which points out the kinds of issues that we have been addressing in our own work on lemmatization for Nxaʔamxcín is found in Knowles and Don (2004).

to link all related wordforms and allow easy navigation between entries in the dictionary.

At this point, therefore, it appears that no fully workable target LMF encoding is as yet available to us. However, we could certainly make use of the RELISH encoding by working around some of the difficulties outlined above; TEI elements and attributes could easily be converted into feature structure representations, at the cost of some clarity, and empty <Lemma> elements could be used. Whether the results would be truly interoperable with other systems is difficult to determine, since at the time of writing we do not actually know of any software systems which fully support the RELISH schemas.

4.2 TERMINOLOGICAL INTEROPERABILITY. As we discussed in the previous section, while TEI has a great many lexicon-related tags whose names and definitions carry semantic value, the LMF data model relies on feature structures, and devolves to using them at a relatively shallow level. In its RELISH incarnation, both LMF and TEI (equivalent to ISO) feature structure encoding can be used in LMF feature structures. Superficially, the use of feature structures appears to promote interoperability, because it is trivial to convert between feature representations such as these (LMF and TEI/ISO respectively):

```
LMF
  <feat
     att="partOfSpeech"
     val="commonNoun"
  />
TEI/ISO
<f name="partOfSpeech">
  <symbol value="commonNoun"/>
</f>
```

FIGURE 26. Equivalent representations of a feature structure in LMF and TEI encoding.

However, as we have noted above, the ease of conversion is only true on the assumption that the terminology used in the source and target representations is identical, or is aligned in a reliable way. What this means, then, is that the use of feature structures shifts the burden of interoperability from the encoding to the feature category and naming system, or ontology.[32] It becomes essential, in other words, that anyone attempting conversion be able to refer to a common reference ontology which aligns the concepts in the source and target representations. Rather than attempting to align two XML structural representations, our conversion task instead becomes primarily a question of aligning terminologies. In this section, therefore, we discuss our attempt to make the feature structures in the

[32] Ontology is defined as follows on the E-MELD website (accessed April 18, 2013): "[T]he term ontology has a completely different meaning in information technology. An ontology here is essentially a formal statement of the relationship between terms, a working model of the entities and the interactions between those entities in some particular domain of knowledge. Its purpose is not to define meaning, but to allow computers to navigate human knowledge in a way that mimics intelligence. What it *is*, then, is not anywhere near as important as what it allows a computer to *do*. And two of the most useful things it does are that it allows a computer to respond usefully to linguistic queries, and to compare linguistic data in a way linguists understand." (E-MELD)

Nxaʔamxcín database interoperable by connecting the terminology that we use in the database with that of the GOLD ontology and ISOcat. As we show below, we found that there were various ways in which the analysis of the grammar and morphology of Nxaʔamxcín which was already encoded in the Lexware-derived database, and which reflects standard Salishanist assumptions (laid out, for instance, in Czaykowska-Higgins and Kinkade 1998) about the structure of Salish languages, did not match the GOLD ontology. We begin by briefly discussing GOLD, and then turn to how it aligns with the Nxaʔamxcín database.

4.2.1 GOLD, ISOCAT AND RELCAT. The need for developing a common reference ontology for linguistic markup was discussed more than ten years ago in Farrar, Lewis & Langendoen (2002) and has been instantiated in the work of E-MELD which proposed and began developing "an ontology of morphosyntactic terms with multiple inheritance and a variety of relations holding among the terms" (E-MELD). Interestingly, E-MELD was partly developed as an alternative to TEI. Thus, Farrar, Lewis & Langendoen state that "while standardization efforts such as the TEI and the CES [Corpus Encoding Standard, a standard for linguistic encoding related to TEI] provide for 'best-practice' methodology, the majority of the linguists still prefer their own standards and will no doubt be reluctant to adopt any form of standardization." The E-MELD website further says that rather than "proposing specific markup recommendations as in the Text Encoding Initiative", the E-MELD group at the University of Arizona "proposed constructing an environment for comparing data sets" whose central feature was the construction of an ontology (E-MELD).

GOLD (the General Ontology for Linguistic Description[33]) has developed from E-MELD, and is thus an attempt to create a stable, documented linguistic ontology to enable diverse projects to align their use of terminology and concepts in the interests of data-sharing and interoperability. GOLD's ontological database is organized in a hierarchical manner, such that, for instance, a concept such as Proper Noun appears as part of an inheritance structure thus:

```
Thing
    |_ Abstract
        |_ Linguistic Property
            |_ Part Of Speech Property
                |_ Noun
                    |_ Proper Noun
```

FIGURE 27. The class hierarchy of Proper Noun in the GOLD ontology.

While GOLD provides a structured ontology, it offers no formal method for linking to a stable definition of a term. The ISOcat Data Category Registry aims to fill this gap by providing a formal system for associating terms from such ontologies with feature structures in a specific encoding, through the use of stable ISOcat URIs associated with specific concepts. These associations are encoded through the use of two special attributes, @dcr:datcat and @dcr:valueDatcat (see Windhouwer and Wright 2012 for a detailed introduction to ISOcat and data categories).

[33] http://linguistics-ontology.org

Our quest for interoperability with LMF/RELISH, then, has primarily been an effort to map the terms we use—explicitly through our own feature structures, and implicitly through TEI element and attribute names—to equivalent specifications in the ISOcat registry, and primarily to the representation of the GOLD ontology in ISOcat. For example, imagine a feature such as this:

```
<f name="voice">
<symbol value="passive"/>
</f>
```

FIGURE 28. A passive voice feature encoded as a TEI feature structure.

We can formalize the association between our use of the terms 'voice' and 'passive' with the GOLD ontology concepts 'Voice Property' and 'Passive Voice' by looking up the GOLD concepts in ISOcat, retrieving the unique PIDs (Persistent Identifiers) associated with them,[34] and inserting them into our encoding:[35]

```
<f name="voice" dcr:datcat="http://www.isocat.org/datcat/DC-3551">
  <symbol value="passive" dcr:valueDatcat="http://www.isocat.org/datcat/DC-
   3375"/>
</f>
```

FIGURE 29. A passive voice feature with the two key terms linked to ISOcat registry entries.

Our work in mapping features to the GOLD ontology reveals three kinds of matchings: one small subset of features needed for encoding Nxaʔamxcín can be mapped straightforwardly onto the GOLD ontology; a second, larger subset has no match in GOLD—these features can be proposed as new additions to the GOLD ontology, thus essentially creating a match; and the third and largest set includes features which are mismatched or can be mapped onto GOLD only loosely. Crucially, the mapping mismatches arise from language(-family) specific analyses of morphological structures. In the next subsection we describe three examples of these mismatches.

4.2.2 MISMATCHES IN TERMINOLOGY. The first example of a mismatch between the terminology that we use to describe the morphosyntactic properties of Nxaʔamxcín and the terminology represented in the GOLD ontology involves two Nxaʔamxcín reduplicative prefixes: one of these prefixes reduplicates the first consonant of a root morpheme and is most commonly labeled 'diminutive' in the literature; the other reduplicates the first CəC of a root morpheme and is commonly labeled 'augmentative' or 'distributive.' In the ISOcat registry representing GOLD, diminutive and augmentative are size categories. In Nxaʔamxcín, however, they do not just represent size: Diminutive may refer to smallness when affixed to nouns (e.g., *ttwʔít DM+√twʔít* 'little boy'), but when affixed to predicates

[34] http://www.isocat.org/datcat/DC-3551 and http://www.isocat.org/datcat/DC-3375 respectively.

[35] In fact, rather than encoding these associations repeatedly wherever they occur in dictionary entries, it makes more sense to encode them once, in our TEI feature structure declaration file, but the concept is the same.

(e.g., *ƛ̓ƛ̓x̓ʷúp DM* + √*ƛ̓x̓ʷúp* 'win a small bit') it can mean 'partial completion' of an activity, accomplishment or achievement (Willett 2003: 278-279). Augmentative /distributive refers to plurality when affixed to nouns (e.g., *tuʔtwˀít AUG* + √*twˀít* 'boys'), but can also indicate repetitive or distributed action, or an increase in intensity of a state or activity, as in *suwˀsawˀs* AUG + √*sawˀ-n-t-Ø-s* 'He asked her over and over' (Willett 2003: 273-276).

The second example of a mismatch between the ontology reflected in GOLD and the ISOcat Data Category Registry and categories assumed by many Salishanist scholars lies in the categorization of voice and valence. GOLD does not make the same distinctions between voice and valence as Salishanist grammars do. Specifically, as in other Salish languages, Nxaʔamxcín words morphologically encode distinctions in transitivity. Thus, a word like *k'ʷáʔəncás* 'he bit me', is composed of the morphemes *k'ʷáʔ-n-t-sá-s,* where *-t-* crucially marks 'transitive', on the predicate. In most scholarship on Salish languages, transitive markers like *-t-* have been categorized as valence-marking or valence-changing morphemes, where 'valence' refers to the number of direct arguments required by a predicate. For instance, in her grammar of Nxaʔamxcín morphology, Willett (2003: 120ff) characterizes intransitives as the lowest, monovalent, valence category, and considers transitive, causative, applicative, and external possession morphemes as 'valence-changing' because they all specify more than one argument associated with a particular stem. In contrast, Willett (2003: 145ff) categorizes middle, reflexive and reciprocal as semantic voice types, following Givón's (1994, 2001) distinctions between pragmatic and semantic voice. Thus Willett's analysis of Nxaʔamxcín, which is reflected in the feature structures of the Nxaʔamxcín database, crucially distinguishes between voice and valence categories. In GOLD, however, causative and applicative are classified together with such categories as middle, reflexive, reciprocal, and passive as direct subconcepts of VoiceProperty, and are not treated as valence-marking categories.[36] GOLD can be said to represent a different understanding of grammatical roles and grammatical relations than the analysis found in Salishanist literature such as Willett (2003).

The third mismatch between the GOLD ontology and the terminology and analysis used in the feature system of the Nxaʔamxcín database is also related to the distinction between voice and valence. In particular, as just discussed, the GOLD categorization for voice/valence markers like passive or applicative is at least partly taken from a supra-category which is labeled Morphosyntactic, and which in turn is dominated by the category Linguistic Property: thus, Linguistic Property>>Morphosyntactic>>Voice Property>>PassiveVoice, ApplicativeVoice, etc. However, this Linguistic Property supra-category does not dominate or include plain transitive or intransitive markers. Instead, GOLD has a different supra-category, namely Linguistic Unit, which contrasts with the supra-category Linguistic Property, and which itself dominates sub-categories Grammar

[36] The VoiceProperty definition in GOLD is as follows: "VoiceProperty is the class of properties that concern the grammatical encoding of the relationship between the verb and the nominals in a subject-predicate configuration. It selects a grammatically prominent syntactic constituent –subject– from the underlying semantic functions. In accusative languages, the basic strategy is to select an agent as a subject (Shibatani 1988: 3). It can be said that all voice systems mark the affectedness/nonaffectedness of sentential subjects (Klaiman 1988: 30)." (See http://linguistics-ontology.org/gold/2010/VoiceProperty).

Unit, Morpheme, Bound, and Derivational. It is this supra-category which ultimately dominates two lower level categories Intransitivizer and Transitivizer (thus: Linguistic Unit>>Grammar Unit>>Morpheme>>Bound>>Derivational>>Intransitivizer and Linguistic Unit>>Grammar Unit>>Morpheme>>Bound>>Derivational>>Transitivizer). The fact that passive, applicative, etc. are subconcepts of Linguistic Property, while Transitivizer and Intransitivizer are subconcepts of Linguistic Unit means that simple transitive/intransitive valence and voice properties are treated as different kinds of entities by the GOLD ontology. But the morphological analysis followed by scholars that have previously worked on Nxaʔamxcín does not make the same kinds of categorial distinctions between transitive/intransitive Linguistic Units and Linguistic Property VoiceProperties. Moreover, neither the lower level GOLD category Intransitivizer nor the category Transitivizer is directly equivalent to the Nxaʔamxcín transitive markers.

This example of differences in the terminology and relationships between terms found in GOLD and those found in Salishanist analyses of Nxaʔamxcín suggests that attempting to align and harmonize the terminology used in GOLD/LMF and that used in the Nxaʔamxcín database is not entirely straightforward. Because the differences involve supra-categories and thus the hierarchical organization of the ontology itself and not just differences in names, trying to align terminologies cannot be achieved simply by adding categories to GOLD or by re-defining existing categories. Rather, trying to resolve these differences either requires a re-thinking of the current GOLD hierarchical structure, or, it requires re-'naming' and thus re-working the morphological analysis of Nxaʔamxcín to fit the GOLD framework.

One further obstacle to feature-structure interoperability should be mentioned. Although it appears that feature structure encoding is simple to the point where it should not itself be a vector for ambiguity, when we look more closely, this is not the case. The *TEI Guidelines* chapter on "Feature Structures" (version 2.3.0, current at the time of writing) includes two examples showing variant encodings of what appears at first glance to be the same feature:

```
<f name="number">
  <symbol value="plural"/>
 </f>

<f name="singular">
  <binary value="false"/>
 </f>
```

FIGURE 30. The same feature value?

This is followed by the comment: "Whether one uses a binary or symbolic value in situations like this is largely a matter of taste." We would argue that this is not the case. In fact, while the first encoding unambiguously asserts that the target form is plural, the second says only that it is not singular; it makes no claim as to whether it is plural or not. It might, for instance, be dual. This case illustrates another level of fragility arising out of the use of features structures: even in cases where two resources use identical terminology, and agree on its meaning, the decision to encode in one particular way as opposed to another

can hinder convertibility. What in one encoding may be the value of a feature ('singular' or 'plural') may in another encoding be the name of a feature.

In this section, we have illustrated several differences in the ontology needed to categorize the grammatical analysis of the morphology of Nxaʔamxcín and the analysis represented in the standard GOLD ontology. Differences between terminologies are certainly not insurmountable, since it is perfectly possible to attempt to find ways to map one set of terms onto another, and in this way to work towards standard-harmonization and terminological interoperability. In fact, RELcat, http://lux13.mpi.nl/relcat/site/index.html, exists precisely to support the creation of such mappings. However, as we pointed out above, trying to align the analyses of linguistic structures represented in a standard ontology like GOLD with those proposed for a Salish language like Nxaʔamxcín is more problematic. If one takes GOLD as the standard, then trying to fit the analysis of Nxaʔamxcín to that of GOLD might not be appropriate for the categories that exist or for the analysis of categories that has been proposed for the language. Attempting to harmonize ontologies thus risks imposing a (potentially inadequate) analysis onto the language. In addition, in constructing the database/dictionary described in this paper, we are working within time constraints due to funding timelines and, most especially and importantly, by the fact that this project has already been spread over many years, and the Nxaʔamxcín language community needs to have complete and usable lexical resources sooner rather than later. In such a context, terminological interoperability may be an excellent goal, but we have had to conclude that it is not a priority in our project at the present time.

5. CONCLUSION. In this paper we have described the evolution of a lexical resource project for Nxaʔamxcín, an endangered Salish language, from the project's inception in the 1990s, based on legacy materials recorded in the 1960s and 1970s, to its current form as an online database that can be transformed into different print and web-based formats. In the course of this description we have illustrated how we are using TEI P5 for data-encoding and archiving. We have shown that, as a mature, reliable standard which is also flexible, TEI is in fact a valuable tool for lexical and morphological markup and for the production of lexical resources.

As we pointed out at the beginning of this paper, there is general agreement among linguists working in language documentation and description that documentation, including creation of lexical resources, benefits from conformance to standards. We have therefore described our attempts to conform to current standards being used by many scholars working with endangered languages and their communities, focusing in particular on interoperability in the format and terminology used in markup. Our experience suggests that achieving interoperability between TEI and other standards such as Lexical Markup Framework (LMF) or the GOLD ontology for terminology is a challenge. While it is possible to achieve interoperability, ultimately it is difficult to do so convincingly, especially when it comes to terminology and the analyses that lie behind particular uses of terminology. In our work we have had to face the question of whether attempting to achieve interoperability should be a priority, and have had to conclude that in the case of a lexical resource project such as the one described for Nxaʔamxcín, interoperability is a more distant goal. The experience with standard-harmonization presented in this paper suggests that those of us working in this area need to continue to discuss such questions as what exactly conformance to standards

means (for instance, are there different degrees of conformance) and how conformance can be achieved in practice.

Acknowledgements

Many people have assisted us in our work on the database and dictionary and in the preparation of this paper. We would especially like to acknowledge the late M. Dale Kinkade and the speakers who worked with Dale in the 1960s & 70s. Without them our work would never have come into being. We would also like to thank the late Mary Marchand, the late Agatha Bart, the late Tillie George, and Elizabeth Davis, who urged Dale and Ewa to work on a dictionary in 1991. We are grateful to the members of the Nxaʔamxcín Language Program of Colville Confederated Tribes who are working with us on the project: Pauline Covington Stensgar, K'saw's Ernest Brooks and Sharon Covington, and Albert Andrews and Guy Moura of Colville Tribes who have supported us in key ways in our work. We would also like to acknowledge the various students who worked on inputting the materials into Lexware and XML, including Heba Ghobrial, Cathy Howett, Elisa Wolowodyk, Nicola Bessell, Ruth Dyck, Ben Gerson, and most recently Caitlin Bird McMillan, programmers Bob Hsu and Greg Newton, Colville Tribes Culture Committee for its support of the project, and, for financial support, the Social Sciences and Humanities Research Council of Canada, University of Victoria Faculty of Humanities, and the University of Victoria Humanities Computing and Media Centre. Thanks also to two reviewers and to Nicholas Thieberger and Andrea Berez for encouragement and comments.

References

Aristar-Dry, Helen, Sebastian Drude, Menzo Windhouwer, Jost Gippert, and Irina Nevskaya. 2012. "Rendering Endangered Lexicons Interoperable through Standards Harmonization": the RELISH Project. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*. http://www.lrec-conf.org/proceedings/lrec2012/pdf/878_Paper.pdf. http://pubman.mpdl.mpg.de/pubman/item/escidoc:1480877:2/component/escidoc:1480876/RELISH-contribution-final.pdf. (6 May 2013.)

Bates, Dawn and Deryle Lonsdale. 2010. Recovering and updating legacy dictionary data. In Joel Dunham and John Lyon (eds.), *Papers for the 44th International Conference on Salish and Neighboring Languages, Missoula, MT*. University of British Columbia Working Papers in Linguistics Vol 27. 1-12.

Bird, Steven and Gary Simons. 2003. Seven Dimensions of Portability for Language Documentation and Description. *Language*, 79(3) (Sep., 2003). 557-582. http://www.jstor.org/stable/4489465. (6 May 2013.)

Broeder, Daan, Oliver Schonefeld, Thorsten Trippel, Dieter van Uytvanck, and Andreas Witt. 2011. A Pragmatic Approach to XML Interoperability – the Component Metadata Infrastructure (CMDI). In *Proceedings of Balisage: The Markup Conference 2011*. Vol. 7 of Balisage Series on Markup Technologies. http://www.balisage.net/Proceedings/vol7/html/Broeder01/BalisageVol7-Broeder01.html doi:10.4242/BalisageVol7.Broeder01. (20 July 2013.)

Budin, Gerhard, Stefan Majewski, and Karlheinz Mörth. 2012. Creating Lexical Resources in TEI P5. *Journal of the Text Encoding Initiative* [Online]. Issue 3, November 2012, Online since 15 October 2012. http://jtei.revues.org/522; doi:10.4000/jtei.522. (6 May 2013.)

Crystal, David. 2011. *Dictionary of Linguistics and Phonetics*. John Wiley and Sons. Sixth Edition.

Czaykowska-Higgins, Ewa. 1998. The morphological and phonological constituent structure of words in Moses-Columbia Salish (Nxaʔamxcín). In Czaykowska-Higgins, Ewa and M. Dale Kinkade (eds.), *Salish Languages and Linguistics: Theoretical and Descriptive Perspectives*. Trends in Linguistics: Studies and Monographs 107, W. Winter (Gen. Ed.). Mouton de Gruyter: Berlin/New York. 153-196.

Czaykowska-Higgins, Ewa and M. Dale Kinkade. 1998. Introduction. In Czaykowska-Higgins, Ewa and M. Dale Kinkade (eds.), *Salish Languages and Linguistics: Theoretical and Descriptive Perspectives*. Trends in Linguistics: Studies and Monographs 107, W. Winter (Gen. Ed.). Mouton de Gruyter: Berlin/New York. 1-68.

E-MELD. http://emeld.org/tools/ontology.cfm. (18 April 2013.)

Errington, Joseph. 2003. Getting language rights: The rhetorics of language endangerment and loss. *American Anthropologist* 105(4). 723-732.

Farrar, Scott, William D. Lewis, and D. Terence Langendoen. 2002. A common ontology for linguistic concepts. *Proceedings of the Knowledge Technologies Conference*. Seattle, WA. http://staff.washington.edu/farrar/documents/inproceedings/FarLewLang02a.pdf. (6 May 2013.)

Givón, Talmy (ed.). 2004. *Voice and Inversion*. Amsterdam: J. Benjamins.

Givón, Talmy. 2001. *Syntax: An Introduction* (2 vols). Amersterdam: J. Benjamins.

Hale, Kenneth, Michael Krauss, Lucille Watahomigie, Akira Yamamoto, Colette Craig, Laverne Jeanne, and Nora England. 1992. Endangered Languages. *Language* 68(1). 1-43.

Hill, Jane. 2002. "Expert Rhetorics" in advocacy for endangered languages: who is listening, and what do they hear? *Journal of Linguistic Anthropology* 12(2). 119-133.

Holmes, Martin and Greg Newton. 2008. Rescuing old data: case studies, tools and techniques. Conference presentation at Digital Humanities 2008, Oulu, Finland. http://www.ekl.oulu.fi/dh2008/Digital%20Humanities%202008%20Book%20of%20Abstracts.pdf. (6 May 2013.)

Hsu, Robert. 1985. *Lexware Manual*. Manoa: University of Hawai'i Department of Linguistics.

LEGO. http://lego.linguistlist.org/. (8 August 2013.)

Kinkade, M. Dale. 1981. *Dictionary of the Moses-Columbia Language (Nxaʔamxcín)*. Nespelem, Wash: Colville Confederated Tribes.

Kinkade, M. Dale. 1982. Transitive inflection in Moses-Columbia Salish. *Kansas Working Papers in Linguistics* 7. 49-62.

Kinkade, M. Dale. 1989. Inchoatives in Columbian Salish. *Working Papers for the 24th International Conference on Salish and Neighboring Lanugagues* 24. 114–119. Steilacoom, Washington.

Kinkade, M. Dale. 1991. *Upper Chehalis Dictionary*. University of Montana Occasional Papers in Linguistics no. 7. Missoula: University of Montana.

Klaiman, M. H. 1988. Affectedness and control: a typology of voice systems. In Masayoshi

Shibatani (ed.). *Passive and Voice*. Amsterdam: John Benjamins B.V. 25-84.
Knowles, Gerry and Zuraidah Mohd Don. 2004. The notion of a "lemma". *International Journal of Corpus Linguistics* 9(1). 69-81.
Perley, Bernard C. 2012. Zombie Linguistics: Experts, endangered languages and the curse of undead voices. *Anthropological Forum* 22(2). 133-149. http://linguistics.berkeley.edu/~fforum/readings/perley_zombielinguistics_2012.pdf. doi:10.1080/00664677.2012.694170. (9 August 2013.)
Romary, Laurent. 2010. Using the TEI framework as a possible serialization for LMF. RELISH workshop, August 4-5, Nijmegen. http://hal.archives-ouvertes.fr/docs/00/51/17/69/PDF/NijmegenLexicaAugust2010.pdf. (6 May 2013.)
Romary, Laurent. Forthcoming. TEI and LMF Crosswalks.
Romary, Laurent. and Werner Wegstein. 2012. Consistent modelling of heterogeneous lexical structure. *Journal of the Text Encoding Initiative* [Online]. Issue 3, November 2012, Online since 15 October 2012, connection on 28 March 2013. doi:10.4000/jtei.540. (6 May 2013.)
Shibatani, Masayoshi. 1988. Introduction. In Masayoshi Shibatani (ed.), *Passive and Voice*. Amsterdam: John Benjamins B.V. 1-8.
Stührenberg, Maik. 2012. The TEI and Current Standards for Structuring Linguistic Data: An Overview. *Journal of the Text Encoding Initiative* [Online]. Issue 3, November 2012, Online since 15 October 2012. doi:10.4000/jtei.523. (20 July 2013)
TEI Consortium. *TEI P5 Guidelines* Version 2.3.0 17 January 2013. http://www.tei-c.org/Vault/P5/2.3.0/doc/tei-p5-doc/en/html/. (9 August 2013.)
Thieberger, Nicholas. 2011. Building a lexical database with multiple outputs: examples from legacy data and from multimodal fieldwork. *International Journal of Lexicography*, 24(3). 463-472. doi:10.1093/ijl/ecr027. (6 May 2013.)
Thieberger, Nicholas. 2013. Reusing manuscript vocabularies, an example from Western Australia. Conference presentation at the 3rd International Conference on Language Documentation and Conservation (ICLDC 3). University of Hawai'i. https://scholarspace.manoa.hawaii.edu/bitstream/handle/10125/26189/26189.pdf. (20 July 2013.)
Thompson, Laurence C. & M. Terry Thompson (compilers). 1996. *Thompson River Salish Dictionary*. University of Montana Occasional Papers in Linguistics, no. 12. Missoula: University of Montana.
Willett, Marie Louise. 2003. *A Grammatical Sketch of Nxaʔamxcín*. PhD Dissertation, University of Victoria.
Windhouwer, Menzo and Sue Ellen Wright. 2012. Linking to linguistic data categories in ISOcat. In Christian Chiarcos, Sebastian Nordhoff and Sebastian Hellmann (eds.), *Linked Data in Linguistics - Representing and Connecting Language Data and Language Metadata* (LDL 2012), Berlin; New York: Springer. 99-107. doi:10.1007/978-3-642-28249-2_10.

Ewa Czaykowska-Higgins
eczh@uvic.ca

Martin D. Holmes
mholmes@uvic.ca

Sarah M. Kell
skell@uvic.ca