MayanWiki: Facilitating Consensus and Linguistic Research through an Openly Editable Corpus[1]

Robbie Haertel
Brigham Young University

1. Introduction

The writing system used by the ancient Maya civilization has intrigued researchers and aficionados for centuries. Now that it has mostly been deciphered, the emphasis in the field of Mayan epigraphy has shifted to a study of the system of phonological, morphological, and grammatical rules that once governed the language that the hieroglyphs encode (Haertel, 2007). Linguistic study of this type could be facilitated by a publicly available, comprehensive, electronic corpus of texts to investigate phraseology, frequency information, and collocations, as is done in more widely studied languages such as English. Such a resource would assist not only Mayan epigraphers, but linguists, archeologists, anthropologists, students, and hobbyists. However, a corpus of the hieroglyphs presents special challenges. For one, new texts are continually discovered. More importantly, since Mayan linguistic epigraphy is in its infancy, there is considerable disagreement concerning such issues as phonology, morphology, etc. Unfortunately, a privately run database reflects only the viewpoints of the maintainer and is difficult to manage under these circumstances. Such a corpus does not necessarily serve the community as a whole; instead, a corpus with decentralized control is needed.

A wiki provides the decentralized control necessary to address these issues. In particular, a wiki based corpus could accelerate the convergence of readings to a consensus if it were used and edited by enough people with sufficient expertise. However, a wiki is only the means by which data is added to the corpus and subsequently modified. In order to be a truly useful linguistic tool, the corpus must also allow for detailed data analysis through sophisticated search tools. This paper discusses MayanWiki as a solution to this problem. MayanWiki is a corpus of transcribed and transliterated hieroglyphic texts intended for linguistic inquiry. MayanWiki employs a wiki for data entry and modification and a relational database that has been specifically designed for efficient storage and retrieval of the data for detailed linguistic analysis. Once the database is more fully populated by users, it will become a valuable tool allowing the textual data to be manipulated in ways that will facilitate scientific discovery of new and interesting linguistic patterns. Most importantly, it will continually evolve to reflect the latest research in the field.

1 This paper is a revised and shortened version of the author's Master's thesis in Linguistics, presented 2007 to Brigham Young University (Haertel, 2007).

This paper begins by more thoroughly motivating the need for a wiki based approach combined with a carefully designed database. To this end, it is first necessary to discuss the basic unit of annotation, which in MayanWiki is the grapheme for transcriptions and the morpheme for transliterations; Section 2 briefly addresses this issue. Section 3 then establishes criteria for a useful corpus based on these levels of annotation and Section 4 examines previous work in terms of their failure to meet these criteria. A discussion of how the wiki and database in MayanWiki meet these criteria appears in Section 5, which also includes a brief justification for the use of a relational database for implementation. Discussion regarding the plausibility of a wiki based corpus follows in Section 6; this section also discusses the role of students and hobbyists in a wiki based system. Finally, a conclusion is given in Section 7 that includes consideration of the applicability of a wiki to other corpora.

2. Basic Level of Annotation

Before the creation of any corpus, it is first necessary to decide on the appropriate unit for annotation. Possible units for transcription include glyphs, characters, morphemes, or words, but could also be as fine-grained as sub-characters or as coarse as phrases, sentences, or even entire documents. Use of a more fine-grained representation typically provides more information than coarser representations. For instance, when glyphs are chosen as the basic unit, information regarding spaces between glyphs (including possible alternative word boundaries) can easily be represented. However, when words are the basic unit, information regarding spacing between glyphs is usually lost. On the other hand, more fine-grained representations are typically much more tedious to obtain, and are thus usually more costly. It is often the case that the more fine-grained units must be grouped by an annotator in order to form the more coarse-grained units. The more fine-grained the unit, the more difficult and time-consuming this can be. Hence, when choosing the basic unit of annotation, a balance must be sought between the information that can be extracted from the corpus and the amount of work required to obtain this information.

One problem regarding the unit of annotation for a corpus of Mayan hieroglyphs is the lack of support for computerized representations. There is no standard font that contains all of the hieroglyphs, and there are certainly no “hieroglyphic keyboards”. This is a problem in general for syllabic and logographic languages like Mayan, and also a problem for many other ancient languages. Certainly, scanned images could be used to represent glyphs in texts, but these can be prohibitively time consuming to obtain and include in a corpus. To address this problem in the field, glyphs are typically transcribed with their phonetic value using Roman characters. Then, based on a set of invertible spelling rules, this transcription is transliterated into

standardized Mayan morphemes and words (using a phonemic alphabet consisting of Roman characters).

As previously mentioned, the focus of Mayan epigraphy is on understanding linguistic aspects of the hieroglyphs. The form of data most favorable to thorough linguistic study is the data derived from transcriptions and transliterations which contain phonetic and grammatical information. Thus, the basic units of annotation in MayanWiki are the grapheme for transcriptions and the morpheme for transliterations. Since texts can be transcribed and transliterated relatively easily, a larger amount of data can be obtained from these than from other forms of annotation and the data will be directly relevant and useful to linguistic study. Of course, this means that, without additional annotation, the data in MayanWiki cannot be used to answer certain questions not directly related to linguistic inquiry, such as the distribution of particular variants of syllables or logograms.

3. Criteria for a Useful Corpus

In order for a corpus of transcriptions and transliterations of Mayan hieroglyphic texts to be useful for the study of language, certain criteria must be met; not just any corpus will suffice. The first criteria and ultimate goal are that the corpus should be comprehensive and in electronic form. Furthermore, it is also necessary that the entire corpus be publicly accessible from a single central location. Yet, a central corpus often introduces additional problems if privately maintained, namely that it is not consensus-based, it is difficult and expensive to maintain, conflicting submissions are difficult to resolve (although privately maintained databases typically do not allow submissions), and there are licensing issues; these problems are discussed more thoroughly below. Hence, it is necessary that control and responsibility of the corpus be decentralized. Finally, a useful corpus must be designed to facilitate meaningful linguistic study through the application of corpus linguistic principles. The latter three criteria are explained in further detail in the following sections.

3.1 Central Access

Currently, there is no publicly available, large-scale, electronic collection of transcriptions, transliterations, or translations of Mayan hieroglyphic texts. Surprisingly, no effort has been made to create even a non-electronic corpus of transcriptions. What little data exist are scattered across multiple publications. These two problems, lack of coverage and lack of centrality, cripple the progress of the field. Under current circumstances, it is necessary to manually locate material that contains transcribed texts (which will in turn require searching the archives of several distant libraries), and then to scour the thousands of pages of print to extract a few transcriptions. Often, the transcriptions are out-of-date or otherwise incorrect. This process is time consuming, expensive, and unreliable. Even when the texts have been collected, it is very difficult to manipulate the data in ways that

can lead to new insights. In short, the current situation strongly resembles corpus-based studies of yesteryear that have been derogatively labeled 'pseudo-procedures' (Abercrombie, 1965).

Electronic resources can cause similar difficulties when they are scattered across multiple web sites, or even fragmented within a single web site through multiple search engines or poor search facilities. For effective research the corpus must be available from a single central location, with a single, useful search engine. Sometimes this involves a meta-search engine that will collect the data from various sources. However, only a very few transcriptions are currently publicly available, and hence a meta-search engine would be of little use.

3.2 Decentralized Control

In most cases a central database—like the one needed for the hieroglyphs—is privately populated and maintained by the owner of the database, frequently a single researcher or a few collaborators (hereafter referred to as the "maintainer"). This is problematic for several reasons. First, a database maintained by a small group is inherently not consensus-based. This is important in a field like Mayan epigraphic linguistics where disagreement and uncertainty abound. There currently are, and probably always will be, at least a few respected researchers who disagree about transcriptions, spellings conventions, morphological analyses, etc. Because of their misgivings, these researchers are unlikely to use a database maintained by someone with differing views, thereby undermining the purpose and existence of a central resource. Little progress is made under these conditions.

Even if one were to suppose that a single maintainer is capable of producing a resource that is widely used, the burden of updating the database to reflect current research and discoveries of new texts lies with that maintainer. For instance, imagine that an archeologist-epigrapher discovers several new texts during an excavation. He or she would then need to send photographs or drawings and optionally a transcription and transliteration to the owner of the database. The owner of the database would then need to perform the onerous task of importing the data (if they even care to do so), and even transcribing it in the case that no transcription was provided. A similar scenario would occur with the publication of a new article, which could necessitate a large number of changes in the database. Few people have the time available to make such changes and additions to the database on a continual basis, especially considering that "submissions" would be coming from multiple submitters, often simultaneously. This is probably why private databases rarely accept submissions. Even if a private maintainer had the time and funding necessary to perform this task, it will certainly take longer for the data to appear in the database than if the original submitter had added it directly to the database.

This leads to the third issue: a privately maintained database has no mechanism for resolving conflicting submissions. Usually, the maintainer's preference would

be used which, as mentioned previously, will frustrate use of and submission to the database by researchers that disagree with the maintainer's point of view.

The final potential problem with a privately maintained database is that people would probably only be willing to submit artistic data if their work was attributed to them and if they were able to own the copyright—at least for the photographs and drawings. Unfortunately, privately-maintained databases rarely offer this type of control.

However, a central resource need not suffer from these problems simply because it is central. The key is to allow access to the central database while keeping ownership and maintenance of the content decentralized. This means that the content is stored in a single database and browsing and searching the texts are done from a single place (i.e. program or web page), rather than requiring that users collect linguistic data across multiple databases or sites. However, anyone, including hobbyists and non-specialists, should be allowed to add, edit, and otherwise contribute content to the database in a way that facilitates collaboration, but remains consensus-based.

3.3 Linguistic Investigation

Surely, any corpus that is to be useful for the study of language should of course be searchable in ways that are linguistically meaningful. Given the success of corpus linguistics, particularly in the last twenty years, any corpus not based on sound corpus linguistic principles would be inadequate. Since a corpus is only as valuable as the information that can be extracted from it, even a well-designed corpus that is stored efficiently is useless if the access software does not provide the ability to mine the available information in meaningful ways. In other words, the value of any corpus depends not only on its content, but on the ease with which the contents can be manipulated and searched. In Hunston's (2002) words:

> If a corpus represents, very roughly and partially, a speaker's experience of language, the access software re-orders that experience so that it can be examined in ways that are usually impossible. A corpus does not contain new information about language, but the software offers us a new perspective on the familiar. (p. 3).

With a well-designed database, and appropriate access, creative minds are able to manipulate and transform data in ways that can shed new light on old problems, inspire new hypotheses, and provide evidence for new and existing theories.

Within the context of linguistic corpora, and particularly computerized data, the three principal ways in which corpora are re-ordered and manipulated is through the study of frequency, phraseology, and collocation (Hunston, 2002). The frequency or relative frequency of a word can be used to compare the distribution of words and phrases in different sub-sections of a corpus; for instance, in monumental versus vessel inscriptions or Early Classic versus Post-Classic writings. Phraseology is most often studied through concordance lines which "bring together many instances of

use of a word or phrase, allowing the user to observe regularities in use that tend to remain unobserved when the same words or phrases are met in their normal contexts." (Hunston, 2002, p. 9). Collocation is a similar concept, but with an emphasis on identifying statistical tendencies of words that co-occur and thus entail meaning not necessarily present in individual occurrences of the words. A corpus of the hieroglyphs should minimally allow these manipulations of the linguistic and glyphic data, both in the way the data are stored and through the access software.

4. Previous Work

MayanWiki is not the only corpus of hieroglyphic transcriptions to have been proposed. This section uses the criteria from the previous section to discuss the three most recent attempts at creating a publicly available corpus of transcriptions: the Maya Epigraphic Database, the Maya Hieroglyphic Database, and the Maya Hieroglyphic Codices.

The Maya Epigraphic Database (Alvarado, 1994) represents a milestone in the creation of a corpus of the hieroglyphic texts as "an experiment in networked scholarship." Besides clearly enumerating the benefits of a computerized resource available on the Internet such as, "replicability, searchability and transformability," the creator also recognizes the importance of centralized access and decentralized control as explained above:

> [...] the archive is in an equally real sense a public and collectively authored entity. In principle, all transcriptions are submitted individually and edited collectively. The sharedness of the medium means that transcriptions will tend to be standardized according to the consensus of participants (Alvarado, 1994).

This includes the recognition that, "Disagreements are of course to be expected, and indeed applauded." (Alvarado, 1994). This database can in many ways be considered the most influential predecessor to the current work.

Unfortunately, despite such a mature point of view on the need for a collectively created, consensus-based corpus, after over ten years of existence, no texts (other than a single text used as an example for submissions) are available from this web site.[2] Perhaps the primary reason for this failure is its pre-maturity: it predates Wikipedia—the first highly successful use of collaborative information—by approximately 5 years. Moreover, at that time, few households had internet connectivity and although researchers had this facility, it certainly was not the norm to perform research in this manner. In short, the world was not ready for this inspired innovation. There are other factors that have prevented this resource from being used. First, the unit of annotation chosen was the glyph (in contrast to the grapheme). This required the use of a cumbersome and difficult encoding

2 Due to lack of maintenance and recent updates, it is possible that some texts were previously available, though it is not likely that there were ever very many.

scheme. Furthermore, this level of transcription is based on the obsolete Thompson numbers rather than phonetic values and hence further annotations would need to be added in order to perform meaningful linguistic research. Finally, the lack of a searchable interface within texts is an unsatisfactory oversight that precludes serious use of the corpus as a tool for linguistic (or other) research.

Another commendable project is the Maya Hieroglyphic Database (MHD) (Macri, 2001). The database aims to be a comprehensive corpus of all known texts that includes line drawings, transcriptions, transliterations, and translations with additional metadata including date, site, and region. If the same information included in the catalog (Macri and Looper, 2003) is also directly available in the database, as is likely the case, then the database also includes related entries from multiple Yukatek and Chol sources and extensive bibliographic information. This is a very rich resource, and perhaps the first to contain phonetic transcriptions. Despite its enormous potential utility, the MHD suffers from several problems. Principally, in spite of its projected 2004 release on the internet, the database is not yet publicly available. In fact, the lack of updates to the web site for several years makes one wonder if the project will ever be released.[3] Even if released, however, this project is privately maintained and suffers from the problems enumerated above for such projects, not the least of which is the lack of ability to be updated by others. This is important because the database is based on the non-standard, unused cataloging system created by the authors. Finally, although it is impossible to know for sure without access to the actual database and the web interface, it doesn't appear that this database or its access software will fully allow for the type of searches established by corpus linguistics, which are essential to understanding the language of the hieroglyphs.

The final and most recent database is a sister project to the MHD known as the Maya Hieroglyphic Codices (MHC) (Vail and Hernández, 2005). This database only encompasses the codices (a very small portion of the overall corpus), and to date, only the Madrid codex is viewable and searchable online. Like the MHD, it includes transliterations, transcriptions, translations, and photographs. It also includes searchable metadata related to the iconography. Notwithstanding the richness of information contained in the database, it is not useful for serious linguistic inquiry. Although it is possible to search by glyph or lexeme, the search engine is fraught with the type of problems present in privately maintained databases. For instance, the search engine returns results solely in Yukatek Maya, despite the general consensus that the codices contain considerable amounts of Ch'olan (Wald, 2004). Even the transcriptions are outdated (e.g. ji is often transcribed as hi in the corpus, despite clear evidence that they are distinct; see Grube, 2004). Most importantly, using this interface, it is not possible to directly study other aspects of language, including

3 The principal investigator of the MHD did not respond to my email inquiry about the projected release date.

frequency and collocation. Indeed, linguistic research based on this system could be termed a modern day "pseudo-procedure" in comparison to the corpus linguistic based approach outlined previously.[4] And if this is any indication of the limitations of the MHD, the same can be said of it. Nevertheless, the MHC deserves due recognition as the first (and only) publicly available, searchable database that contains linguistic information.

The problem with current resources can be summarized as follows. First, there is no publicly available database containing even a significant fraction of transcriptions of known texts. Second, most of the databases are privately maintained and therefore biased and fraught with errors; the one corpus that offers decentralized control suffers from other problems, including an inappropriate unit of annotation. Third, although some corpora allow for basic searches that could be used to study some aspects of phraseology, none of them have facilities for performing serious linguistic inquiry as described earlier. Clearly there is a need for a corpus that meets the previously established criteria.

5. MayanWiki

MayanWiki seeks to meet all of the aforementioned criteria as a publicly available, wiki based, central corpus of hieroglyphic texts based on sound corpus linguistic design. At the heart of MayanWiki is a carefully designed database that allows data to be stored and retrieved in a manner conducive to linguistic research. MayanWiki also includes a flexible search engine to facilitate such research. Finally, the wiki frontend allows for decentralized control since content is user submitted and openly editable by anyone. These three aspects of MayanWiki are discussed in turn.

Creating a database is usually simple and straightforward. However, designing a database well requires more effort. The database employed by MayanWiki has been engineered to efficiently store and retrieve data in ways propitious to the study of language. It also has provisions for handling reconstructed and unreadable data which makes it particularly useful for ancient languages such as Mayan. Importantly, because the database focuses on storing linguistic data, the design can be incorporated into corpora of other languages as well. Moreover, the database was designed first at a conceptual level, and hence is not specific to any one type of database, e.g. a relational or XML database. It is not the purpose of this work to provide the details of the database; a more thorough exposition, including the conceptual, logical, and physical design, is found in Haertel (2007).

As always, the choice of whether to use a relational database or XML database depends on a project's particular needs and goals. MayanWiki implements the linguistic database as a relational database primarily because MediaWiki, the wiki

4 That is not to say that other valuable research is not possible. For instance, this database appears to provide a wealth of iconographic information that could be invaluable to iconographers.

software, is already implemented as a relational database and hence the two are more tightly integrated this way. Although relational databases and XML databases are essentially equal in capability, relational databases are typically able to process larger amounts of data more quickly. This allows for a very large amount of annotation to be added to a relational database without noticeable degradation in performance. Furthermore, a relational database can be exported as XML in situations where standardization is required, particularly when data is exchanged. For these reasons, a relational database best matches the goals of MayanWiki, but this may not be the case for all linguistic databases—even wiki based corpora.

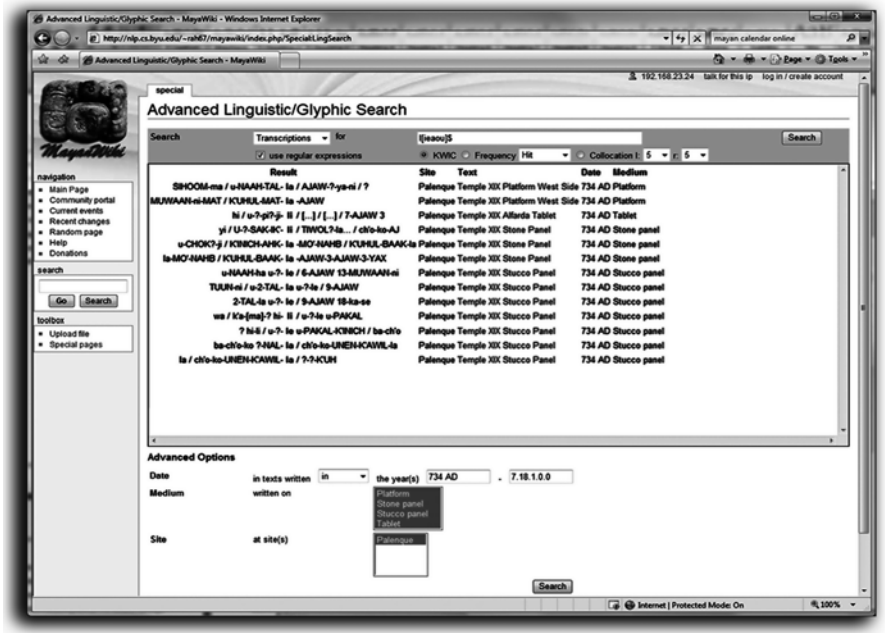

Figure 1: Example search results in MayanWiki (searching for "l[ieaou]$" at `http://nlp.cs.byu.edu/~rah67/mayawiki/index.php/Special:LingSearch`).

MayanWiki's search engine leverages the design of the database to allow for powerful and flexible linguistic searches (see Figure 1 for an example). The basic mechanism for studying phraseology in MayanWiki is the use of concordance lines that show particular search terms (strings of graphemes or morphemes) surrounded by their context. MayanWiki also allows search terms to be studied by comparison of frequency in the entire corpus or across sites, media, or date ranges. The search engine also has a facility whereby a list of graphemes or morphemes that occur within a specified window of the search term can be displayed in order to facilitate the study of collocations. For all types of searches, it is possible to use regular

expressions and limit results to specific sites, date ranges, or media. The results can also be sorted by any combination of columns (meta-data) returned within the results (i.e. the search term(s), site, date, media, etc.). This flexibility will allow for data to be manipulated in ways that will facilitate new understanding of the language of the hieroglyphs.

The relational database is the heart of MayanWiki, but the wiki is the means through which data is viewed, entered, and modified. The choice to use a wiki as the medium for this resource is advantageous in several ways:

- Data are user submitted. One of the major hindrances to achieving the goal of a central repository of all glyphic data is that it is not feasible for a single person, or even several, to transcribe, transliterate, and translate the entire corpus. If this task is instead left to the larger group of Mayanists, the task is much more feasible. A wiki format makes this plausible.
- Consensus based. Scientific progress only happens with consensus. Typically, proposals regarding decipherments, spelling rules, syntactic elements, etc. are made based on available evidence. The acceptance or rejection of such proposals ultimately depends on the consensus within the community. A wiki is explicitly based on this same principle, namely, that over time, the interpretations based on user submitted data will converge based on consensus; conflicting viewpoints are resolved over time.
- Modifiable. A wiki is designed to allow anyone to contribute (although controls are available to avoid vandalism). When anyone can contribute, more data are made available, and existing data are readily correctable. Existing texts are easily updatable to reflect new or amended decipherments, spellings, etc. Finally, adding new data as it becomes available through new archaeological finds or other means is straightforward.
- Public discussion. Some wikis, such as the one employed in MayanWiki, include the ability to discuss every page (i.e. text, image, or other information). This is important because new ideas or disagreements can be discussed publicly and permanently where all can participate and view the discussion.
- Private pages. Sometimes, consensus takes a very long time. Other times, certain proposals may not be mainstream. In either case, it is possible for users to propose new readings in their own private space that does not conflict with the generally accepted transcriptions and transliterations.
- Change tracking. A history of every change ever made to a text is recorded by the wiki. This makes it easy to undo accidental or malignant changes. Additionally, it provides an automatic history of the progress of the field.
- Watch lists. The wiki implemented in this project includes a watch list. Subscribed users are notified of every change. This not only checks vandalism, but also allows users to receive the latest updates to progress in the field.

- Flexible copyrights. A wiki can allow for flexible licensing, most notably, a Creative Commons license, which typically allows free use when proper attribution to the author is given. This protection should encourage researchers to submit their drawings and photographs, while still retaining the benefits of being freely available.

In short, the wiki media allows central access to texts, while control is decentralized, as discussed earlier.

Clearly, MayanWiki's powerful search, highly engineered linguistic database and wiki frontend allow MayanWiki to meet all of the criteria of a successful corpus of hieroglyphic transcriptions useful for linguistic research.

6. Feasibility of a Wiki

The use of a wiki to create and maintain a corpus is novel, but not without challenges. First, many specialists are uncomfortable with the idea that their work can be edited by anyone. Also, the underlying assumption is that a wiki will allow the data in the corpus to converge to the truth, but this need not be the case. This section discusses these topics.

The idea that anyone, including students, hobbyists, and non-specialists, can modify the texts contained in the database may at first seem to be a major disadvantage to the use of a wiki. This has been used as criticism against the highly successful Wikipedia. However, research has shown that by-and-large (though not without exception), the content on Wikipedia is surprisingly accurate (Giles, 2005; Rosenzweig, 2006) and devoid of vandalism (Viegas et al., 2004). Reasons for this include Wikipedia's insistence on neutrality, the use of talk pages for "meta-discussion" about articles, the fact that it is easier to undo vandalism than to vandalize, and the existence of watch lists that allow for almost immediate removal of vandalism (see Lih, 2004; Viegas et al., 2004). These same principles apply to MayanWiki, since it employs Wikipedia's software. Furthermore, wikis can be configured to only allow registered users to modify data. Thus, in the case that some users become a hindrance to research, such a policy can be enforced. However, this creates the burden of deciding who to grant editing privileges to; it also has the potential to turn the wiki into a privately maintained database. Hence, care should be taken when using this option and it should only be used when necessary.

The fundamental premise of MayanWiki (and wikis in general) is that the data will converge to consensus if used by enough people with differing views and enough expertise over a long enough period of time; the more people there are that are actively involved in editing content on a wiki, the less time it takes to reach consensus. One reason for the success of Wikipedia is that it is used by a very large number of people with a large spectrum of varying views. Obviously, the number of people that are capable of transcribing Mayan hieroglyphic texts is much lower than those that use Wikipedia. Nevertheless, compared to many other (non-Classic) ancient

languages, there is a large number (perhaps one to two dozen) of researchers whose principal field of study relates to the Mayan hieroglyphs and a similar number of students entering the field. Furthermore, there is a large number (on the order of several hundred) of hobbyists who are also interested in the Mayan hieroglyphs, although their proficiency with the glyphs is at varying levels. Nevertheless, there are enough active researchers, students, and aficionados and few enough texts for consensus to be reached in a relatively short period of time.

However, even though there are a sufficient number of people involved in the field, this does not guarantee the success of the wiki. There are two strategies that

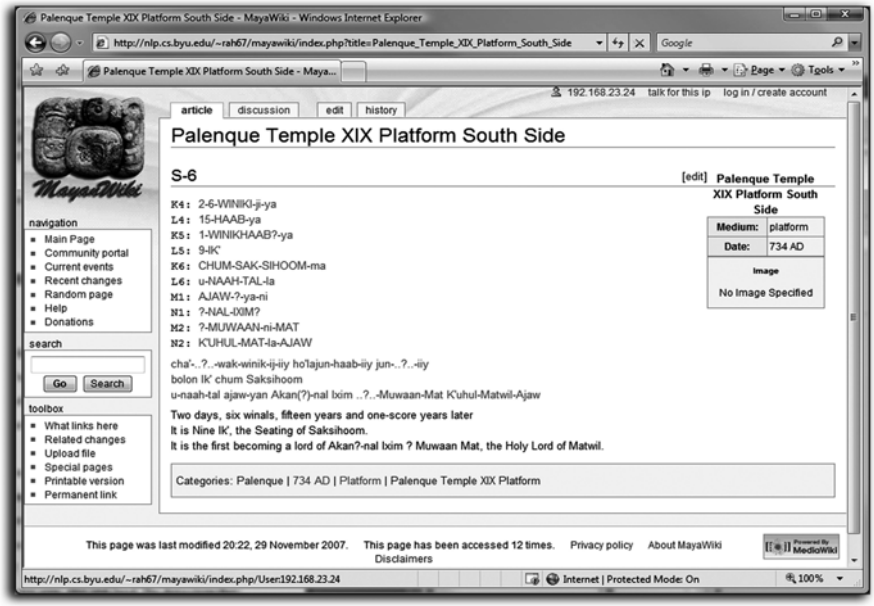

Figure 2: Example transcription, transliteration, and translation taken from Stuart (2005) `http://nlp.cs.byu.edu/~rah67/mayawiki/index.php/Palenque_Temple_XIX_Platform_South_Side`.

can further improve the chances of MayanWiki's success. First, the database must be populated as quickly and extensively as possible. Second, a policy of "conservative transcriptions, innovative explanations" must be firmly established. Each of these strategies is discussed in turn.

The major weakness of MayanWiki in its current state as a prototype, is that it presently contains only a handful of texts. Most serious researchers will likely ignore MayanWiki until it contains a respectable amount of data. However, because users are allowed—even encouraged—to submit texts (see Figure 2 for an example submission), MayanWiki is fully capable of becoming a comprehensive resource

in relatively little time compared to the enormous amount of time required for a single researcher to populate a database by himself or herself. If even only a handful of knowledgeable students or researchers were to continuously contribute data, much progress will still be made. One strategy for getting data—albeit slightly less accurate than desirable—is to encourage students to add content as part of the learning experience in an introductory course to hieroglyphic writing. Hobbyists could also contribute texts as they practice reading and transcribing. This process of populating the database is the first key to the success of MayanWiki.

Epigraphers, especially seasoned ones, can be skeptical of the work produced by other schools of thought, even when these opinions affect a small percentage of the data. Thus, if neutrality in transcriptions is not maintained, many researchers will choose not to use MayanWiki and it will effectively be reduced to a privately maintained database. This problem can be avoided if users are encouraged to transcribe texts as conservatively as possible—that is, based on accepted, published decipherments, etc. This way, the data will be perceived as being less problematic. This squares with Wikipedia's principle of absolute neutrality. Nevertheless, differences of opinion are inevitable, and in fact, such differences are ultimately the source of new discovery. Researchers are thus strongly encouraged to propose innovative ideas and alternative readings. However, this should be done outside the context of the more conservative, generally accepted data that is analyzed by the search engine. More to the point, users should be encouraged to propose innovations in a convincing, clear manner with supporting data (which can conveniently be obtained from MayanWiki itself) on discussion pages or on their own user pages (also a part of MayanWiki). Researchers should further be encouraged to offer additional supporting evidence or counter-evidence in order that all theories get fair treatment from all parties. Such discussion will eventually materialize into published articles and the accepted theories will make their way into the data themselves. Indeed, this policy of "conservative transcriptions, innovative explanations" is essential not only to the success of MayanWiki, but to the progress of the field. Like Wikipedia's emphasis on neutrality, this principle should be encouraged and enforced by the main contributors to the project.[5] This can be accomplished primarily through feedback on discussion pages and reverting changes deemed non-conservative.

7. Conclusion

The purpose of this work has been to introduce MayanWiki as a corpus of transcriptions of Mayan hieroglyphic texts specifically intended for linguistic inquiry. First, this paper established basic criteria for such a corpus, namely, that it must be

5 The main contributors to the project are those who actively submit new content or edit old content—they are not appointed or elected. It is assumed that these contributors will be intimately familiar with the policies set forth by MayanWiki and probably the top researchers in the field.

publicly available, in electronic format, and centrally accessible, but with decentralized control. It must also minimally allow for analysis of phraseology, frequency, and collocations. It was then shown that existing corpora/databases fail in most of these categories. Next, MayanWiki was presented as a viable solution that meets all of the criteria. The specially engineered linguistic database coupled with the powerful search engine allow the corpus to be amenable to linguistic research. The wiki frontend allows data to be submitted and enhanced by anyone. Finally, it was shown that the wiki approach is feasible in the field of Mayan linguistic epigraphy. Key to this is quickly populating the database, and students and aficionados could make this possible. Also, a policy of "conservative transcriptions, innovative explanations" must be maintained if researchers are to take MayanWiki seriously.

To conclude, the applicability of the ideas presented herein to other languages is considered. As previously mentioned, the database itself is applicable more broadly than to just the Mayan hieroglyphs, and the database is easily adapted for finer-grained annotations or other languages (see Haertel, 2007). However, the decision to use a wiki depends on several factors. One advantage to the wiki format is that it is simple to add new texts. In cases where new texts are constantly being discovered, a wiki may be appropriate. Another consideration is the level of disagreement about the data contained in the corpus. This can itself be affected by the chosen level of annotation: (sometimes, more fine-grained annotations are more subjective and hence have more disagreement, although the opposite case is often true as well). For corpora in which the data is highly disputed, a wiki tends to be more appropriate. Of course, individual researchers might feel like their voice cannot fairly be heard in a wiki. However, if their ideas are indeed correct, they will eventually be accepted by the community at large, especially if the ideas are first published in reputable journals and other venues. A wiki could help resolve many of the disagreements and gradually bring the field to a consensus on disputed matters.

Another consideration that was previously mentioned is the number of people that are actively performing research in the area for which the corpus will be used. If the project is mainly being used by one researcher or a small group of researchers, a wiki may not be as appropriate since in that case the corpus is essentially a private database. However, a group of researchers may choose to use the wiki format to facilitate easy modification of the corpus and simply only allow registered users (i.e. members of the group) to perform edits. On the other hand, if there are a large number of researchers, students, and/or hobbyists, a wiki has the potential to be seen by enough people often enough to converge to correct readings relatively quickly. In this case, a wiki is desirable.

Since linguistic interpretation of glyphic data is highly disputed in the field of Mayan linguistic epigraphy and there are a large number of researchers, students, and hobbyists, a wiki format is appropriate, even advantageous, to MayanWiki. Indeed, the wiki combined with the linguistic centered database will allow MayanWiki to

become not only a useful tool for the study of the language of the hieroglyphs, but a prototype for other wiki based corpora as well.

References

Abercrombie, David. 1965. *Studies in Phonetics and Linguistics.* London: Oxford University Press.

Alvarado, Rafael C.. 1994. The Mayan Epigraphic Database Project. http://www3.iath.virginia.edu/med/ (1.8.2007).

Giles, Jim. 2005. Internet Encyclopaedias Go Head to Head. *Nature* 438 (7070): 900-901.

Grube, Nikolai. 2004. The Orthographic Distinction between Velar and Glottal Spirants in Maya Hieroglyphic Writing. In Søren Wichmann, *The Linguistics of Maya Writing*. Salt Lake City: University of Utah Press, pp. 61-80.

Haertel, Robbie. 2007. *MayanWiki: An Online, Consensus-based Linguistic Corpus of the Mayan Hieroglyphs.* Unpublished M.A. thesis, Department of Linguistics and English Language, Brigham Young University, Provo, UT.

Hunston, Susan. 2002. *Corpora in Applied Linguistics*. Cambridge, UK: Cambridge University Press.

Lih, Andrew. 2004. Wikipedia as Participatory Journalism: Reliable Sources? Metrics for Evaluating Collaborative Media as a News Resource. In: *Proceedings of the 5th International Symposium on Online Journalism*, University of Texas at Austin. Retrieved 16-17.4.2004). http://online.journalism.utexas.edu/2004/papers/Edwards.pdf (1.8.2007).

Macri, Martha J.. 2001. Maya Hieroglyphic Database Project. http://nas.ucdavis.edu/NALC/mhdhome.html (1.8.2007).

Macri, Martha J. and Matthew G. Looper. 2003. *The New Catalog of Maya Hieroglyphs, Volume One: The Classic Period Inscriptions, volume 1.* Norman (OK): University of Oklahoma Press.

Rosenzweig, Roy. 2006. Can History be Open Source? Wikipedia and the Future of the Past. *The Journal of American History* 93: 117-146.

Stuart, David. 2005. *The Inscriptions from Temple XIX at Palenque*. San Francisco: The Pre-Columbian Art Research Institute.

Vail, Gabrielle and Christine Hernández. 2005. The Maya Hieroglyphic Codices, Version 2.0. http://www.mayacodices.org (1.8.2007).

Viegas, Fernanda B., Martin Wattenberg, and Kushal Dave. 2004. Studying Cooperation and Conflict between Authors with History Flow Visualizations. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 575-582.

Wald, Robert F.. 2004. Languages of the Dresden Codex. In Søren Wichmann, *The Linguistics of Maya Writing*. Salt Lake City: University of Utah Press, Salt Lake City, pp. 27-58.