Problems and prospects in the Penobscot Dictionary

Conor McDonough Quinn

University of Maine-Orono

conor.mcdonoughquinn@maine.edu
www.conormquinn.com

1. Introduction

• 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.

• **Penobscot Dictionary Project** (NEH #PD-50027-13; co-PIs Conor Quinn and Pauleena MacDougall).

= A collaborative effort of the Penobscot Indian Nation, the University of Maine, and the American Philosophical Society to revise and publish (both digitally and in print) a manuscript dictionary of Penobscot, an indigenous language of central Maine.

• 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

• 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.

• 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 (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.

• 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 (c), we examine accessibility from two perspectives: the text's own internal structuring and content; and 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.

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.

• Focus for the second on how some simple but still underutilized digital text manipulation tools---called "regular expressions"---can radically facilitate recovering and structuring the data into a digital lexicon, and correcting systematic and semi-systematic errors.

• And you can do this yourself: no need for expensive experts.

• Working manuscript draws from two sources:

Siebert's personal printout copy from the 1980s project

Contains some handwritten emendations.
Now archived at the APS, appears to be the most up-to-date version of the manuscript.

Set of 5.25" disk files, archived at and in 2011 recovered by the APS.

A slightly earlier backup draft.

While otherwise close to complete, it noticeably lacks the separate Dependent Nouns section, and also a section from the start of "k" until the "|kati-|" entry, equalling about 4.5 pages, and some smaller gaps more recently discovered.

• A full digital version corresponding directly to the Siebert printout therefore requires carefully comparing the two ms. and re-entering missing material.

• 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 and 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. This makes our workflow transparent to future users, both for back-tracking introduced errors, and also to provide a model for similar efforts.

• 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	
@	=	ɑ	alpha	[= IPA /ɤ/]
$	=	č	c-haček	
*	=	ʷ	superscript w (except a few isolable asterisks proper, in historical reconstruction)	

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

Luckily, the replacive ASCII symbols almost completely 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

• Originally just layout/design elements, these have provided a way to re-establish a digital data structure for the ms. This is because some are used uniquely for distinct parts of the dictionary data structure, i.e. 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.

Each entry's primary part of speech is drawn from a restricted vocabulary, is always formatted in italics, and is consistently positioned after the headword, making it automatically recoverable as well.

<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 + position = part-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>

• 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 more flexible is needed.

• This is a set of digital tools both freely available and easy to learn, but also quite powerful. 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 name. So 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). (Only a few of these need to be learned.)

a. <P2><BO>.*?<KB> *= add in the "anything" part*

• First, we replace the "..." with a special code,

.*?

that means, basically, "this part can be anything" (Xa). (Only a few of these need to be learned.)

a. <P2><BO>.*?<KB> *= add in the "anything" part*

b. (<P2><BO>)(.*?)(<KB>) *= chunk it up with parentheses*

• Then we use parentheses to divvy the whole thing up into separately manipulable chunks (Xb).

• First, we replace the "..." with a special code,

.*?

that means, basically, "this part can be anything" (Xa). (Only a few of these need to be learned.)

a. <P2><BO>.*?<KB> *= add in the "anything" part*

b. (<P2><BO>)(.*?)(<KB>) *= chunk it up with parentheses*

• Then we use parentheses to divvy the whole thing up into separately manipulable chunks (Xb).

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

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

Replace: <entry><hw>\2</hw> *= use the "\2" to spit back the second chunk only*

- That's it. Just three things.

• 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>

• 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> content 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 immediately start experimenting. If we can define the unique pattern we're looking for, 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, and a final visual scan checking for uniformity and errors---which stuck out nicely thanks to the patterns created by regular expressions.

• The resources are all free: not just the tools themselves, but also extensive online tutorials, reference works, and help forums.

• Most strikingly, for our initial purposes---i.e. recapitulating a print dictionary---we find we do not need a database application at all. This bare-bones file, made just of plaintext with appropriately structuring markup, 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 set up 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.)

2.4 Current work and future plans.

• Currently have two ongoing tasks.

(a) 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.)

(b) Completing the remaining structuring of the miscellany of material still lumped together 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 current data structure design. From there, we can provide the most structurally uniform base for content editing.

• Here we are particularly looking for any suggestions or advice as we lay the foundations of the structure we will ultimately be fairly committed to.

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 (including the regular expression algorithms used) will be archived with final digital document with guides to how to search them. One open questioN: should these be separate files or an integrated component in its own (very large) field, so that they are never separated from the core document?

• An archive-quality scan of the Siebert printout ms. is crucial not only for final content, but also as a philological tool for 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 at least a daily basis. With a ms. that has yet to reach 3MB in size, this is quite practicable: another advantage of plaintext minimalism in the working stages of development.

3.2 What kinds of editing?

• What sort(s) of editing can and/or should we do?

• Obvious typos and errors (though what we consider "obvious" here could be wrong, too). Less clear-cut...

• 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.

• One solution: provide the whatever forms are attested, plus an abstracted stem form, since that would clearly be a (legitimate) abstraction, and not mistakable as a claimed piece of real data, but still usable for general lexicographic purposes.

• Can't recheck usage or translation directly with native speakers; only way to check questionable data in this area is searching on textual attestation. Since this is not solid primary data, we remain wary of changing original definitions (<sense> etc.) even when all such data suggests it.

• Default protocol is to leave the primary material as unedited as possible, and simply annotate heavily.

• This will include flags for headword data (etc.) that is likely problematic, and certainly for any cases where we have in fact changed the data, 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

• Normalization is a major 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 variants out as substandard, or devised elaborate but incompletely supported scenarios to validate them as reflecting other, primary contrasts. These include at least two.

• PD ms. marks some items as reflecting coastal vs. inland subdialects. However, the distinctive lexical and phonological features of putative coastal forms are nearly all the same as what we would expect from PsmMl-influenced usage (i.e. broader application of contraction, káhkakohs rather than mkàsess for 'crow', etc.). Since coastal dialects most likely would be more PsmMl-influenced, the challenge is determining whether these variants actually reflect an old (and quite possible) dialectal distinction vs. more recent language contact and shift effects.

• Siebert's claim that speakers with the innovative TI 3s Cj form *-tok*, in contrast to the general historical reflex *-tɑkʷ*, use the first for TI and the second for OTI. This is flatly contradicted by his original field data (where we can see an earlier stage where he analyses the innov ative variant as a substandard, "wrong" form); it suggests that -tok is just an across-the-board innovative replacement 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 ms. 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 1899). This presumably motivates a note indicating this semantic/usage variation, both under the "čiláhčəli" and "wihkʷə́skehso" entries.

• Editorial stance re normalization is again annotation over modification, and annotation of any significant modification. Normalizational modification is motivated by the need to ensure that linguistic searches do not miss relevant but somehow variant forms. One solution may be to provide an alternative form explicitly labeled as NOT ACTUAL PRIMARY DATA that can still serve as a pointer back to a categorically relevant but somehow variant form. Any advice in this direction would be particularly welcome.

3.4 Phonological issues

• Not enough time to present in full, but will be on web-posted version. Summary:

- /a, ɑ, ə/ allophonic ranges sometimes confused in Siebert; now better understood
- schwa coloring as a normalization issue
- PsmMl comparison can uncover real transcription errors, but also possible genuine variation

3.5 Annotational design

• Annotation over modification...so 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 require a distinct categorical treatment.

• An unrestrained diversity of <note> types is also undesirable, since we aim to keep the data structure maximally simple and transparent. For now, we are restricting <note> categories to no more than the three discussed above (= <ednote>, <lingnote>, <siebnote>). Suggestions here are also greatly welcomed.

4. Accessibility

4.1 Overview

• Accessibility is not just physical access to the dictionary resource(s) themselves. A barely usable dictionary is little better than no dictionary at all, and can do direct harm to revitalization efforts.

• Accessibility of form and content is therefore just as important as physical dissemination. Thoughtful accessibility design applies both to the text's own internal structuring and content and also to its external presentation, in development and final form alike, to its user communities. In particular, while we are interested in using digital tools to addressing the fundamental problems of lookup in a polysynthetic, head-marking language---i.e. a morphological parser and morpheme lexicon---we still think that a fuller accessibility comes from solid pedagogical outreach. Which in turn goes beyond just training learners in Algonquianist linguistic analysis and terminology, and instead rethinks the same understandings in terms pragmatic, lay learner-familiar reference points. This reframing, we think, may also open doors to new insights into the phenomena themselves.

4.2 Accessibility: internal structuring and content

• We identify two major concerns for dictionary-internal accessibility. First is the orthography itself; second is the structural nature of the language. The dictionary is essentially unusable except to users who are clear on both points; and so a guide to orthography and core grammar is integral to real meaningful use of the resource. One unsolved problem: how can we make it so users actually read the guide and internalize it, rather than going directly to the dictionary and finding it too difficult to use?

4.3 Accessibility: internal structuring and content: orthography

• Established Pb orthography is quite opaque to English-based users:

• Noncontrastive obstruent voicing is not written despite being quite salient to English speakers; also, two essentially English-like consonants are written with non-English letters (k^w, č; also h^w).

• Vowels used with unfamiliar continental/IPA sound values, and also include two non-English letters (ə, α).

• Diacritics for pitch-accent and (noncontrastive) length are unexpected, uninterpretable, and typically ignored.

• *How we deal with this...*

• Reports from learners suggest that many believe (likely from English educational norms) that a writing system is an imposed claim about right vs. wrong ways to write words down. We explain instead that the sounds that matter for one language may not for another: the technology of writing therefore can and must always be flexible. So we explain specifically what voicing contrasts are, and show how they systematically alternate even in the same word, motivating the choice to equally systematically not write /b d g/ etc. But we then make clear that this is no more "correct" than writing out the voiced forms, and encourage beginners to do just that, in order to better understand the voicing alternation pattern. And to return a bit of power: clarifying that learners do have the right to thoughtfully rework their uses of the writing system to fit their changing needs.

• Similarly, we acknowledge that more English-based vowel spelling norms would also work, but would then make the long words even longer (e.g. from doubled "ee" and "oo"). Finally, we give examples of words distinguished solely by accentuation, and provide clear hints for easily producing and perceiving those contrasts. In short, we present the orthography as a well-thought-out tool for writing down Pb sounds, but not the only possible one, nor claiming any kind of inherent authority.

4.4 Accessibility: internal structuring and content: grammar

• Second concern: pervasive structural differences mean that relatively little basic Pb expression directly matches English.

• Users cannot, as they often expect, simply look up words corresponding to English and string them together. Nor can they decode even the most basic Pb sentences without understanding that (and how) Pb words constantly change shape to reflect the relations they are part of.

• No matter how little jargon is used, traditional linguistic analysis-based approaches seem only to succeed in convincing potential speakers that the language is too complex to be learned without massive dedication of time and effort.

• A grammar guide absolutely MUST counter this misconception immediately, by being matter-of-fact, rather than exoticizing.

• All explanations can and should use only non-technical terms. This can be easily done by introducing language patterns first as what they pragmatically accomplish, and only afterwards as the fiddly details of their formation. Here we can and should avoid abstracted morphological breakdowns: these artificially load learners up with too many bits and pieces to track in real time. Instead, a simultaneously bottom-up and top-down approach is advised.

• This starts from absolute minimal USABLE forms, small chunks which are mastered first before being built upon, simple increment by mastered simple increment---to showcase that anything that seems complex is just several simple things happening at the same time.

• Alongside this is the top-down approach, e.g. introducing the whole ditransitive (double object) construction early on, precisely because other verbal patterns are then just simpler subsets of it. And also because when whole constructions are internalized as pre-processed chunks, real-time recognition and production both work much better.

• Whenever possible, we present Penobscot in terms of English-familiar language patterns. This turns out to be possible more often than not, though not always in instantly obvious ways (cf. Quinn 2006 on the exact parallel between obviativity and clausal subordination). But more than anything, we promote a focus first on understanding what is said in Pb itself, rather than on how to translate from English.

• We do provide access to technical Algonquianist terminology, but reframe it so that it is never taken as a precondition to understanding the patterns themselves.

• (Dictionaries often attempt to instruct the reader in the technical maelstrom that is Algonquianist terminology; we think these terms should be introduced only after the concepts they name are explained in clear, everyday-language terms. A list of terms like "obviative", "animate", etc, followed by brief definitions and examples, is not enough.

• Instead, the key morphosemantic components of the language need to be demonstrated in a brief but effective set of explanatory narratives, highlighting their real-world consequences for meaning. Technical terms can then be provided at the end, and also in a final appendix for quick lookup, with references back to the narrative, situational-usage-based explanations.

• One challenge here is that the original dictionary ms. is permeated with technical terms, particularly in its part-of-speech component.

• The digital, simple-XML-based version is flexible; it permits multiple alternative presentations: with technical terms minimized, replaced with more transparent alternatives, or simply hidden.

• In print, we look to the PMD's insightful use of small, lower-case abbreviations to provide technical information to specialists while minimizing its visual impact for lay users.

4.5 Accessibility: internal structuring and content: learner lexical prioritization

• We expect to create topical (= "thematic") subdictionaries by adding topic codes into the PD entry fields. Which topical categories are of top priority needs to be determined---any suggestions?---since these need to be a limited few to be practicable within this grant.

• One category, however, stands out: tagging and perhaps even roughly ordering a "top X-hundred words" of core, highest-frequency vocabulary.

4.6 Accessibility: external presentation

• In terms of accessibility in the external presenation, we are currently thinking much about

- optimal user interfaces
- typographical and graphical layout
- comparing (and contrasting) both of the above regarding the needs of print and web format

- a morphological parser and morpheme lexicon (see §4.7)

• In the web presentation in particular, we are aiming for an uncluttered design, with a nice color scheme and soft clean lines: not hard on the eyes, with key components front and center. We also want to ensure that the search function is easy to use, and does not return too much, any more than it does too little. Any suggestions in this area are welcomed.

• The optimal approach is not just the design of the resource itself.

• A maximally accessible dictionary is one that is fully contextualized by the user communities---and not just in its final form, but also in its development. To this end we are offering a series of community workshops throughout the length of the project. The exact nature of this engagement is essential: not just introducing the dictionary as is, but actively inviting feedback on how it can be reshaped to better serve these communities. Some questions we are currently asking are:

- What do you want most in the dictionary?
- Especially for teaching?

- What problems have you had in using the current versions?

- What changes would you like to see in the PD?
    - In the layout, presentation, etc.?
    - In the content, etc.?

- How would you like to go about teaching the technical linguistic terminology? What alternatives would you like, if any?

• Seeing it as a fundamental part of making a truly usable dictionary, we will be integrating a small-scale course in basic Penobscot into the workshops.

• The course is intensely minimalist: a set of seven foundational grammatical patterns, carefully selected as those off of which most others can be readily incremented, and each introduced as an immediately pragmatically useful construction. For any given paradigm, instead of the typically overwhelming complete chart of pronominal forms, we introduce its workings using only 1s and 2s, since together these are the minimum needed for a YOU and ME conversation. This allows learners to grasp the principle of the pattern without burdening them immediately with all its possible forms.

(1) Seven-point approach

1	possessor marking, with demonstrative identificational construction	[that is {my, your}...]
2	IdpIdc use of possessor marking (simplest minimum)	[{I am, you are} named ...]
3	Cj + demonstrative deictic constructions of the above	[that is what {I am, you are} named]
4	ditransitive commands (Imperatives)	[give ME, give HER]
5	Idp statement versions of the above	[you give ME it, I give YOU it]
6	Cj + demonstrative deictic constructions of the above	[that is what you gave ME, etc.]
7	topic continuity (backreference) and change particles	[that (established) is what I am named] [that (new topic) is what I am named]

(1) Seven-point approach

1	possessor marking, with demonstrative identificational construction	[that is {my, your}...]
2	IdpIdc use of possessor marking (simplest minimum)	[{I am, you are} named ...]
3	Cj + demonstrative deictic constructions of the above	[that is what {I am, you are} named]
4	ditransitive commands (Imperatives)	[give ME, give HER]
5	Idp statement versions of the above	[you give ME it, I give YOU it]
6	Cj + demonstrative deictic constructions of the above	[that is what you gave ME, etc.]
7	topic continuity (backreference) and change particles	[that (established) is what I am named] [that (new topic) is what I am named]

• Mastering these seven points sets the stage for pretty much everything else. 3rd person arguments (and their often very different patterns) increment and contrast off of the now-familiar 1|2 patterns; plurals expand on the patterns of singulars; obviatives come in in the parallel patterns of nominal and verbal argument structure; negation concord fits as one new element into the Idp, Cj, and Imperative frames; and early knowledge of Idp vs. Cj makes it possible to introduce the rich range of constructions (esp. interrogative) based around Cj and Relative Roots. The Idp ditransitive similarly grounds learning the AIO pattern, and in turn, the Subordinative. The pragmatic, real-life communicative value of each of the seven constructions above should hopefully be self-explanatory.

• Particles are in this initial core because these integral elements are often systematically ignored in teaching, and typically difficult for students to pick up on their own. Stem-construction is omitted because it is relatively transparent, fun to learn, and fits easily into usage, which is what the remaining bulk of the course concerns itself with. Same too with gender, which sneaks in the back via matching demonstratives, numerals, pronominal marking, and verbal Finals.

• Students have often stated "I just want to talk...", so a minimal but targeted presentation of key language patterns works best when constantly looped back into real-life situation usage. Here we plan to borrow liberally from the Metallic approach (Sarkar and Metallic 2009), which grounds the spoken in the visual, and also bases itself on working from a minimal core.

4.7 Accessibility: morpheme lexicon and morphological parsing algorithms

• Obvious accessibility issue: inherent difficulty of dictionary lookup in a heavily inflected polysynthetic language, where a simple lexeme can potentially have thousands of context-determined forms.

• Well-known problem (Poser 2002, Maxwell and Poser 2013).

• Digital lookup interface solution: a morphological parser that can (in principle) take any inflected form and redirect the user to the corresponding citation form.

• Implementing a full parser of this sort is outside our present grant's scope.

• Can at least optimize the resource in that direction.

• In particular, for learner accessibility, we are already providing:

(a) a bound-morpheme lexicon (of Initials, Medials, and Finals

(b) a transparent account of the functional lexicon (= derivational and inflectional), its allomorphies (individually and as patterns/configurations) and their morphotactics

• Both of these are essentials for a morphological parser.

• But we still think the best parsers to train are human ones.

• Effective tools for them to learn and practice must cover three points: how to help learners:

(a) grasp the importance of bound morphology at all

(b) grasp the patterns themselves

and above all...

(c) not find themselves intimidated by or resistant to these patterns

• Efficient computational algorithms for morphotactic and allophonic patterns ideally translate readily into clear representations for teaching. For both computational and pedagogical purposes, then, we are working to produce a minimalist, reductive set of such patterns (2).

(2) Preliminary reductive algorithms for morphological parsing

- weak vowels (= a, ə)

deletion and IC alternation (vs. non-alternations like (C)ahC < PEA *(C)āhC)
deletion after vowels (in composition)
fusion to /o/ after Cw (often diacritic!) and some Caw (vs. listable shifts to /α/)

- weak consonants (= glides)

intervocalic glide loss: this is abstracted/diachronic, but feeds...
...univocalization (= deletion within VV sequences) and also....
...gives rise to systematic contrastive penult accent
related alternations of /e/ as /ew/~/ay/ (> univocalizing deletion)

/wC/ to /CC/ or /hC/: domains of application are restricted
partial-harmony of schwa across /h/

- consonant-vowel secondary articulation effects

palatalization of /t/ before most /i/ (exceptions listable; as also rare palatalization before /o/)
delabialization of /k^{w}/ before and after /o/ (incl. weak-vowel fusion-derived /o/)

• As they are nearly exhaustive, we see these generalizations as the start of a set of plain-language algorithms that can be handed off to learners or computational linguists alike. Within them, we find that the linchpin is the representation/implementation of weak vowels, as their alternations present the most pervasive challenges to simple concatenation-based parsing. Equally important in representation/implementation are word-edges---both as phonological boundaries and morphological ones, e.g. definitional to the Initial, Medial, Final and Pre {verb,noun}categories---and zero morphs. The above observations may also be applicable in whole, in part, or in essence to other Algonquian languages. Needless to say, we are very interested in collaborating with anyone else working in this realm.

• While the morphological parser solution is one we do hope to see realized, we still think that accessibility in this area is still fundamentally more about solid pedagogical outreach. Here traditional Algonquianist terminology and analysis has tended to be more obfuscating than helpful, furthering the view of (and by!) linguists as practioners of obscure analysis disconnected from real language use. Translating Algonquianist analytical insights back into the pragmatic uses of the basic patterns, and prioritizing both the minimalist and the naturalistic in presentation, should go a long way in bridging this avoidable gap. And being forced to recast standard-terminology-framed analyses in new, plain-language terms is likely to bring about new insights into the phenomena themselves.

5. Conclusions

• Annotation over modification

• Accessibility targeted in every component, from within and without.

• Humans over parsers, but still try to set up for a parser.

• Rethinking Algonquian analysis for genuinely accessible lay presentation helps linguistic analysis at least as much as it does finally meet the learner need.

• Plenty more to say. Have at.

6. References

Fidelholtz, James. 2003. Contraction in Mi'kmaq verbs and its orthographical implications. Ms., Universidad Autónoma de Puebla.

Francis, David A., & Leavitt, Robert. 2008. *Peskotomuhkati Wolastoqewi latuwewakon / A Passamaquoddy-Maliseet dictionary*. Orono, ME: University of Maine Press.

LeSourd, Philip. 2000. The Passamaquoddy "Witchcraft Tales" of Newell S. Francis. *Anthropological Linguistics* 42 (4):441-498.

1993. *Accent and syllable structure in Passamaquoddy*. New York: Garland.

Maxwell, Michael & Poser, William. 2013. Morphological interfaces to dictionaries. Ms., University of Pennsylvania.

Quinn, Conor. (in progress). Tonogenesis in the Northeast: pitch-accent in Penobscot and its neighbors. Ms., University of Maine.

2006. Referential-access dependency in Penobscot. Ph.D. thesis, Harvard University.

1999. Some unresolved issues in the Penobscot materials of Frank T. Siebert, Jr. In David H. Pentland, ed. *Papers of the 30th Algonquian Conference*. Winnipeg: University of Manitoba. 288-322.

Poser, William. 2002. Making Athabaskan dictionaries usable. In Gary Holton, ed. *Proceedings of the Athabaskan Languages Conference -2002*. Fairbanks: Alaska Native Language Center, Univ. of Alaska. Working Papers #2. 136-147.

Sarkar, Mela, & Mali A'n Metallic. 2009. Indigenizing the structural syllabus: the challenge of revitalizing Mi'gmaq in Listuguj. *Canadian Modern Language Review* 66 (1): 49-71.

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

1996. *Penobscot dictionary*. Ms., American Philosophical Society.