2013

Problems and Prospects in the Penobscot Dictionary

Conor Quinn

Problems and prospects in the Penobscot Dictionary

0. Abstract / 1. Introduction

[REWRITE THIS AS INTRODUCITON, ADDING SEPARATE SHORT PARAGRAPH LAYING OUT THE PLAYERS AND SUPPORTERS]

Siebert 1980 discusses technical issues in developing the Penobscot Dictionary, a project unfortunately not completed at the time. We happily report on a new effort to complete this work, and detail its challenges both old and new.

The project has three major goals:

(a) recover, archive, and disseminate versions reflecting the document in its most complete forms from the 1980s project outcomes

(b) provide an error-corrected edition linked to those mss., permitting trackback of editing changes
.

(c) disseminate the resource in forms maximally accessible both to the Penobscot Nation and outside scholars alike

For (a) we discuss the digital+print manuscript sources, showing how recovering legacy data, structuring it into a digital lexicon, and correcting systematic and semi-systematic errors all can be radically facilitated through minimal but powerful digital text manipulation tools (regular expressions), which are both freely available and easy to learn. This opens the door, we suggest, to cheaper and more broadly accessible dictionary-making, especially for groups with limited resources of work time and software.

For (b) we lay out the editorial process, showcasing how documentation of intermediate stages is integral to the final product. We then examine problems of the transcriptional record (e.g. phonemic normalization issues, and the limits of comparative phonology for resolving uncertain transcriptions) and conclude that rich editorial annotation is preferable to invisible normalization.

For (c), we examine accessibility both from the text's own internal structuring and content and from its external presentation (in development and final form alike) to its user communities. We present our high-tech solutions to dictionary lookup for a polysynthetic, head-marking language---a morpheme lexicon and morphological parsing algorithms---but emphasize that real accessibility comes from solid pedagogical outreach. This goes beyond teaching learners to recapitulate Algonquianist linguistic analysis and terminology, and instead rethinks categories like "obviative" and "animate" from pragmatic, lay learner-familiar reference points. We suggest that this can also offer new insights into the phenomena themselves.

[SHORT PARAGRAPH LAYING OUT THE PLAYERS AND SUPPORTERS]

2. Recovery

2.1 Sources and their processing

Manuscript recovery has two components: the digital+print manuscript sources themselves, and the tools for processing them. For the latter, we will examine how the simple but still underutilized digital text manipulation tools known as regular expressions can radically facilitate the recovering and structuring the data into a digital lexicon, and correcting systematic and semi-systematic errors.

Our working manuscript draws from two sources.

First is Siebert's personal printout copy from the 1980s project, containing some handwritten emendations. It is now archived at the APS, and appears to be the most up-to-date version of the manuscript.

Second is a set of 5.25" disk files, archived at and in 2011 recovered by the APS. This material is a slightly earlier backup draft. While otherwise close to complete, it noticeably lacks the pre-appendix of Dependent Nouns, as well as a section from the start of "k" until the "|kati-|" entry, equalling about 4.5 pages [= how many entries?]. [Also new missing 2pp discovered by Pauleena]. These and other differences mean that a full digital version corresponding directly to the Siebert printout requires carefully comparing the two ms. and re-entering missing material. (More on this in §REF.)

The original digital files themselves have already undergone two stages of recovery and structuring.

First is the APS-commissioned recovery of the original 1980s files (spring 2011). These are plaintext ASCII, and include formatting markup from the original Gutenberg word-processing application.

Second is the Penobscot Nation DCHP-commissioned preliminary tagging of that material into machine-ready (i.e. XML) dictionary fields (fall 2012).

We consider it crucial best practice to archive all the intermediate stages in this process, and also document the processing itself, and to make these available as part of the overall digital resource. In this way, our workflow is transparent to future users, both to make introduced errors back-trackable, and also to provide a model for other efforts of this sort.

Some highlights of this process are worth noting.

2.2 Basic ASCII to Unicode replacement

The 1980s files use replacive ASCII strategies that correspond to the current standard Penobscot orthography Unicode. Examples include

#	=	ə	schwa
@	=	α	alpha [= IPA /ɤ/]
$	=	č	c-haček
*	=	ʷ	superscript w (except a few isolable cases where asterisk indicates asterisk proper, for historically reconstructed forms)

(This is not an exhaustive list; and accentual diacritics in particular are slightly more complexly coded, but manageable in essentially the same way.)

Luckily, in almost all cases the replacive ASCII symbols correspond one-to-one with current Penobscot Unicode code points. So a simple global replacement for each of these correspondences produced a directly legible version of the digital manuscript.

2.3 Recovering data structure from formatting markup: the value of regular expressions

Importantly, the Gutenberg-ASCII text also includes extensive formatting markup, of the following sort:

<P2>	marks paragraphs
<BO>...<KB>	marks bold face

<UFI>...<UFP> marks italic face

Though they are just layout/design elements, these provide a way to re-establish a digital data structure for the ms. This is because some of them are used uniquely for distinct parts of the dictionary data structure (i.e. of entry, headword, part of speech, etc.).

For example, the paragraph marker is only used at the start of entries, and so becomes an effective tag for the initial edge of an <entry> field. Similarly, boldface is only used for Penobscot-orthography material, and so its tags become an effective marker for the same.

Parts of speech combine three features that make them automatically recoverable: they are drawn from a restricted vocabulary, and are always in italics. (Particularly the primary part of speech for the entry, as it is consistently positioned after the headword.)

<P2>	marks paragraphs	→	initial edge of <entry>
<BO>...<KB>	marks bold face	→	anything (and only what is) in Penobscot
<UFI>...<UFP>	marks italic face	→	+ restricted set = parts of speech

So in many cases, the precise configuration and/or relative position of these formatting tags unambiguously demarcates certain dictionary components. For example,

<P2><BO>...<KB>

unambiguously demarcates the beginning of an entry, followed by its headword, i.e. what we can relabel explicitly as

<entry><hw>...</hw>

Now most of us are familiar with Find-Replace as a tool that can easily make the [# → ə] type of replacement. But to search out and use these positional combinations of formatting tags to recover the dictionary's structure, e.g. to do this:

<P2><BO>...<KB> → <entry><hw>...</hw>

something stronger is needed.

What we use is a simple but powerful digital tool that is both freely available and easy to learn. Called "regular expressions", they do not require any special programming skills, or expensive special programs. Most word processors offer some version of them, as do free text editors like TextWrangler.

They do one simple thing: they let us do Find-Replace operations on any pattern we can define. So if we want to carry out the above replacement, we do just two things.

First, we replace the "..." with a special code .*? that means, basically, "this part can be anything" (Xa). Then we use parentheses to divvy the whole thing up into separately manipulable chunks (Xa).

a. <P2><BO>.*?<KB>

b. (<P2><BO>)(.*?)(<KB>)

This allows us to automatically find every example of the this pattern, and spit back out the second of these three chunks---which we code as \2---with changes we want on either side of it. In other words:

Find: (<P2><BO>)(.*?)(<KB>)

Replace: <entry><hw>\2</hw>

Working from this kind of automated searching (but also with some hand-corrections), it was possible to process the Gutenberg-ASCII files into a preliminarily usable form. This is a tag-structured (= XML) file fundamentally composed of <entry> elements, with the following familiar internal structure:

<entry>

<hw>ačitáwəssin</hw>

<pos>AI</pos>
<subpos>stat.</subpos>

<other> he lies with his head lower than his feet; 〈BO〉nətačitáwəssin〈KB〉 I...</other>

</entry>

With this, the ms. can already be displayed on a web browser in a familiar dictionary format (separate entries, stand-out headwords, etc.), and its major components can be searched on. (What remains now is structuring the <other> element completely, i.e. separating out translations, examples, and other remaining material.)

The point here is that this requires no computer skills to speak of: anyone can learn just a few basic codes (like .*? = "pretty much anything") and their patterns, and immediately start experimenting. If we can define the unique pattern, regular expressions can find it and manipulate it for us.

The time saved is massive. Recovery of the 16,000-entry ms. into this internet- and search-ready form took only about 25 hours, and this includes developing the search-and-replace patterns themselves, visually scanning for uniformity, and error checking.

The resources are all free: and not just the tools themselves, but also massive online reference materials, courses, and forums.

Most strikingly, for our preliminary purposes---i.e. recapitulating a print dictionary---we find we do not need a database application at all. A bare-bones plaintext file with appropriately structuring markup (as above) is enough to provide us with all core components of the dictionary. And with regular expressions, we can do all kinds of editorial and linguistically relevant "smart" searches like "find all ANs that end with /kʷ/". The file itself is small (easy to email, quick to back up and archive), and works on anybody's platform.

The main attraction of this minimalist approach is that it makes dictionary-making cheap and broadly accessible. With just a few key skills in handling plaintext and regular expressions, underfunded projects can save greatly in human work-hours and software expenses, and make a practical and richly usable digital dictionary in a short time at relatively little cost. (Which makes getting support for further bells and whistles much easier.)

[FT: The remaining Gutenberg markup in the <other> field now uses mathematical angled brackets ("〈〉") instead of plain angled brackets ("<>") so that each can be searched on separately. This also helps in validating/testing the XML itself.]

2.4 Current work and future plans.

Currently, then, we have two ongoing tasks. One is comparing the digital ms. to the Siebert printout ms. and creating a separate file of re-entered material. (Keeping them separate for now is philologically more cautious.) The other is completing the remaining structuring of the miscellaneous material currently structured within the <other> field.

We aim to complete this structuring effort before actually editing content, in case something not yet encountered in the ms. requires revision to the basic structure already designed. From there, we can provide the most structurally uniform base for content editing.

[Questions, issues? Feels like there is more to be considered here. Probably just needs more work with the material....]

3. Editing and archiving

3.1 Overview

Given the state of the mss. outlined above, our key tasks in editing and archiving are to provide an accurate, well-edited, and fully-structured final document that can be readily tracked back to its primary sources (= the digital and printout mss.).

As noted in §2, earlier stages will be archived with final digital document with guides to how to search them. Whether these should be separate files or an integrated component in its own (very large) field (so that they are never separated from the core document) remains an open question.

An archive-quality scan of the Siebert printout ms. is crucial not only for content, but also as a means to philological trackback from the final document. Entry-by-entry trackback links would be ideal, but are impracticable both in terms of time cost and incomplete isomorphy between the two mss. Instead, a field in each entry providing the page number (= scan page number/anchor) can give instant trackback to the printout ms. scan page. This should suffice for philological purposes, and is relatively quick to implement.

We currently have no special version-tracking software, and would welcome advice in this direction (particular with regard to TshwanaLex, which looks promising). In the meantime, our plan simply to archive date-and-time-stamped drafts on a daily basis. With a ms. that has yet to reach 3MB in size, this is reasonable practicable: another advantage of plaintext minimalism in the working stages of development.

3.2 What kinds of editing?

Perhaps the biggest question is exactly what sort of editing we can and/or should do. Obvious typos and errors are one thing (though what we consider "obvious" here could be wrong, too). There are a number of other cases that are less clear-cut, however.

For example, some <inflection> examples in the ms. are very likely wrong/artificial. Identifying these and distinguishing them from genuine variation is difficult. Relatedly, the default format for entries requires complete inflectional forms and part-of-speech information that have not always been documented, and may not be recoverable. This will certainly be the case for supplementary lexical material drawn from texts and other sources. At present, it seems reasonable to provide the whatever

forms are attested, plus an abstracted stem form (since this is a legitimate abstraction, and not actually claimed to be real data).

As we cannot recheck usage or translation directly with native speakers, our only way to check questionable data in this area is searching on textual attestation (and, further afield, at least checking on cognate use in PsmMl). Since none of these approaches constitutes solid primary data, we remain wary of changing original definitions (<sense> etc.) even when all such data suggests it.

[For cognates, we can also consult with Passamaquoddy-Maliseet (and perhaps even Mi'gmaq) speakers, but it is always possible that the meanings and usages of have changed. This comparative information could nonetheless at least be included in a note.]

Our baseline solution is simply to leave the primary material as unedited as possible, and simply annotate heavily. This will include flags that the headword data (etc.) is likely problematic, and certainly of any cases where the data has in fact been changed, with the rationale and the original ms. form both provided. One way or another, both need to be available and searchable, since otherwise users may not be able to find information that may actually exist.

3.3 Normalization/normativity

Normativity is a further issue for editing. As mentioned above, it is not always possible to distinguish genuine variation (dialectal, famililectal, idiolectal, and stylistic; as well as free variation) from simple error (primarily on the part of the recorder, but possibly also the speaker).

Siebert seemed to have a strong antipathy to variation itself. He often either tried to edit it out as substandard forms, or devised elaborate but still incompletely supported scenarios to validate them. These include at least two. First, his claim of clearly delimited coastal vs. inland subdialects---marked in the dictionary at points, but contradicted by a much more diffuse distributional attestion of those variants. And second, that speakers exhibiting the innovative TI 3s Cj form -tok, as against otherwise general historical reflex -tɑkʷ, use -tok as the TI form and -tɑkʷ as the OTI. This is flatly contradicted by his original field data (much of which shows initial over-writing of -tok as a substandard, "wrong" form), which suggests simply that -tok is an across-the-board innovative variant of -tɑkʷ. (The data here is messy whichever way one looks, in part because /ɑkʷ/ is evidently relatively easy for English-based recorders to mishear as /ok/.)

We also have instances of variation simply not documented by Siebert. For example, the PD has "čiláhčəli" only as 'ovenbird (Seiurus aurocapillus L.)', and 'robin (Turdus migratorius L.)' only as "wihkʷə́skehso". But one speaker I worked with (JF) was very clear that "čiláhčəli" was his term for 'robin', and a cognate form with a related designatum is also found for PsmMl (PMD; Chamberlain REF). This presumably motivates a note indicating this semantic/usage variation, both under the "čiláhčəli" and "wihkʷə́skehso" entries.

Here too, then, our editorial stance is annotation over modification, and annotation of any significant modification. The main motivation for normalizational modification is ensuring that linguistic searches for some category or type of word do not miss relevant forms that simply vary along some such parameter. Here the solution may well be to be to provide an alternative form explicitly labeled as NOT ACTUAL PRIMARY DATA that can nonetheless serve as a pointer back to a categorically relevant but somehow surface-variant form. Any advice in this direction would be particularly welcome.

3.4 Phonological issues

[IN PRESENTATION: Not enough time to present in full, so simply throw up a table/list of the key issues, and point the audience to the online version for followup; also Q&A period.]

3.4.1 Overview

Phonological issues---what Siebert would have called "issues of the phonemic record" are substantial enough to warrant separate discussion.

Here we confine ourselves to segmental issues, i.e. vowels and consonants. Our understanding of the suprasegmental system is so minimal that we can provide only limited critical evaluation of its documentary attestation; here the editorial strategy is likely to be entirely annotational.

We find Siebert's basic phonemic analysis (Siebert 1988) underlying the PD ms. and field notes to be relatively unproblematic. Most remaining issues are in certain details of representation rather than phonemic contrast, and are largely recoverable from the data as is, and so do not impinge on the presentation of the dictionary data.

More significant are the contrasts therein that Siebert himself (Siebert 1988:REF, 1996:REF) noted general and personal difficulty in recording.

3.4.2 Vowels

For example, Siebert reports that the vowels /a, ɑ, ə/ are in some cases difficult to tell apart, and handwritten corrections and re-corrections in his field notes reflect this accordingly.

Some phonetic and phonological observations can help here. Chief among them that /a/ in most Northeastern-areal languages, be it reflex of PEA *ā (as in PsmMl, Mq, and before /hC/ in Pb) or PEA *a, has as it primary phonetic target a low back vowel /ɑ/. This makes sense in terms of its origins in the PA four-vowel system, where phonologically [low] and [back] (or equivalent) would be its key contrastive features, making it no surprise that that a solidly back vowel would be the target, rather than the more familiar central /a/ vowel familiar from Latin-type five-vowel systems and their kin. Crucially, the slightly higher/ʌ/ is a common allophone of /a/. (This seems particularly the case in closed syllables, though this aural impression has yet to be mechanically confirmed.) This makes it possible to confuse with the nearby nearby /ɑ/, which, in contrast, tends to remain distinctly higher: nearly always in the realm of /ɤ/, and is typically a bit more fronted/centralized. The traditional pseudo-equation of wedge /ʌ/ with schwa /ə/ (i.e. schwa wrongly as "the vowel in _but_") in English-based beginner phonetics could have also influenced Siebert to record [ʌ] at times as /ə/ rather than /a/, even if he likely knew better overall.

Schwa phonetics in the Northeast deserves special note. Its target is never very close to /ʌ/, and instead diverges from cardinal mid central /ə/ chiefly in high and front directions, i.e. towards /ɪ/ and /ɨ/. Its lack of backness is again in keeping with its diachronic-phonological origins as a non-high front vowel.

It is systematically the only vowel that regularly reduces, either to an almost subsyllabic degree of shortness, or to form syllabic sonorants (see LeSourd REF:REF for this in PsmMl; I have also observed the same surface phonetics in Listuguj Mq, where the tradition of analysis is quite distinct (REF).)

Schwa in Pb cannot appear in absolute word-initial or word-final position and also has a markedly different behavior than the other vowels with regard to accentuation (LeSourd 2000, Quinn in progress). Alongside etymology, carefully noting these phonetic realization and phonological distributional factors greatly helps reduce the search space when problems of distinguishing /ə/ from /a/ and /ɑ/ arise.

Schwa assimilates in color to following glides: /əw/ = [ow], /əy/ = [iy] are distinguishable from /ow/ and /iy/ only in accentual properties (and only in rather limited cases). This creates a normalization issue, as Siebert generally recorded /əw/ and /əy/ in all but the few clearly morphophonologically motivated distinct cases, but sometimes wrote /ow/ and /iy/---particularly when carrying an accent---

even in forms that are unlikely to have had contrastive /ow/ or /iy/, but just the surface-colored /əw/ and /əy/. This is a particularly vexing problem with regard to offering search-optimizing normalization.

Whole or partial comparison typically can resolve many /a/~/ə/~/ɑ/ uncertainties, except where orthographic-phonological analysis obscures, as in the case of the EAb and WAb nasal cognate to Pb /ɑ/, which is often not reliably contrasted in documentation before nasal coda consonants. PsmMl can help here in that its cognates to Pb/a/ and /ɑ/ primarily have the more readily distinguished reflexes /ə/ and /a/ respectively. And much of its documentation is the result of extensive careful recording and rechecking, in its largest part by highly trained native speakers.

Even then, non-correspondences do not always clearly diagnose and resolve recorder error, as some can may represent genuine reshaping.

(X) Documentation errors or language-specific reshapings?

	Pb		PsmMl		*Pb-after-PsmMl
a.	apilatəwan	'...mushroom'	wapilatuwan [PMrsrc]		*ɑpilɑtəwɑn
b.	manakʷan	'rainbow'	mənəkʷan		*manakʷɑn
c.	kəsipəwɑn	'REF'	sipun	'black fly'	*kəsipəwan

By regular sound correspondences, PsmMl suggests the reconstructed forms given above. But for (Xa), the fact that PsmMl transparently has wap- 'white' as an Initial (Pb wɑp-) is already a clear indication of reshaping. And Pb further has -atəw(-e, -an, -is) in a number of 'round object' terms (atowətəwe 'summer squash', atəwətəwan 'winter squash', atəwis 'ball and cup game [REF:sort of; recheck]'), and cognate to terms meaning 'ball' and 'mushroom' in several other languages (REF).

For (Xb, c), two similar original nominalizer endings PEA *-an and PEA *-ān (the latter associated with intransitive verbal stems in PEA *ē) are perhaps easily switched around, since the root components are not clear for either. (Except in Pb, where kəsip- is at least found in EAb kezib-, and -əwe probably is a fossil of PEA *-əwē 'AN speak, make noise', cf. čipakəwe 'REF'. This may also be the source of ssip- in ssípahkʷəss 'sliver-cat', with an intermediate ksip- reduced to ssip- as a semantically opaque/ unproductive Initial.)

3.4.3 Consonants: the /hCC/ problem

In the consonant domain, The question of whether or not /h/ is contrasted before /CC/ remains unresolved (cf. Quinn 1998, LeSourd REF). Siebert's early transcriptions systematically lacked it. He then determined for himself that the contrast exists (reported in Siebert REF), with subsequent corrections as in (X).

(X) Correcting to /hCC/

<seski-sɑpakʷtəhɑ> corrected to <seski-sɑpahkʷtəhɑ> (S:20(Bears):51b)

This certainly seems morphologically motivated in that, lacking a deletion rule, the Medial -ahkʷ- 'sticklike object; wood' would give the corrected form.

Extant recordings do not clearly support the existence of /hCC/, but most recordings are of speakers with strong PsmMl influence, a language in which /hCC/ reduction to /CC/ is solidly demonstrated (LeSourd REF:REF). The one speaker for whom there appears to be some evidence for /hCC/ is precisely the one whose word-accentuation in a recorded performance of PsmMl makes it clear that he is not a native speaker of PsmMl. The quality of these sound recordings attesting a putative /hCC/, however,

make it difficult to be sure one way or another.

nahstok (ADawehsohsak: ca. 3:50)
nahstok (ADawehsohsak: ca. 4:12)

wəmehstónkɑnɑ 'they finished talking' (A.N.:ésahsit)
kətahčəwi-nɑči-mikahkhátipəna 'we have to go fight' (A.N.:Text 4)
[nənɑtashatipəna (S:75:80) WHO IS THIS?]

Furthermore, neighboring languages like Mohawk are also consistenly reported to have such clusters (Michelson 1989). In absence of clearly damning evidence, we remain philologically conservative, as it were. Retaining /hCC/ clusters may in fact be phonetically/phonemically wrong (just Siebert's morphological restoration, but the error becomes that of too much data rather than too little. Leaving the issue still clearly available for question, whereas removing it makes users unable to consider it at all.

3.5 Annotational design

In sum, our editorial policy seeks to modify the original data and instead provide rich annotational commentary to problematic points (i.e. it aims to play a Talmudic role). The open question at this point is how to design the annotational component.

An undifferentiated <note> field is probably unwise, since it can be useful to distinguish editorial-philological notes from purely linguistic (usage, cross-reference, etc.) notes. Furthermore, the ms. itself has occasional notes from Siebert that demand a separate categorical treatment.

An unrestrained diversity of <note> types is also undesirable, since we aim to keep the data structure maximally simple and transparent. Hence for the present, we are restricting <note> categories to no more than the three discussed above. Suggestions here are also greatly welcomed.

4. Accessibility

[TO BE RE-INSERTED]

5. Conclusions

[TBD]

6. References

Quinn, Conor.
(in progress) Tonogenesis in the Northeast: pitch-accent in Penobscot and its neighbors. Ms., University of Maine.
1999. Some unresolved issues in the Penobscot materials of Frank T. Siebert, Jr. Papers of the 30th Algonquian Conference, ed. by David H. Pentland, pp. 288-322. Winnipeg: University of Manitoba.

Siebert, Frank T., Jr. 1980. The Penobscot dictionary project: Preferences and problems of format, presentation, and entry. Papers of the 11th Algonquian Conference, ed. by William Cowan, pp. 113-127. Ottawa: Carleton University.