Towards a model of maximal accessibility in linguistic documentation work

Conor McDonough Quinn

Proceedings of Conference on
Language Documentation & Linguistic Theory 2

Edited by Peter K. Austin, Oliver Bond, Monik Charette, David Nathan & Peter Sells

13-14 November 2009 School of Oriental and African Studies, University of London

Hans Rausing Endangered Languages Project
Department of Linguistics
School of Oriental and African Studies
Thornhaugh Street, Russell Square
London WC1H 0XG
United Kingdom

Department of Linguistics:
Tel: +44-20-7898-4640
Fax: +44-20-7898-4679
linguistics@soas.ac.uk
http://www.soas.ac.uk/academics/departments/linguistics

Hans Rausing Endangered Languages Project:
Tel: +44-20-7898-4578
Fax. +44-20-7898-4349
elap@soas.ac.uk
http://www.hrelp.org

ISBN: 978-0-7286-0392-9

This publication can be cited as:

Conor McDonough Quinn. 2009. Towards a model of maximal accessibility in linguistic documentation work. In Peter K. Austin, Oliver Bond, Monik Charette, David Nathan & Peter Sells (eds) *Proceedings of Conference on Language Documentation and Linguistic Theory 2*. London: SOAS.

or:

Conor McDonough Quinn. 2009. Towards a model of maximal accessibility in linguistic documentation work. In Peter K. Austin, Oliver Bond, Monik Charette, David Nathan & Peter Sells (eds) *Proceedings of Conference on Language Documentation and Linguistic Theory 2*. London: SOAS. www.hrelp.org/eprints/ldlt2_22.pdf

Towards a model of maximal accessibility in linguistic documentation work

CONOR MCDONOUGH QUINN
University of Nizwa

1. INTRODUCTION

The goal of this paper is to present a novel approach by which to facilitate heritage-community access not just to the products of documentary linguistic work, but also to their means of production. In this presentation we show how to share with a non-technical audience the means to easily link together a rich and complicated range of linguistic source material in a form that is both stably archivable and broadly accessible/disseminable. The approach suggested is a simple one: the skills and technology involved are minimal and readily learnable, even to the most computer-illiterate/computer-unconfident.

Specifically, we demonstrate how to handle a core problem in community language work, namely, how to effectively collect, organize, and distribute multiple presentations and versions of the same source material. A simple, workshop-style introduction shows non-specialist language workers how to easily create basic annotated interlinear glossed texts, forming master documents which are not only searchable, but can also readily incorporate linked audio and image/video files, or image files of original documents, as well as any number of alternate or earlier versions of the same core material, e.g. direct transcriptions, and phonemicized versions thereof, plus corrections. Particularly highlighted for the non-technical audience is this approach's flexibility: that alongside this capacity for extensive annotation and cross-linking of documentary materials comes a capacity for exceedingly fluid presentation, allowing users to create , from one single master document, the following outcomes (essentially the at the touch of a button):

- interlinear glossed texts
- facing-page (synoptic) bilingual texts
- monolingual texts or translations
- selective presentations (e.g. school-appropriate dictionary entries)

The master document can be printed out as a human-legible paper backup, just in case all digital data systems become defunct.

To the non-specialist, this task would seem to require either extensive technical computer knowhow, or at least an expensive and/or complex database tool. In this talk we showcase how one can, in a wholly non-technical, broadly accessible manner, share the basic notion(s) and application(s) of XML-based data management - and its benefits of being free, open-source, platform-neutral, fundamentally human-legible, flexible, and simple to use and troubleshoot - to precisely the audience least confident in their ability to use such a tool.

The technology involved is not new. What we show here, using a very brief but effective introduction to the small set of skills needed to create documents of this type, is that this approach can be detechnicalized and presented in a form that makes it usable to those with limited background and resources in digital data management. This is offered as one step in the ongoing process to democratize access not just to the products of documentary linguistic work, but also to the tools of its creation.

2. ACCESSIBILITY

Community-level access to the means of production of language materials is the chief aim of this work. This is part and parcel of a baseline commitment to active heritage-community empowerment in this domain, as a replacement for the still all-too-frequent scenario of prolonged dependence on outsider academics for the production of core language maintenance and revitalization materials.

Helping heritage community members empower themselves to do linguistic materials production on their own - what one might call the Ken Hale method, though he was by no means the only exponent of this approach - has so many clear advantages that they need not be stated at length. We might at least briefly note, however, that in this model, first and foremost, the primary stakeholders get to run their own show. Ownership and authorship rights remain more clearly local or at least primarily local; and the goals and intentions of academic linguistic research (particularly of outsiders) become more transparently understandable (these last two going a long way towards addressing suspicions and grievances about theft of cultural wealth); and opportunities for community members' direct access to further resources (academic degree study, grant funding, etc.) improve substantially. We can also show precisely where valuable reusable skills can be shared: my recent field methods students have remarked on how they now can and will use XML-type data management in other aspects of their work life. This sort of enhancement to basic computer skills seems likely to have a comparable doubling value for many heritage language community members as well.

The present approach also has the benefit of putting an outsider or local linguist more clearly into the hired-consultant model of academic linguistic expertise, rather than in the social-conceptual frame of the authoritative scholarly expert. The communities that I have worked with certainly much prefer the former, on both a political and a personal level. As do I, since I think it is a fairer and more accurate framing of the baseline social reality in most language documentation situations.

Finally, attention to accessibility as means to community-level empowerment in language work not only puts power where it ought to be, while addressing concerns about outsider power and priorities, but also serves to develop support for documentary linguistic endeavors in general. People are much more likely to think an effort worthwhile if they can not only understand the product, but also have a direct hand in its creation.

3. INTRODUCING NON-SPECIALIST USERS TO STRUCTURED LABELING OF DATA, AND THE FLEXIBLE PRESENTATION THEREOF

This introduction assumes an audience with access to digital tools and a specific minimum set of skills to use them: producing basic word processing documents, manipulating files and folders to save and move such documents, and using the internet. It is far from certain what percentage of community language workers around the world have these skills and internet access, although the latter is increasing daily. The point here of course is not to assume universal access to digital tools, but simply help non-specialists get the most out of what they do have available to them, and to speed them to a working level of productive ability if these tools and skills are introduced for the first time as part of the documentation effort.

The skills presentation itself lays primary emphasis on audience empowerment: that there are common problems they face in their language work that can be solved using simple tools, ones requiring little more than the basic computer skills and technology they already have. Here the emphasis is on the idea that the new technology lies in the concept, not the computer - i.e. that the tool they need is just a piece of mental technology, a technique rather than a device: namely, structured labeling of data. Performed in what we might call a relentlessly informal style, the presentation constantly loops back to this basic notion that any ordinary user of computers can easily implement the approach.

It might seem sensible to attract the interest of the audience right from the start by showcasing the power of this technique, i.e. to start by flipping through the many nifty presentational outcomes it can produce. However, audiences may well be used to seeing whiz-bang digital tools demonstrated to them by the technologically proficient, only to find that the tools are far from easy for they themselves to use. Hence we explicitly choose *not* to show first what clever things can be done, but instead start by stating some familiar problems, promising a solution, and guiding the audience to the point where they feel that they themselves can implement the relevant techniques.

For the remainder of this section, I will offer a summary of a sample presentation of this approach, with notes and comments on why the presentation is structured as it is. I assume that the present reader is more or less familiar with the concepts contained in the talk: the purpose of the summary here is to highlight *how* these concepts can be quickly and accessibly presented to a non-specialist audience.

The talk begins by stating its goal explicitly: the desire to maximize accessibility, by empowering people to do a whole lot with very little. Namely, by using the simplest tools possible. The advantages of simple systems are emphasized: that they do not break as easily, and are easier to fix; that they share/disseminate more easily; and that they stand a better chance of long-term survival.

This insistent emphasis on simplicity pervades the presentation, since the audience we are most trying to reach is precisely the one most intimidated by the

slightest hint of technological complexity. The next step, in fact, is explicitly naming as the direct addressees of the talk precisely those people convinced that they cannot do anything with a computer beyond word processing and email. Then, we highlight perhaps the most burdensome problem they face that this approach is a decent solution to: the laborious task of data migration and all its attendant joys of anarchic formatting mutation.

With their interest piqued, we immediately lay out the core concept: that the technology is not in the computer, but rather in the idea, the simple idea of labeling one's information. Humor being a useful aid to memory, my present means of conveying this notion is to present an image of a cow, then to label it with opening and closing <cow> tags:

<cow>[image of a cow]</cow>

This allows the presenter to briefly note the formal properties of XML tagging - angle-brackets and forward slashes - by means of an informal and explicit demonstration.

Now that data types can be labeled visibly for the audience, the presentation moves on to examine a familiar case: the implicit data structures of a dictionary (entries, headwords, parts of speech, senses, definitions, example sentences), demonstrating this by contrasting a sample dictionary passage in typical dictionary layout and formatting against the same passage with all the elements and structures explicitly identified through tags.

This in turn allows us to show one specific and fundamental advantage of this kind of data management approach: easy reformatting of presentation of extensive structured data. Here we contrast the problem of hand-reformatting the entry layout of an entire dictionary, versus simply restating what the presentational form of each tag will be. The advantage becomes clear: reformatting at the press of a button, not the slog of hours of highlight-and-click.

With this, we are ready to give the audience the core of the presentation: a minimalist data structure for an interlinear text. This consists of nothing more than an enclosing <line> element containing three sub-elements: <tgl> for target language, <eg>[1] for language of wider communication, and <note> for notes. Looking at a short sample of a real interlinear text, then making the same data-structure-demonstrating contrast as was done for the dictionary sample, we show that it can be nothing more than this:

[1] Here '<eg>' is in fact preferred over the more neutral <lwc> for 'language of wider communication' when presenting to an audience with that specific LWC. This is to minimize the number of opaque abbreviations involved. Needless to say, for an audience expecting to use a different LWC, this component would be adjusted accordingly.

<line> contains:

<tgl> target language
<eg> language of wider communication; here English, hence <eg>
<note> notes

And that is the key to this presentation: to get the audience to the point where they can grasp this simple data structure and feel confident that they can reproduce it (or something like it) in their own language work. Since this the essence of the system—this alone is enough for a non-specialist worker to produce effective XML-structured data: human-legible as is, readily automatable for flexible digital presentation (either by the producer directly or when handed off to a specialist), and, by dint of its simplicity, robustly archivable.

Well structured data does not do anything by itself, however. Hence at this point in the presentation it is also necessary to familiarize the audience with the minimum formalities needed for this data structure to be presented in various formattings using current XML/XSL technology. It is possible to reduce these to just four main points, which are what we then present. Namely:

(i) The need to add current XML-document headers at the head of a document, along with a link to a stylesheet document. This is presented as minimal gobbledygook that the audience can simply cut and paste without needing to understand. In the current version, this reduces to just two lines of code:

<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheettype="text/xsl"
href="LinguisticDocumentationStylesheet1.xsl"?>

(Note that later in the presentation the audience will find out more about the stylesheet component here.)

(ii) The need to give the XML document a single root element: this is conveyed simply by saying, 'Make sure you give the whole document itself one big label: that's to enclose it.'

(Here, strictly for the specific purposes of the upcoming prefabricated stylesheets, it is necessary to walk the audience through the essentially arbitrary choice of using <collection> as the root element label, and then within that root element, using <text> elements to wrap inside them all <line> elements that form a single text.)

(iii) The need to have the XML document be plain text and have the *.xml* extension to its filename. Again, simply as this: 'Save your document as plain text; and then make name (filename) end in *.xml*. '

(Typically saving as plain text rather than in the word-processing application's specific file format also requires a demonstration, but this is simple, as most users already find familiar the basic kind of Save window in which this parameter is set.)

(iv) The need to keep all data elements related by and to the core XML document in the same folder. Actually this is not really necessary: working around it is a simple matter of giving full pathnames, but imposing this constraint maintains simplicity for basic-level users.

Following the instructions this far creates an XML document that can actually be presented in a browser, using an appropriate XSL stylesheet. Stylesheet construction is more complex than basic XML data structuring, and it is not worthwhile - and in fact likely to result in counterproductive information overload - to present the workings of an XSL system. Instead, we offer prefabricated XSL stylesheets set up to take XML of the above basic interlinearization data structure (currently available online at my website) and format it as browser-ready HTML. From there we walk the audience through the simple process of linking the core XML document to each and any of the available stylesheet documents.

This component of the presentation would seem to give the lie to the idea that the technology is readily rendered accessible to a non-specialist audience. But in fact it is: because taking the audience to this point means that they can at the very least produce the core structured XML data, and also have a sense of how that can then be linked up to system that will flexibly present it. So even if the specific stylesheets offered with this presentation are lost or otherwise not accessible to the audience, it is short work for any tech specialist to work up effective equivalents - given that the non-specialist users are now able to provide them with clear, legible, and usable structured XML documents. In short, while the prefabricated XSL component is something of a hack, it provides the non-specialist audience with the means to see the results of the core skills to be conveyed - structured labeling of data - realized with minimal extra effort as usable and practical presentational outcomes.

It is only once stylesheet linking is established that we then showcase examples of useful presentational outcomes, namely:

- interlinear glossed texts
- facing-page (synoptic) bilingual texts
- monolingual texts or translations

In working through each, we emphasize again and again to the audience that all of these different forms come from one single master document - that they essentially need do nothing more than create that one core document, and then leave creating all these different presentations as a separate task, one they perhaps

need not even perform themselves. This simplification and reduction is, of course, the working principle of the presentation.

With the audience familiarized with real instances of this flexibility in presentation, of how much can be done with even the simple interlinear-text data structure they have learned, we can briefly showcase further simple expansions of the system beyond this basic level. For a salient and memorable example, we can quickly introduce how audio and image/video components can be very simply added in, using an appropriate stylesheet and adding a component to the XML that simply structures the filenames of the relevant multimedia components.[2] By the same token, the audience can now be shown how alternative versions of the same text - original transcriptions, phonemicized versions, alternative spelling system versions - can be aligned together by doing nothing more than placing relevantly labeled elements together inside each respective <line> element. The same applies for cross-references to dictionaries and other connected resources: here we can show how each element can in turn be selectively presented or suppressed using different stylesheets, as for, say, creating school-appropriate dictionary from an unexpurgated original, or for presenting the same text in the orthography preferred by a specific set of users. The key point, as always, is to emphasize how little it takes in terms of personal educational and technological investment to make these options possible.

From there, as in all lectures that hope to not only reach but stay with their audiences, we close by reviewing and summarizing the core concepts of the presentation. I usually finish with the cow.

4. CONCLUSION

This work is based on the idea of getting the most out of radical simplicity. Active detechnicalization of the production methodology for basic linguistic documentation not only increases the number of individuals who can gain and apply meaningful production skills, but also dovetails nicely with the fact that the resultant simple documents are on the whole more likely to be archivably and distributionally robust. Hence what we offer in this kind of presentation is not a dumbing-down, but rather a bare-bones approach designed specifically so that even the most technophobic can, with an absolute minimum of training, grasp the essence of a simple XML document—i.e. structured semantic tagging—and, with only basic word-processing skills and resources, immediately create and present such documents (with the help of pre-built XSL templates) for online and offline access using nothing more than a web browser.

[2] While this particular presentation focused on text manipulation, since this currently is the simplest to implement, a particularly skilled or talented user can also utilize this technology to manage purely audio and image/video components to produce working materials for wholly unwritten languages, or for users not literate in whatever writing systems may exist.

For simplicity's sake, the non-specialist audience learns in this presentation only that bare minimum: how to prepare and present text data. The complications arising with audio and image/video components are sidelined in order that the core skills can be conveyed with an immediate and clear payoff. While this is a limited outcome, it is still a substantial step forward, in extending the range of non-specialists who can produce linguistic documentation material that is both robustly archivable and also, at the very worst, very readily handed off to a computer specialist for easy automated presentational manipulation. As a simple, open-source technique, broadly sharing this kind of approach could also help to avoid the loss of hours upon hours of work limited or even lost due to its dependence on platform - and/or application-specific components. Compare this approach, for example, with the time and effort needed for the aforementioned specialist to take language-worker-produced Microsoft Word documents and try to reproduce all their formatting in a broadly browser-friendly form, while also creating an archival version, and keeping all three documents updated together.

I would also suggest that offering new but simple and accessible digital information management skills to non-specialists may also encourage at least some of them to go still further with the technology. The key point, of course, is that the agents doing all the core work with this technology are no longer just a handful of outsider or outsider-trained specialists. With this approach, a broader spectrum of the speech community instead can make a direct and concrete contribution to the documentation of their language and the production of materials they themselves determine to be useful to their language maintenance efforts.

The present technique is still rough around the edges. There is still more to do to make this a polished, user-friendly set of procedures and documents. And certain basic language documentation/revitalization materials, such as word-for-word interlinear glossed texts, are still rather laborious to produce under the simplest versions of the approach. Needless to say, I welcome criticisms and suggestions for improvement. In particular, I welcome visits to the current online manifestation of the project, at *http://www.conormquinn.com*. The specific links are:

- an example of the actual performed presentation:
  http://www.conormquinn.com/cmquinnWAIL2009.mov.zip

- a collection of the basic XML template and associated useful stylesheets:
  http://www.conormquinn.com/LinguisticDocumentationTemplateAndStylesheets.html

The presentation performance available above demonstrates this approach by drawing its content from a collection of early Penobscot (Eastern Algonquian, central Maine, U.S.A.) texts, including the original text and phonemic retranscription of a pre-phonemic collection of traditional literature (Speck 1918), alongside a set of similarly redacted rare native-speaker writings from the 19th and 20th centuries. In so doing, it displays the facility with which one can create a

rich set of digital documents - all in a searchable, flexibly presentable, and robustly archivable form that could potentially link together all known documentation for this language - a document set that nonetheless is not bound to a specific computer platform or indeed to computer technology at all.

This small effort, then, is one step towards a fuller model of maximal accessibility in linguistic documentation work. Another technique for expanding community access to linguistic documentation is detechnicalizing the jargon in collected linguistic data - or at least offering accompanying detechnicalized abstracts/summaries. Ensuring that community members are aware of their rights of access to public archives, and how to use such archives, is still another, this one learned from the summer 2008 Breath of Life workshop at the University of California at Berkeley. All of these examples are components of an approach, or better, a viewpoint, from which specialist documentary linguistic work, both in the form of its productive skills and the outcomes thereof, can become much more accessible to the communities from which it derives. It is hoped that this presentation makes a useful contribution to how we can not only meet the basic obligations of access that we have to these communities, but also go further by creating possibilities for active empowerment, and in so doing, realize a generous spirit in academic work.

REFERENCES

Speck, Frank G. 1918. Penobscot transformer tales. *International Journal of American Linguistics* 1(3), 187-244.