Athabascan Textbases: A Navajo Case Study

Kip Canfield
Department of Information Systems
University of Maryland, Baltimore County

Introduction

There are a large number of collected and written texts in various Athabascan languages that form a substantial literature that could be used for both scholarship and education. This is especially true of the Navajo language for which there are a large number of written texts, many that are public domain or out of copyright protection. This paper describes and evaluates a project to acquire these texts in electronic format, in the standard orthography, and develop a dictionary lookup tool for use with these texts.

The primary problems for such a project are:

- Text acquisition for many text types, with differing orthography and no electronic source.
- Difficulty in developing dictionary look-up tools due to the polysynthetic character of verbs.

Collected texts can take many forms and use many different orthographies as noted by Gary Horton with reference to the Alaska Native Language Center Archive.[1] Navajo materials suffer from this same barrier to standardization. For this pilot study, we use Navajo texts that are typewriter written with a non-standard orthography.

The Navajo language has a polysynthetic structure that poses special problems for making dictionaries.[2] Although many may differ on technical definitions of polysynthetic, we only mean to say that the task of picking out a morpheme such as a stem from the complex morphology of a verb is difficult to implement for a paper-based or on-line lexicon for an Athabascan language. As Melissa Axelrod puts it: “If you think of polysynthesis as being a matter of high synthesis and

[1] Approaches to digitization and annotation: A survey of language documentation materials in the Alaska Native Language Center Archive, Gary Holton, University of Alaska Fairbanks 3rd E-MELD (Electronic Metastructure for Endangered Languages Data) workshop on language engineering, 7/11/2003, available http://emeld.org/workshop/2003/paper-holton.html.

[2] These issues are well summarized in McDonough, J. M. (2000). How to use Young and Morgan’s “The Navajo Language.” In K. M. Crosswhite & J. S. Magnuson (Eds.), *University of Rochester Working Papers in the Language Sciences, 1 (2),* 195-214, available http://www.ling.rochester.edu/faculty/mcdonough.html.

high fusion, however, I think all Athabaskan languages qualify as polysynthetic."[3] William Poser has extended this discussion of problems in Athabascan dictionary making into the on-line realm and has suggested a morphological analyzer component.[4]

These previous studies are extended here as a practical implementation and evaluation of a Navajo textbase. The complete workflow of the project is described along with an evaluation of the dictionary lookup tool.

Methods

The following steps were used for this project and are detailed below:

1. Scanner acquisition of images of the original texts.
1. Optical character recognition of the text images and post-edit.
1. XML encoding of the texts using the Text Encoding Initiative.
1. Use of an XSLT stylesheet for web display of the texts.
1. Development of an automated look-up tool for the lexicon.

A goal of this project is not only to produce a textbase, but to develop tools and a workflow that are accessible to linguist and educational specialists in Athabascan languages. Texts from *Navaho Texts* by Sapir and Hoijer are used for this pilot project.[5] Figure 1 shows a page fragment from this work.

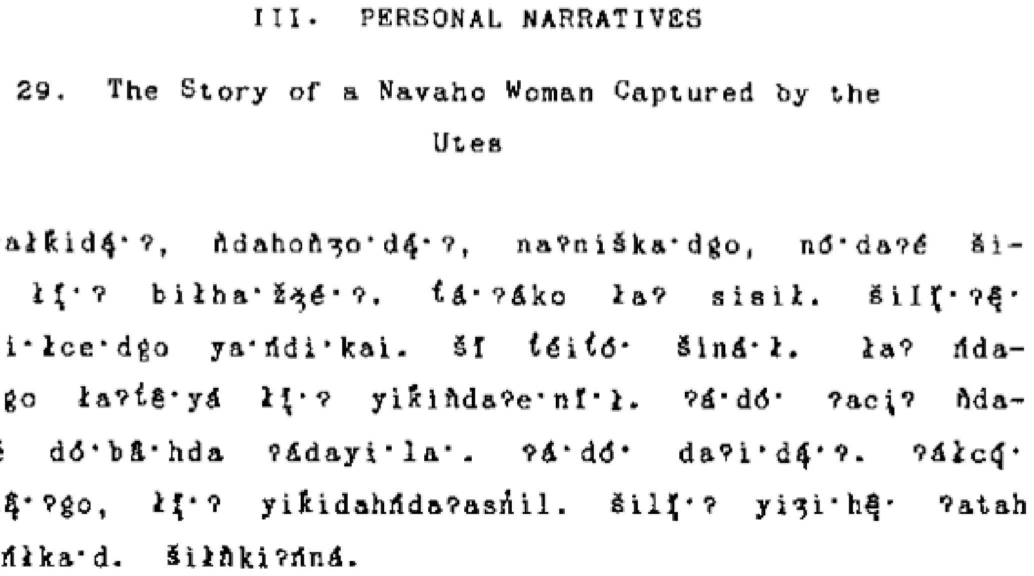
This image is acquired from an inexpensive consumer scanner using a Xerox copy of the page from the book. Although this poor quality input results in many speckle artifacts, these are rarely cleaned using an image editor unless they are of a size greater than a period in the text. Note that the orthography differs considerably from the standard one. For comparison, the first Navajo line from the page in Figure 1 is the following (translated as 'A long time ago, when people where fleeing in small groups (i.e. a time of war), I was herding, when Utes rode up to me on horses.'):

'ałki'dą́ą́', ndahondzoodą́ą́', na'nishkaadgo, nóóda'é shikijį́' łį́į́' biłhaazhjéé.

[3] Post #1091, Tue Jun 1, 2004 5:02 pm, Melissa Axelrod on the lexicographylist, available http://groups.yahoo.com/group/lexicographylist/.

[4] Making Athabaskan Dictionaries Usable, William J. Poser, Athabascan Languages Conference & Workshop on Athabascan Lexicography June 16-18, 2002, Fairbanks, Alaska, available http://www.ling.upenn.edu/~wjposer/.

[5] Sapir, Edward and Harry Hoijer (1942), Navajo Texts. William Dwight Whitney Series. Baltimore: Linguistic Society of America.



III. PERSONAL NARRATIVES

29. The Story of a Navaho Woman Captured by the Utes

ʔałḱidą́·ʔ, ńdahońʒo·dą́·ʔ, naʔniška·dgo, nó·daʔé ši-
kiʒį̨ʔ łį̨·ʔ biłha·žʒé·ʔ. ƛá·ʔáko łaʔ sisił. šilį̨·ʔę̨·
yidaʔni·łce·dgo ya·ńdi·kai. šį́ ƛéiƛó· šiná·ł. łaʔ ńda-
ʔałʔahgo łaʔƛę̂·yá łį̨·ʔ yiḱińdaʔe·ní·ł. ʔá·dó· ʔacį̨ʔ ńda-
yi·sƛé dó·bą̂·hda ʔádayį·la·. ʔá·dó· daʔi·dą́·ʔ. ʔáłcą́·
daʔi·dą̂·ʔgo, łį̨·ʔ yiḱidahńdaʔasńil. šilį̨·ʔ yiʒi·hę̨̂· ʔatah
nikiʔńńłka·d. šiłńkiʔńná.

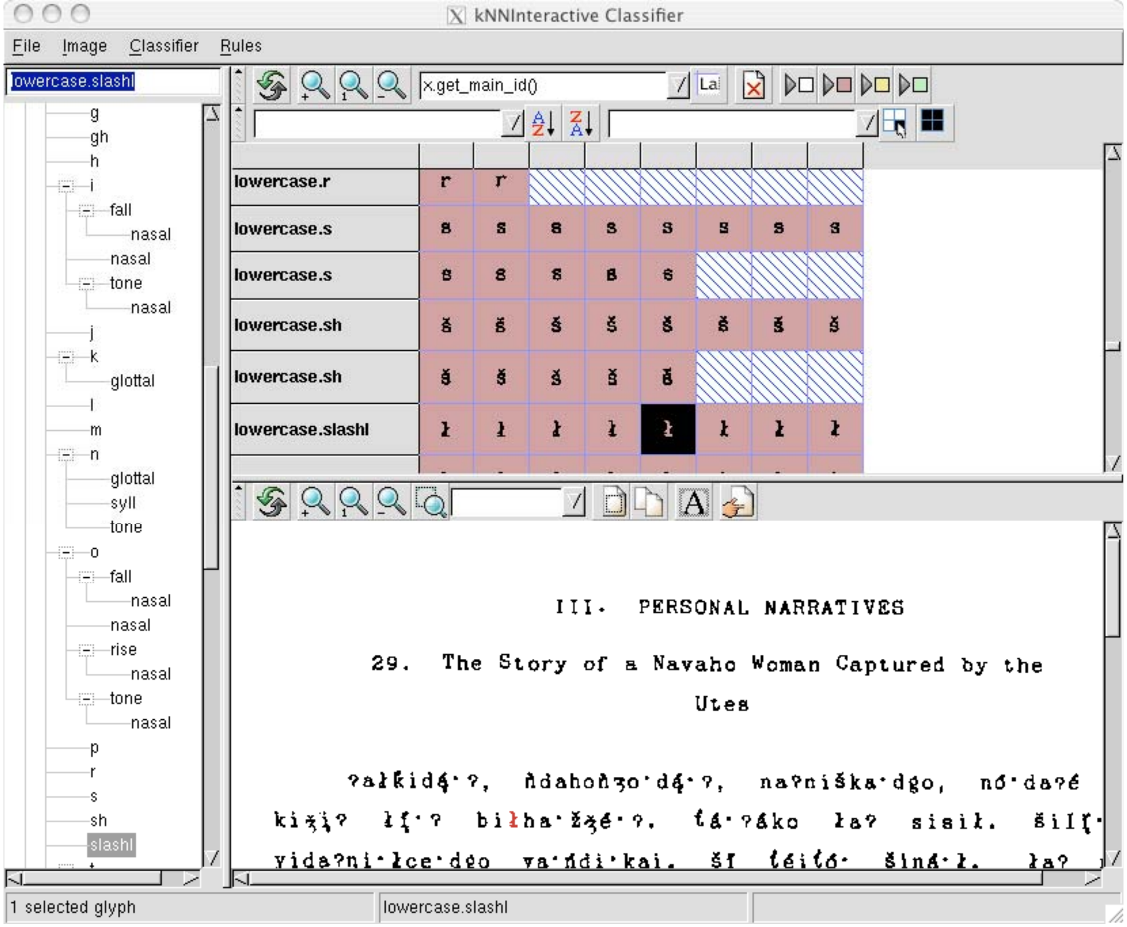
Figure 1. A page from *Navaho Texts*

After the image of a page has been scanned, it must be recognized using optical character recognition (OCR). The usual programs used for this are inadequate since they are pre-trained on typical characters and fonts used for European languages. For this project, the open source Gamera system was used which is written in the Python language.[6] Gamera allows arbitrary characters to be trained using an implementation of the k-nearest neighbor algorithm whose weights are optimized using a genetic algorithm.[7] From the Gamera website: 'Developing recognition systems for difficult historical documents requires experimentation since the solution is often non-obvious. Therefore, Gamera's primary goal is to support an efficient test-and-refine development cycle.' The GUI-based training interface is accessible to non-programmer domain experts. Figure 2 shows the interface that allows iterative classification of characters from actual text images.

[6] Gamera is a framework for the creation of structured document analysis applications by domain experts. It combines a programming library with GUI tools for the training and interactive development of recognition systems. Available http://dkc.mse.jhu.edu/gamera/.

[7] Droettboom, M., K. MacMillan and I. Fujinaga. 2003. The Gamera framework for building custom recognition systems. Symposium on Document Image Understaning Technologies, 275-86, available http://dkc.mse.jhu.edu/gamera/.

Note that in Figure 2, the left-hand side of the window shows that the Navajo characters in the standard orthography have been defined and then can be mapped to characters in the text image. For example, when the slash-L in the text image in the lower pane in clicked, (it appears lighter in the image below and red in the actual interface) that instance of the character is found in the upper pane of recognized characters. If a character is recognized incorrectly, the user can fix it directly in that interface. The more and more pages that are recognized results in fewer and fewer errors. One continues training until the error rate is acceptable.



III. PERSONAL NARRATIVES

29. The Story of a Navaho Woman Captured by the Utes

ʔałk̓idą́·ʔ, n̓dahon̓ʒo·dą́·ʔ, naʔniška·dgo, nó·daʔé
kiʒį̨ʔ łį̨·ʔ biłha·žʒé·ʔ. t̓á·ʔáko łaʔ sisił. šilį̨·
yida?ni·łce·dgo ya·n̓di·kai. šį t̓éit̓ó· šiná·ł. łaʔ

Figure 2. A Gamera Screen shot.

So the inputs for the Gamera system are very accessible to domain experts and require no programming. The output configuration for a particular project requires Python programming, however, to define the character mappings. Gamera offers a rich library of routines to make this fairly simple. A goal for this current Navajo textbase project is to develop a non-programmer interface to the output configuration. The output is a mapping from the recognized characters of the text image to ASCII text that follows the Times New Roman Navajo font.[8] Comments about the success of this OCR step are made in the Results section below.

Once the recognized text is in electronic form, an XML encoding is needed for the text output. The Text Encoding Initiative (TEI) '[…] is an international and interdisciplinary standard that helps libraries, museums, publishers, and individual scholars represent all kinds of literary and linguistic texts for online research and teaching, using an encoding scheme that is maximally expressive and minimally obsolescent.'[9] The TEI is an obvious choice for an encoding, but since it is a standard format, it would be easy to transform the TEI encoded texts into any other format programmatically. Similarly, the text encoding for the Times New Roman Navajo font, can be transformed for any other font should a better one become available.

A sample of *Navaho Texts* that has been encoded using TEI is shown in Figure 3. It is transformed to HTML using the XSLT stylesheet that is available at the TEI website, augmented with a CSS file that includes the Navajo font.[10] That display has been set so that each sentence of the original is on a separate physical line, but that is easily changed. The transformation to HTML allows web access to all the texts in the textbase, but retains the high maintainability of the structured XML source.

The final step in the workflow is to develop and use a lookup tool for the lexicon that allows a user to click on a word and see the correct dictionary page or easily navigate to it. As noted above, looking up words in a Navajo language lexicon is not straightforward. The two major works for the Navajo lexicon are by Young and Morgan and are exceptional print resources.[11] There is also a project to put the *Analytical Lexicon* on-line that is partially completed.[12] The

[8] The TrueType font is freely available from several sources including: http://chinleusd.k12.az.us/navajo.html and http://ling.kgw.tu-berlin.de/Navajo/. The font is described at http://www.sanjuan.k12.ut.us/Fed%20Programs/readmemac.html.

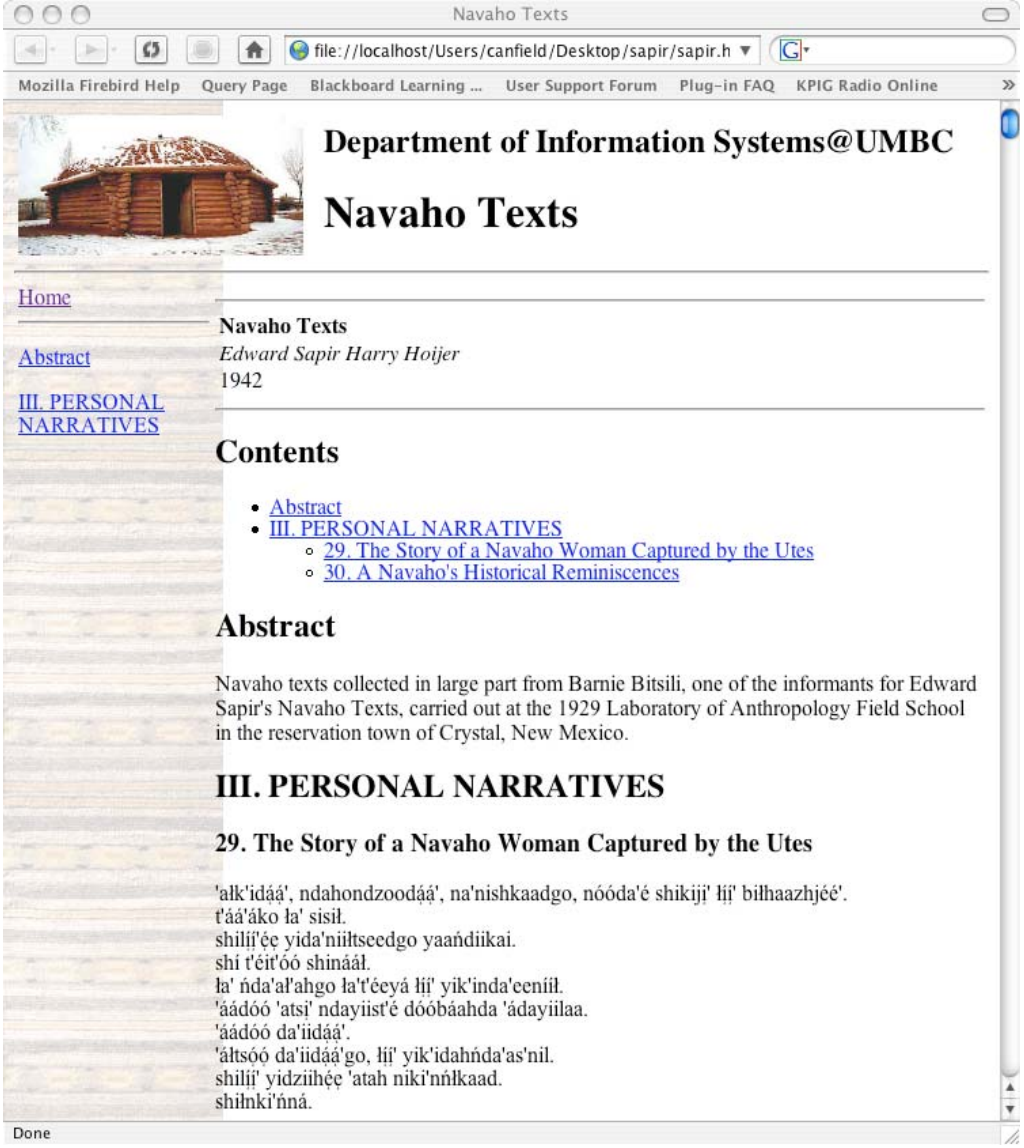
[9] http://www.tei-c.org/.

[10] All documents for this paper can be viewed at http://haggis.umbc.edu/canfield/dine which is available at the publication date. XSLT is the Extensible Style Language Transform standard which can transform any XML document into any other, including an (X)HTML one. CSS is the Cascading Style Sheet standard that allows elements (tags) to be formatted. See http://w3c.org/.

[11] Young, Robert W. and William Morgan, Sr. (1987) The Navajo Language: a Grammar and Colloquial Dictionary. Albuquerque: University of New Mexico Press. and Young, Robert W., Morgan, William Sr. and Sally Midgette (1992) Analytical Lexicon of Navajo. Albuquerque: University of New Mexico Press.

[12] Available http://www.speech.cs.cmu.edu/egads/navajo/. This project was never fully completed as its developers (S. Burke and J. Lachler) note, but it remains a very useful resource.

dictionary lookup tool developed here tries to map a verb stem parsed from a word to a page (URL) in this on-line *Analytical Lexicon* (AL).



Department of Information Systems@UMBC

Navaho Texts

Home

Abstract

III. PERSONAL NARRATIVES

Navaho Texts
Edward Sapir Harry Hoijer
1942

Contents

- Abstract
- III. PERSONAL NARRATIVES
  - 29. The Story of a Navaho Woman Captured by the Utes
  - 30. A Navaho's Historical Reminiscences

Abstract

Navaho texts collected in large part from Barnie Bitsili, one of the informants for Edward Sapir's Navaho Texts, carried out at the 1929 Laboratory of Anthropology Field School in the reservation town of Crystal, New Mexico.

III. PERSONAL NARRATIVES

29. The Story of a Navaho Woman Captured by the Utes

'ałk'idą́ą́', ndahondzoodą́ą́', na'nishkaadgo, nóóda'é shikiji' łį́į' biłhaazhjéé'.
t'áá'áko ła' sisił.
shiłį́į'ę́ę yida'niiłtseedgo yaańdiikai.
shí t'éit'óó shinááł.
ła' ńda'ał'ahgo ła't'ééyá łį́į' yik'inda'eeníił.
'áádóó 'atsį' ndayiist'é dóóbáahda 'ádayiilaa.
'áádóó da'iidą́ą́'.
'áłtsǫ́ǫ da'iidą́ą́'go, łį́į' yik'idahńda'as'nil.
shiłį́į' yidziihę́ę 'atah niki'ńńłkaad.
shiłnki'ńná.

Figure 3. The TEI Encoding.

As noted by Bill Poser in the article sited above, it would be very useful to have a morphological analyzer to pick the verb stem out of a verb and map it to a verbal root in Athabascan since this is a non-trivial process even for native speakers. A simple example of such a parser is presented here. It is always a problem for any natural language processing work to find a balance between simplicity and practicality and completeness or effectiveness since natural language is complex. The algorithm developed here for Navajo does not perhaps balance these as much as come down decidedly on the side of simplicity. The effectiveness is reported in the next section.

Pseudo-code for the algorithm is shown in Figure 4. The algorithm is implemented in the Perl programming language.

```
1. Get the word
2. Look in a list for direct word lookup (no parsing)
3. If found
    a. Display lexical entry
4. Otherwise
    a. Parse the word (assume it is a verb)
    b. Match the longest common substring to a list of all stem shapes
    c. Score each match
    d. Rank the matches by score
    e. Link each stem match to the URL for the corresponding root in the AL
```

Figure 4. The parsing algorithm pseudo-code.

For step 4a, each substring of the verb is compared to a list of all stem shapes. A simple score is attached to each match in step 4c where the score = (index position of the substring) * (length of the substring). This privileges matches that are towards the end of the word and longer substring matches. The ranked matches are displayed with the recommended one being the one with the highest score. Figure 5 shows example output from the batch version of the Perl program that shows a correct parse.

```
Word=na'nishkaadgo :
nish - (12) - http://www.speech.cs.cmu.edu/egads2/navajo/entry?nish
kaad - (28) - http://www.speech.cs.cmu.edu/egads2/navajo/entry?kaad
na' - (0) - http://www.speech.cs.cmu.edu/egads2/navajo/entry?na%27
ni - (6) - http://www.speech.cs.cmu.edu/egads2/navajo/entry?ni
The recommended stem is kaad
```

Figure 5. Parser output.

Figure 5 shows that the highest scored match (28) is the correctly recommended stem. Note that ‘kaad’ gets a higher score than ‘nish’ because the word position of ‘kaad’ is later in the word even though the substrings are of the same length. Enclitics such as the ‘-go’ in Figure 5 prevent stems from always occurring as word final. Also note that the URL to the on-line AL contains still another encoding for Navajo characters (a custom Latin1 mapping[13] that is also URL encoded) and the Perl program must also translate between the standard orthography and this custom mapping.

Results

The results of this case study are in two categories:

- The effectiveness of the workflow
- The effectiveness of the dictionary lookup tool

The effectiveness of the workflow is mostly a subjective call. The scanning procedure is very simple and does not require specialized equipment. Any inexpensive consumer level scanner[14] will suffice. The process of scanning does not require any special linguistic expertise and can be carried out as a batch job that produces the image files. The OCR training and classification process using the Gamera system is fairly straightforward and with the output programming pieces pre-done for a project, it can be performed by domain experts. The training process for the OCR can be extended for any arbitrary level of correctness, but since each text still has to be hand post-edited, perfection is not needed or advisable. A general rule of thumb is to try to minimize the text acquisition cost in terms of domain expert time. The author found that using a minimal OCR training level with an approximately 15% error rate, he could do all acquisition steps of the workflow in under 20 minutes for a physical page and batch pre-scanning would have significantly reduced this time. An open question is whether a skilled typist could beat the OCR times, but such a typist (that has to mentally change the original font to the standard orthography) may be difficult to find and, of course, the output would still have to be post-edited by someone with knowledge of Navajo. The TEI encoding is simple and ensures that the textbase will be archival.

A more formal evaluation of the dictionary lookup tool was performed. A sample of the first 300 words of the text shown in Figure 1 was selected and the Perl program parser was run against this sample in batch mode. This resulted in a list of 300 outputs such as that in Figure 5. The author then checked each of the 300 parses for accuracy. Three categories were used to evaluate this output: correct, incorrect, and non-verb. The non-verb category was used for adverbials, nouns, etc. that do not transparently map to verb stems. These would have to be handled by direct lookup and/or another parser. For example, postpositions could be handled with a special and simple parser that recognized the pronoun prefixes as in *shikiją́'* from the example above. The result of this evaluation was that the parser was 92% accurate for Navajo verbs with a breakdown of: correct=124, incorrect=10, and non-verb=166. So the good news is that the verb parsing was fairly successful and the not so good news is that verbs only formed 45% of the

[13] See http://www.speech.cs.cmu.edu/egads/navajo/wr-sys.html.

[14] A USB Microtek scanner was used for this project that is available for around $US 100.

sample so that other provisions need to be made for non-verbs. Many of the non-verbs can be stop-listed such as adverbials and neuter verbs or handled with special parsers such as that proposed for postpositions. Other types of words are more open-ended and problematic such as the many place names represented in the above text from *Navaho Texts*. The problem is how to generally discriminate between verbs and non-verbs for lookup.

There is room for improvement on the very simple scoring algorithm. An example of an incorrectly parsed verb is:

```
Word=yidziih65 :
dziih - (10) - http://www.speech.cs.cmu.edu/egads2/navajo/entry?dziih
ziih - (12) - http://www.speech.cs.cmu.edu/egads2/navajo/entry?ziih
The recommended stem is ziih
```

The recommended stem for *yidziihę́ę* is incorrect and the scoring fails because the substrings are so very close in length and word position. In this case, a user would have to try both. The parser would still arguably be very helpful to a user. Another example of an incorrect parse is:

```
Word=daash99]zahj8' :
daas - (0) - http://www.speech.cs.cmu.edu/egads2/navajo/entry?daas
zah - (24) - http://www.speech.cs.cmu.edu/egads2/navajo/entry?zah
shÓ - (9) - http://www.speech.cs.cmu.edu/egads2/navajo/entry?sh%ee
jÔ' - (33) - http://www.speech.cs.cmu.edu/egads2/navajo/entry?j%ef%27
hÓÓ - (12) - http://www.speech.cs.cmu.edu/egads2/navajo/entry?h%ee%ee
hÓ - (8) - http://www.speech.cs.cmu.edu/egads2/navajo/entry?h%ee
The recommended stem is jÔ'
```

The correct stem of ‘zah’ for *daashį́į́ńzahjį'* is bypassed for *jį’* which is not only not the correct stem, but is an enclitic. This is due to there being a stem shape that is the same as the enclitic. Note that the raw output of each of the examples above are in ASCII and not a font so the ‘Word’ is in the ASCII representation of the Times New Roman Navajo font and the jÔ' (= *jį’*) is in the Latin1 encoding of the on-line AL.

Conclusions and Future Work

The workflow introduced here appears to be useful for domain experts that are trying to create on-line textbases for Athabascan literature. The acquisition methods are accessible to language specialists and there is no expensive equipment or software required. Ultimately these textbases could be used in the schools for language and cultural studies since they are easily implemented as web pages. The dictionary lookup tool goes at least some distance in solving the long-standing problem of helping users to navigate the complex Navajo lexicon. The link to the on-line AL is a simple example of cross-project interaction in the computational humanities. An updated programmatic interface to the AL would be a significant one!

Significant remaining problems for this project are the programming-required parts of the OCR process and needed refinement of the simple parsing algorithm. A major problem for users will be when a non-verb word is chosen that is not filtered by the stop-list or other parser. These problems are represented in the future work planned for this project:

- Add pages to the Navajo textbase progressively to create a significant resource.
- Expand to other works with different orthographies and document the marginal increase in effort required for each.
- Explore the Gamera OCR effectiveness for hand-written augments to typed text.
- Create software for the OCR output configuration that does not require programming.
- Refine the parsing algorithm for verbs.
- Look at problems with stop-list creation.
- Look into other parsers such as one for postpositions.
- Create a web-based textbase using an XML database and word-clickable lookups.

The general effectiveness of this pilot project workflow can only be really evaluated after it has been applied to many different works with different orthographies. The initial results presented here are encouraging. A textbase also needs an overall architecture that is efficient and maintainable. After a significant number of texts and works are acquired, this project intends to put all of the XML documents into a native XML database that will allow efficient search and organization for the textbase.[15] Finally, the HTML pages that result from the XSLT transform of the XML documents will allow a user to click or highlight a word to access the parser and the lexicon. This requires a very simple modification of the Perl program to operate as a web application server instead of a batch program as in the parser evaluation reported here.

[15] Preliminary experiments for this project have used Exist, an open-source native XML database. Information is available at http://exist-db.org/.