# Considerations for the continuation of a Kansa corpus*

Justin T. McBride
*Northeastern State University*

**Abstract:** More than a decade ago, the Kaw Nation Language Department began a grant-funded project to compile the known Kansa texts for the purpose of creating a graded reader book. This project resulted in, among other things, a unique XML-based electronic corpus of the language. While valuable in terms of the initial project, a closer look at the corpus reveals its numerous practical problems, including general incompleteness, an overly specific purpose, limitations of use and potential users, and compatibility with modern computing tools. Nevertheless, it could be expanded and modified to serve as a much more functional Kansa-English bilingual aligned corpus usable by both Kaw Nation citizens and language scholars outside of the tribe. Various features of such an expanded corpus and its possible development are considered.

**Keywords:** Dhegiha Siouan, bilingual aligned corpus, XML, planning

## 1. Background

The Kansa language, known also as either Kanza or Kaw among members of the Kaw Nation, belongs to the Dhegiha branch of Mississippi Valley Siouan, where it occupies a very close relationship to Quapaw, Omaha, Ponca, and especially Osage. There are presently no L1 speakers of Kansa, but the Kaw Nation has maintained an active and successful language revitalization program for more than two decades. The original project described below was one such activity from this program.

## 2. Original corpus project

As reported in McBride (2009a), the Kaw Nation Language Department received a grant from the Administration for Native Americans in 2008 to develop a graded reader document with accompanying audio files and making use of texts in the language. Less than three dozen in number and collected by such field researchers as Dorsey (c. 1880) and Rankin (c. 1974-2011), these few texts represent—with the exception of two prayers—the whole of extended monologic discourse recorded from L1 Kansa-speaking consultants and span several genres; no dialogic texts for Kansa are known to exist. The development of the reader package for Kansa language learners was a large

*I would like to extend my sincere thanks to the attendees of SCLC 39 for making numerous useful suggestions about the project described herein and especially to the editorial reader who provided thorough and thoughtful feedback on my original manuscript. All remaining mistakes are entirely my own.

project that involved a great deal of planning and yielded several useful products. The choices made during this project from more than a decade ago still remain relevant today.

## 2.1. Purpose and products

In the initial planning for the grant, it became obvious that some of the available texts had greater pedagogical value than others. As such, planners at the Language Department decided that a planned reader and supplementary audio CD would offer only a subset of the texts. Moreover, given the rare opportunity to work so closely with the texts, the three linguists associated with the project—then Language Director Justin McBride, then Language Coordinator Linda Cumberland, and the late Language Consultant Robert Rankin—wanted to analyze them as fully as possible. Thus, the three linguists began a systematic morphemic parse of the body of Kansa texts in order to provide the team's Community Advisory Group of tribal citizens—a group whose role was to offer guidance and support solely for this project before disbanding—sufficient knowledge to make an informed decision on which texts to include in the reader while simultaneously generating comprehensive interlinear gloss material for all of the texts.

While the initial plans involved only the assemblage of various electronic files and an audio CD as grant deliverables, a decision was also made around this time to publish the material developed in the project as two separate print volumes. One of these two print works (i.e., McBride & Cumberland 2009) would be the more inclusive of the two and would contain a detailed morphosyntactic analysis and a corresponding comprehensive morpheme glossary for all the texts. This document would appeal to a somewhat more scholarly audience and would be archived for future reference. Owing to the unexpected cancellation of a mandatory grantee meeting at around the same time, grant funds sufficient to print this volume as a short run of only a few dozen copies became available suddenly. The team jumped on this opportunity, and several copies of the first printed volume were produced and archived while others were given away to team members and tribal administrators; still others were donated to select libraries. The second print volume (i.e., McBride & Cumberland 2010) would be the actual graded reader featuring the smaller set of texts; it would be produced following the conclusion of the grant project to ensure a maximum of federally funded effort went into its creation. Thus, the rest of the initial grant project was spent developing this volume, which included illustrations, grammatical explanations, exercises, a glossary, an accompanying CD of audio recordings, and various other learner resources. Additional funding was then secured from the Endangered Languages Fund to print 500 copies of this document for sale to interested learners; many copies of the reader volume remain for sale through the Kaw Nation at the time of this writing.

The two print volumes themselves, while of potential interest to any number of individuals, were in fact merely products of the background system used to compile the text corpus and its analysis. The selection and use of this system are especially germane to the discussion at hand.

## 2.2. Initial corpus considerations

Project linguists at the Language Department were aware of the fact that detailed work on the Kansa texts would require much effort insofar as the extant analysis was both inconsistent and intermittent. Consider, for instance, that several of the texts collected by Rankin (c. 1970-1979) and especially Dorsey (c. 1880) had been fairly well analyzed, but those analyses were not always theoretically

congruent with each other; Dorsey's (c. 1880) Kansa work, less complete than his work elsewhere in Dhegiha (cf. Dorsey 1890), did not always seem to exhibit great consistency even within itself. Additionally, some texts had never been analyzed. Any analysis associated with them would need to be brought into alignment with both Dorsey and Rankin. Plus, while Rankin's (2008) lexical file for Kansa did manage to feature many items from Dorsey, it was not an exhaustive reconsideration of Dorsey's (c. 1880) Kansa work and did not feature items from such other collectors of texts as Spencer (1908) or Morehouse (c. 1908). In short, a much more uniform processing of all texts was needed.

At the outset of the project, two popular software systems offered the analytical parsing and interlinearization functionality needed to complete the required text processing. One of these was SIL's "The Linguist's Toolbox" (Toolbox), and the other was the American Indian Studies Research Institute (AISRI) at Indiana University's suite of programs including Annotated Text Processor and Indiana Dictionary Database (ATP and IDD). Both of these systems had various pros and cons. For example, while the SIL software was free and Unicode-compliant and offered semi-automatic interlinearization through its seamless text and dictionary integration, it was essentially unsupported by the developers, was difficult to configure for complicated functions, and made use of unintuitive procedural workarounds for routine phonological processes. Meanwhile, although the Kaw Nation staff linguists were already familiar with ATP and IDD and had an established partnership relationship with AISRI, which both developed and supported these programs, ATP and IDD were not so seamlessly integrated, were not Unicode-compliant, were similarly difficult to configure, and were also known to have some nagging performance issues.

In considering the advantages and disadvantages of these two software systems, the team arrived at a simple list of program requirements necessary for simultaneously processing the Kansa texts and compiling an electronic corpus for the language. This list would drive our decision as to which system we would choose to complete the project. The chosen system would need to meet the following feature criteria:

- *Free*—the grant budget did not allow for the purchase of specialized software intended exclusively to help with the analysis of texts;
- *Amply supported*—the project's narrow timeframe was not conducive to attempting overly problematic solutions, and the availability of rapid, high quality support to address possible obstacles was considered critical to project success;
- *Unicode-compliant*—regardless of the fonts chosen for spelling Kansa words, the practical orthography for Kansa makes use of various accented vowel characters, <á, à, é, è, í, ì, ó, ò, ú, ù> (representing primary and secondary stress, respectively), and one high-frequency but potentially problematic character, <ⁿ>, whose use in marking vowel nasalization could not be avoided;
- *Cross-platform*—because the Kaw Nation offices, where the work would be completed, were equipped exclusively with IBM-compatible computers while many Kaw tribal citizens preferred Macintosh computers, the solution had to be compatible with both systems;
- *Self-contained*—the solution could not be part of a larger system that end users would also have to obtain and then learn to use; and

- *Small*—the solution would have to create a text corpus that could be easily shared between team members via USB flash drive or e-mail.

## 2.3. Solution

Rather than choose a pre-packaged application that only met only some subset of these requirements, the team ultimately opted for a self-designed XML-based solution that could be developed entirely in-house and still achieve the desired functionality. Extensible Markup Language, or XML, is a coding system that allows pre-existing text (i.e., the content) to co-exist with embedded computer processing instructions (i.e., the code), the latter of which can be used to manipulate the former. Think, for instance, of how a webpage coded in the related Hypertext Markup Language, or HTML, works to display the content of the page in a particular manner when opened in a browser window. In HTML, the code to make a passage of content text appear italicized begins with an opening "tag"  placed before the text in question and ends with a closing tag  placed afterward, as in 'Hello, friends', which yields '*Hello, friends*' in italics. XML works similarly, but it is different in a very important way: XML is generic and lacks established tag codes for the embedded processing instructions. Rather, the codes and the instructions they represent are largely left up to the coder to define by way of a supplementary 'stylesheet' document; this makes XML potentially much more powerful than HTML, which is limited as to what it can do by its range of pre-existing codes. Given XML's flexible nature, an XML-based approach to developing the Kansa corpus would give the team the freedom it needed to create the corpus it desired directly from word processed versions of the texts. XML, moreover, is free, supported by numerous online developer communities, Unicode-compliant, cross-platform, and self-contained and results in comparatively small files that can be easily shared.

After making the decision to go with an XML-based solution, the team began developing the necessary files to realize this plan. Two XML documents had to be developed along with their two corresponding stylesheet documents. One of these two XML documents would include the texts themselves serving as content. Embedded within, the code for this document would also include the line-by-line and morpheme-by-morpheme parse and any notes or other such supplementary material for each text, such as the consultant's name, the date of collection, reference to any illustrations, etc. To populate the ultimate interlinearization of the texts and to generate a constantly updated glossary, a second XML document would in turn include the morphemic units themselves along with a gloss, lexical and semantic classes, and a numeric code for each. To relate the two documents, the numeric codes from the glossary document were referred to in the text document rather than the morphemes themselves. That way, a change in the morpheme document to any single entry would cascade throughout the interlinear analysis in the text document.

The two associated stylesheet documents were written to generate the desired output materials, namely, 1.) a body of parsed and annotated texts and 2.) a glossary of all the morphemes appearing in these texts; the latter document also drew an example sentence for each entry from the former document and listed the location of the sentence within the corpus. Additionally, 10 illustration graphics resided in the same folder as the two XML documents and their two stylesheets. Even with these graphics, the entire folder was under 1.5 MB and could be easily transferred from computer to computer using almost any sharable media—even the then increasingly rare 3.5-inch double-density floppy disk. The only software needed for any user to access the material in a usable format was a free web browser, e.g., Explorer, Firefox, or Safari, capable of compiling XML code,

which, even at the time, was a standard feature on up-to-date browsers.

Of course, interlinear text is not generally what is meant by the term 'corpus,' which refers only to a body of texts. Nevertheless, a simple modification of the stylesheet—the inclusion of two small tags to demote stylesheet code from instruction to comment—could be used to generate text that could be used with any free concordance software such as AntConc or MonoConc. From this, routine corpus work could be done using the admittedly small body of Kansa texts.

#### 2.3.1. Corpus contents

Using the system described above, the following extended monologic texts were converted for use in the electronic Kansa corpus:

- eight myths (for lack of a better word), thirteen personal histories, and three items of personal correspondence from Dorsey (c. 1880);
- one song from Spencer (1908);
- one transcribed speech from Morehouse (c. 1908); and
- five myths transcribed from Rankin (c. 1970-1979).

The only other pieces of extended monologic text from L1 Kansa speakers consist of two prayers that have never been adequately parsed and are generally assumed to be of a sensitive religious nature. These prayers were omitted from the electronic corpus, as were all other known Kansa materials that did not contain extended monologic texts, such as word lists or even sentence-length elicitation responses.

## 3. Practical evaluation

Several observations can be made about the use of this XML-based solution for the problem of compiling and analyzing Kansa texts and generating pedagogical materials from them. On the one hand, it worked! That is to say, the two planned volumes were successfully produced using the newly developed corpus materials. Additionally, the corpus is still available for additional computer-assisted study of the Kansa language through the viewable front-end output of the XML files, which can be manipulated in various ways via the back-end interface to reveal language data in new and thought-provoking ways. On the other hand, there are still many issues that were never dealt with, some that did not even occur to the planning team at the time of the development of the XML corpus.

### 3.1. Problems

The most obvious problem involves the completeness of the corpus. The original set of compiled and parsed texts included only extended monologs of lengths greater than a single clause which were collected by Dorsey (c. 1880) and Rankin (c. 1974-2011) plus two others (i.e., excerpts from Spencer 1908 and Morehouse c. 1908). This is far from the entirety of sentential material collected from Kansa speakers. The largest source of additional material is Rankin, whose (c.

1970-1979) field notebooks alone conservatively contain over five times the amount of material in the corpus at present; compare the approximately 4,200 lines of known notebook material to the approximately 800 lines of compiled corpus texts. These notebooks document elicitation sessions with three separate speaker consultants, both male and female, and span the better part of a decade. Not all the material is appropriate for rigorous corpus-based analysis insofar as it is often no more responses to requests for single Kansa words or phrases. The speakers often struggle to recall such words, resulting in many false starts and obvious mistakes. Worse still, the clausal material that does appear is merely a response to an elicitation and, in terms of broader discourse-level considerations, unconnected to what comes before or after. Nevertheless, sentence-level material is available for all three of these L1 Kansa-speaking consultants.

Moreover, while working with his primary consultant, Maude Rowe, Rankin stopped collecting material in notebooks and shifted over to eliciting responses straight from photocopies of Dorsey's original Kansa dictionary slip files, handwriting Rowe's responses directly on these copies (Rankin c. 1974). While Rankin did allow the Kaw Nation to make further copies from his annotated Dorsey slip files, these have not been systematically examined to collect sentence-level material for inclusion in any digital document. A smattering of additional material may also be available, for example, in the extensive Bourassa (1843)[1] and Morehouse (c. 1908) collections for the Kansa language. At present, no known clausal material from these collections remains unanalyzed, but more research is needed to be sure.

Another problem arises from the purpose behind the corpus. Specifically, the XML solution was developed for very particular goals involving text-based language pedagogy. This was its primary purpose, and general language scholarship was only a happy consequence. Clearly, the choices made with the pedagogical goal in mind affected the design of the system, which in turn creates obstacles that must be dealt with for more routine corpus work. For instance, it has already been mentioned that modifications must first be made to the stylesheets to generate output usable by standard, third-party concordancing software, which must also be obtained elsewhere. The inconvenience of these first steps makes even a simple keyword-in-context search a tedious process.

Given the built-in purpose, even the range of potential uses is somewhat limited. One logical use of the XML materials, for example, would be the subsequent development of a stand-alone multilingual aligned corpus. This category of corpus includes such corpora as Compara, which is a bilingual Portuguese-English corpus that can be queried in various ways in either language (cf. Frankenberg-Garcia & Santos 2003), or *MulTed*, a proposed multilingual corpus composed of TED talk titles and subtitles (Zeroual & Lakhouaja in press). At present, configuring the Kansa materials in this way would be very time-consuming, mostly because of part of speech tagging

[1]There is no convenient means of citing or even referring accurately to the Bourassa materials. Consider the following personal communication from Ives Goddard from August 11, 2008, alerting Robert Rankin of their existence: "The Cullman library in the Smithsonian Natural History Museum has acquired a ms. with vocabularies of Potawatomi, Ottawa, and 'Kaw' which is annotated by Wilberforce Eames but apparently copied by someone else from original mss. of Joseph N. Bourassa. (A ms. related in some way is in the Pequot library in Conn.) The ink is faded and often hard to make out even with the naked eye, but much is readable and interesting. The 'Kaw Dictionary' (on pp. 163-183) is probably copied from the one listed for Bourassa by Pilling (then in the possession of John B. Dunbar), and at least one of the Potawatomi sections is presumably a copy of the Potawatomi vocabulary that Pilling also gives as Dunbar's. The Ottawa is sandwiched between two Pot. sections in the copy we have, and as it is not labeled as such the exemplar may have gone unidentified." After receiving this message, Language Department staff obtained a photocopy of the 'Kaw Dictionary' excerpt mentioned above directly from Goddard at the Smithsonian Institution while on a work-related trip to Washington, DC.

for the Kansa, which would have to be done manually; the current tagging is not strictly at the word-level. The parsing of the Kansa texts is currently morpheme-by-morpheme, meaning that all morphologically complex words in the texts would have to be coded for word-level lexical class. This would require a new level of interlinear analysis that would have to be developed for every word of every text. Note, by the way, that the corresponding English tagging would not be as difficult—it could be done automatically through a part of speech tagging program such as TagAnt or TreeTagger—but an additional line of analysis would have to be added to accommodate the tags for English just as with Kansa. This is to say nothing, of course, of the theoretical concerns about lexical class in Siouan as a whole. For instance, some may argue that Kansa has no adjective class, but only stative verbs, while others may disagree; it is impossible to expend the effort on tagging or expanding the available analysis without opening numerous of such cans of worms, and the results could potentially decrease the potential number of users.

The current XML-based solution already has a very, very small number of users. While the McBride & Cumberland (2009) volume includes the current analysis resulting from the corpus compilation, less than 40 copies were published, and many of these copies reside in archives or are owned by individuals who may lack the necessary experience with interlinear analysis to make ample use of it. This means that the work done for the corpus project is mostly left up to users of the XML-based source files. Given that the coding structure is unique, that there is no convenient query interface (corpus queries can be approximated by simple search functions in word processing based off of the numeric codes associated with individual morphemes), and that manipulation of the source files requires learning XML, any use of the source files outside of their primary purpose involves a steep learning curve. There may only be a handful of people comfortable using these files for anything despite the potential value the files may possess.

Finally, the XML files are no longer easily viewable on browsers. While the display of lengthy local XML code by way of an associated local stylesheet is still possible on some browsers (e.g., Edge and the now obsolete Internet Explorer), it is rare enough that extra steps must taken to do so on some browsers (e.g., Chrome requires Document Type Definitions to compile the files and recommends use of its XML Viewer extension), and it is simply not possible on others (e.g., Firefox).

# 4. Present considerations

With so much material that could be converted for use with the Kansa corpus documents, and with so many serious complications associated with the current corpus, it would seem that the project stands at a crossroads. In order to decide how to proceed, several questions must be answered.

## 4.1. Who would use such a corpus and to what ends?

Clearly, the project should benefit Kaw Nation citizens first and foremost. The work, after all, began as a grant-funded tribal project to produce materials for tribal learners and made extensive use of resources furnished by the tribe to facilitate its completion. Any derived products would also need to be geared toward primary use by Kaw Nation citizens, presumably as an interactive archive of knowledge relating to their heritage language. The final product should, therefore, be targeted to an audience composed mostly of non-specialists in language study without relying heavily on

theoretical terminology, niche technology, or impractical functionality that cannot readily advance language learning. As before, an advisory group composed of tribal citizens could be assembled to provide guidance on how work on the project should progress, all the while keeping the tribe's best interests in mind.

Secondarily, the design should permit Siouan scholars, language professionals, and others with a vested interest in the promotion of understanding of Kansa and its related languages to achieve their goals. For example, while simple corpus tasks such as keyword searches or even line-by-line navigation through texts should be obvious to non-specialist users, the functionality must be robust enough to allow for much more complex use of the language data in ways that may not immediately occur to such users. The interface design must also not appear to hide such functionality from non-specialists. Moreover, Dorsey's (c. 1880) and Rankin's (c. 1974-2011) Kansa materials, which are of special interest to scholars for their relative regularity and overall trustworthiness, should be as comprehensive as possible within the data; materials from as many others as can be managed should also be included.

Another feature that would be potentially valuable to scholars would be the ability to toggle between the practical spellings of Kansa and Siouanist phonemic transcriptions; the former are actually derived from the latter, but they can obscure more complex phonological goings-on within the language, especially with regard to phenomena that may be of cross-linguistic interest. For example, the practical Kansa <p, t, k> characters correspond to the Dhegiha 'tense' stop series that is realized as /pp, tt, kk/ in Kansa, Omaha, Ponca, and Quapaw, and as /hp, ht, hk/ in Osage—not as plain /p, t, k/ in Osage and Quapaw, which correspond to /b, d, g/ in Kansa, Omaha, and Ponca. Consider, for instance, the word for 'similar, alike' in Osage /kǫzékǫ/ and Kansa /góze égo/, practical <góze égo>, where the plain stops surface in both languages; the plain Osage stop /k/ corresponds to Kansa phonemic /g/ and practical <g>. But, consider 'teaching, religious devotion' in Osage /hkihkǫ́ze/ and Kansa /kkikkáze/, practical <kiká$^{n}$ze>, where the tense stops surface in both languages; the tense Osage stop /hk/ corresponds to Kansa phonemic /kk/ yet practical /k/. On a similar note, <aá$^{n}$> in the Kansa practical orthography corresponds to a long, nasal vowel with falling pitch, which Rankin tends to represent as /ą̀ą̀/ in his (1974-1975) notebooks. As such, the practical spelling of the name of the tribe and the language, <Kaá$^{n}$ze>, which tribal citizens have come to accept, corresponds to the more familiar Siouanist transcription /kkąą̀ze/, which tribal citizens may not even recognize as the same word. While these orthographic concerns may appear minor, the division between practical and technical spellings for Kansa is an important one, and it has been the subject of intense internal debate and planning (McBride 2009b).

## 4.2. What form should it take?

This single question is in many ways far more nebulous than the first. One the one hand, a list of desirable design features should be easy to come by. On the other hand, some of the most fundamental considerations that would be helpful for such a design wish list have never been dealt with. For example, it is not clear at present if converting the remaining data to the current corpus format, which could then be modified wholesale to develop the desired corpus tool, is preferable to starting essentially from scratch, saving only those parts of the current corpus needed for the optimal end state, whatever that may be. At any rate, there are still some things that must be accomplished.

#### 4.2.1. Bilingual alignment—for a start

The end product must involve a multi-use bilingual corpus interface that involves multiple levels of analysis. At the most basic level, parallel text must be available for practical and/or phonemic Kansa on one hand and a corresponding English gloss on the other hand. An interlinear parse, perhaps featuring a user-defined depth, should also be immediately available. A close phonetic transcription option may also be desirable, but it would not be available for any of the texts save for those collected by Rankin, whose (c. 1970-1979) audio recordings of the elicitation sessions survive.

#### 4.2.2. Robust user functionality

The version of the product finally released for public use must include various searching, sorting, multimedia, etc. tools usable as necessary by its two main groups of end users, Kaw tribal citizens and language scholars. One potentially very valuable search tool that would be of interest to scholars is the ability to search for two items in the same sentence. In Dhegiha, determiners associated with the subject often—but not always—take the same form as auxiliaries in the predicate; being able to search for both—either by lemma or by part of speech tag—could help to clarify the reasons for this. Such functionality could be achieved by allowing additional n-gram searches on initial search results. Frequency-based sorting of concordance results would also be very helpful, as would the ability to align audio recordings, where available, to Kansa sentences. Beyond this, routine corpus tools, wherever possible, should be included.

#### 4.2.3. Additional information about texts

Elicitation conditions must be made clear. Highly contextualized and discursively complex material from an extended monologic text may be found in the data right alongside single-sentence responses to very simple elicitation requests. While each is valuable in its own right, an instance of the second lacks the cohesion and coherence of a single sentence from the first; comparing the two is a proverbial apples-to-oranges scenario. Meta-data on the corresponding texts, speaker consultants, collection dates and times, and other such general reference data, must be recoverable from any single line of text or even individual words from it. Recovery of second-order data as this can provide users the analytical context needed to judge how best to interpret the primary data.

### 4.3. How should it be delivered?

This question flows from the last one, but specifically frames the consideration in terms of which technological solution will maximize ease of the use while also minimizing conversion obstacles and possible errors. On the one hand, one obvious solution would be a web-based platform making use of SQL databases for storing the data and an interface whose functionality would be enabled by PHP scripts calling on the data. This sort of system is far more common these days than the XML documents and related stylesheets found in the original corpus. On the other hand, migrating the entirety of the current system to a format that is not at all similar would be tedious and time-consuming. Moreover, a stand-alone app usable on mobile devices may be even more popular for end users. Given the very small file size of the current corpus, such a solution may be particularly efficient. At any rate, some of these questions could be deferred until such time as another

advisory group or other such ad hoc tribal panel could be assembled to provide culturally sensitive suggestions and guidance.

# 5. Conclusion

In this paper, I have attempted to demonstrate some of the considerations associated with the creation and possible continuation of an electronic corpus of Kansa texts.

### 5.0.1. Summary

The Kaw Nation Language Department developed an electronic Kansa corpus more than a decade ago for the purpose of creating pedagogical materials for Kaw tribal citizens. This original corpus is still operational, but it suffers from several deficiencies. While representing nearly the whole of extended monologic discourse in Kansa, it is noticeably incomplete, having been developed from many different sources of many different kinds, and its purpose and resultant structure at present are too narrow for a wide range of uses; as such there are few, if any present-day users. It is also difficult to access today given how technology has advanced since its initial development. Nevertheless, given both the incompleteness and the trove of additional material that could be adapted for use with the corpus, now is an ideal time to review its condition with an eye toward its possible use in the future.

Balancing the interests of the two primary stakeholders while expanding on the original project is obviously a key concern here. On the one hand, the corpus should continue to be used for the primary benefit tribal members seeking to help revitalize their heritage language or who may simply wish to learn more about what their fellow tribespeople had to say *in their own words*. Part of this tribe-centered purpose would also directly benefit the Language Department who are constantly looking for new and exciting—but ultimately simple—ways to generate meaningful pedagogical content for their students. On the other hand, it could also be used by scholars and language teachers and learners from outside of the Kaw Nation who, although somewhat secondary to the main purpose, may find the prospect of relatively unfettered access to a body of lesser known Dhegiha Siouan texts appealing and who may use the corpus to advance particular theoretical or practical goals.

Assessing the situation from these vantage points yields a veritable wish-list of features stemming from the initial development of the corpus, yes, but also poised to govern all aspects of its expansion. The resultant system should have the following features:

- *Free*—The corpus should cost no additional money to develop, maintain, and access (implicit here is that the responsibility for development and maintenance should remain with the Kaw Nation Language Department and its affiliates as directed by a panel of concerned tribal citizens—just as it was at the beginning of the project);
- *Supported*—The technology driving the corpus must enjoy ample development support;
- *Unicode-compliant*—Representation of the language, especially if fields are added to allow for technical phonemic transcription, requires an expanded range of characters in any typeface used for the project (implicit here is that the typeface should be free for the end user to access);

- *Cross-platform*—Just as before, the corpus should be usable on many different platforms (implicit here is that mobile devices, which were not a concern at the time of the initial development, are likely to be a major driver of choices in the expansion of the corpus);
- *Self-contained*—Now more the ever, the corpus should not rely on external resources to use (implicit here is that, if downloadable, development of the expanded corpus should include some sort of installation tool to ensure that the various components are in place and working properly, in terms of both what is available at the time of initial download and also what may be added later as a result of updates);
- *Small*—The corpus should have a small digital footprint (implicit here is that whatever delivery technology is used for expanding the corpus does not unjustifiably add to the overall size of the accessible content);
- *Primarily Kaw-oriented*—The expanded corpus must benefit Kaw Nation citizens primarily (implicit here is that Kaw Nation Language Department staff members are likely to be its most prolific users, but those without technical skills in language description or teaching will ultimately benefit from its expansion the most);
- *Secondarily academic*—The corpus should benefit other Kansa language scholars, teachers, and learners, albeit secondarily (implicit here is the assumption that Siouanists are likely to be among its users, and all onboard functionality should at least be congruent with popular theoretical and practical understandings of Siouan languages);
- *Reflective of different attitudes regarding spelling*—The actual content of the expanded corpus potentially alienates prospective users unless both practical and technical spellings are recoverable (implicit here is that a toggling function and the underlying mechanism for ensuring its effective use must be built-in to either the content or code or both);
- *Bilingual*—At a minimum, the corpus should feature bilingual alignment between Kansa and English sentences where appropriate, but additional functionality may extend to finer-grained levels of analysis on either side (implicit here is the belief—which may well be unfounded—that discourse in one language may align sentence-by-sentence with discourse in another language, and perhaps even at lower levels);
- *Feature-packed*—The corpus must provide robust user-functionality (implicit here is that a survey of potential Kaw and non-Kaw users may need to be conducted to be sure of which features will be the most effective for achieving specific goals);
- *Metadata-packed*—The corpus must provide additional information about the texts sufficient for understanding the place of a single sentence within the corpus—either as a stand-alone item or as an element of a larger discourse—and the circumstances of its utterance wherever known (implicit here is a more or less complete understanding of these considerations, not to mention an efficient means of encoding that understanding into the system); and
- *Convenient*—Along the same lines as self-contained and small above, the corpus must be accessible in a convenient and popular digital format (implicit here is that technologies are known to change quickly, and that subsequent development take a long-view with respect to future use of the corpus).

To ensure that all of these conditions are met in the final product and that the corpus retains its usefulness for Kaw citizens, it is also essential that a panel of such tribal stakeholders provide some degree of oversight on the project, exactly as was done before.

### 5.0.2. Looking forward

Until a new advisory panel can convene to provide definitive guidance on how to proceed under the following conditions, entering clausal material from the Rankin (1974-1975) notebooks into the current corpus scheme should not be difficult. What is more, the available morphemic data can already be used to populate much of this new material without having to add new morpheme entries. Completing this task for the available notebooks, therefore, should prove a satisfying preliminary step before work could commence on locating other such sentential data in the Rankin (c. 1974) dictionary materials or the materials collected by other researchers; for instance, looking for heretofore unknown clausal material in the Bourassa (1843) or Morehouse (c. 1908) materials would be an excellent idea. Provided that additional texts for inclusion in the corpus cannot be found among these physical sources of written Kansa, the extensive audio recordings Rankin's (c. 1970-1979) dictionary elicitation sessions could additionally be re-transcribed in an effort to locate clausal material. Whatever the case may be, it is hoped that, by the time the notebook material has at last been entered, a more permanent form for the Kansa corpus will have presented itself.

## References

Bourassa, Joseph N. 1843. A vocabulary of the Po-da-wahd-mih language [manuscript]: With illustrative sentences and a translation of the first three chapters of the Gospel of Matthew, followed by a vocabulary of the Kaw language. Ms. PM2191 .Z5B68 1843a. Washington, DC: Joseph F. Culman 3rd Library of Natural History, Smithsonian Institution.

Dorsey, James Owen. 1890. The Ȼegiha language. *Contributions to North American Ethnology* 6. 1–794.

Dorsey, James Owen. c. 1880. Kansa texts. James O. Dorsey papers. Suitland, MD: National Anthropological Archives, Smithsonian Institution, MS 4800, Box 33, Items 245-247.

Frankenberg-Garcia, Ana & Diana Santos. 2003. Introducing COMPARA: The Portuguese-English parallel corpus. In Federico Zanettin, Silvia Bernardini & Dominic Stewart (eds.), *Corpora in translator education*, 71–87. Manchester, UK: St. Jerome.

McBride, Justin T. 2009a. Compilation of Kansa texts: Overcoming challenges in the drafting of a graded reader. Paper presented at the 29th Annual Siouan and Caddoan Langauges Conference. University of Nebraska-Lincoln, 12 June. Lincoln, NE.

McBride, Justin T. 2009b. Orthography and ideology: Examining the development of Kaw writing. In Daisy Rosenblum & Carrie Meeker (eds.), *Proceedings from the 12th annual workshop on American Indigenous Languages*, vol. 20 Santa Barbara Papers in Linguistics, 30–45. Santa Barbara, CA: University of Santa Barbara Department of Linguistics.

McBride, Justin T. & Linda A. Cumberland (eds.). 2009. *Compiled Kanza texts*. Kaw City, OK: Kaw Nation.

McBride, Justin T. & Linda A. Cumberland (eds.). 2010. *Kaáⁿze wéyaje (Kanza reader): Teaching Kanza literacy through historical texts*. Kaw City, OK: Kaw Nation.

Morehouse, George Pierson. c. 1908. eorge pierson morehouse papers. Manuscripts Collection 453. Topeka, KS: Kansas State Historical Society.

Rankin, Robert L. 1974-1975. Series 2: Kaw (Kansa, Kanza); 2.1: Field notebooks. Robert Rankin papers. NAA.2014-16, Box 2. Suitland, MD: National Anthropological Archives, Smithsonian Institution.

Rankin, Robert L. 2008. Kansa-English lexical file. Ms. Lawrence, KS: University of Kansas.

Rankin, Robert L. c. 1970-1979. Series 2: Kaw (Kansa, Kanza); 2.3: Sound recordings. Robert Rankin papers. NAA.2014-16, Boxes 51-54. Suitland, MD: National Anthropological Archives, Smithsonian Institution.

Rankin, Robert L. c. 1974. Series 2: Kaw (Kansa, Kanza); Kansa-English dictionary vocabulary slip files. Robert Rankin papers. NAA.2014-16, Boxes 42-45. Suitland, MD: National Anthropological Archives, Smithsonian Institution.

Rankin, Robert L. c. 1974-2011. Series 2: Kaw (Kansa, Kanza); 2.2: Files. Robert Rankin papers. NAA.2014-16, Boxes 3-4. Suitland, MD: National Anthropological Archives, Smithsonian Institution.

Spencer, Joab. 1908. The Kaw or Kansas Indians: Their customs, manners, and folk-lore. *Transactions of the Kansas State Historical Society* 10. 373–382.

Zeroual, Imad & Abdelhak Lakhouaja. in press. MulTed: A multilingual aligned and tagged parallel corpus. *Applied Computing and Informatics* .