2007-12-05

MayanWiki: An Online, Consensus-Based Linguistic Corpus of the Mayan Hieroglyphs

Robbie A. Haertel
Brigham Young University - Provo

MAYANWIKI: AN ONLINE, CONSENSUS-BASED LINGUISTIC CORPUS OF THE MAYAN HIEROGLYPHS

by

Robbie A. Haertel

A special project submitted to the faculty of

Brigham Young University

in partial fulfillment of the requirements for the degree of

Master of Arts

Department of Linguistics and English Language

Brigham Young University

December 2007

BRIGHAM YOUNG UNIVERSITY

GRADUATE COMMITTEE APPROVAL

of a special project submitted by

Robbie A. Haertel

This special project has been read by each member of the following graduate committee and by majority vote has been found to be satisfactory.

Date	John S. Robertson, Chair
Date	Mark Davies
Date	Deryle Lonsdale

BRIGHAM YOUNG UNIVERSITY

As chair of the candidate's graduate committee, I have read the special project of Robbie A. Haertel in its final form and have found that (1) its format, citations, and bibliographical style are consistent and acceptable and fulfill university and department style requirements; (2) its illustrative materials including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the graduate committee and is ready for submission to the university library.

Date

John S. Robertson
Chair, Graduate Committee

Accepted for the Department

William G. Eggington
Chair, Department of Linguistics and English Language

Accepted for the College

Gregory D. Clark
Associate Dean, College of Humanities

ABSTRACT

MAYANWIKI: AN ONLINE, CONSENSUS-BASED LINGUISTIC CORPUS OF THE MAYAN HIEROGLYPHS

Robbie A. Haertel

Department of Linguistics and English Language

Master of Arts

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. One of the most important resources for linguistic study of this type is a comprehensive, electronic corpus of texts to investigate phraseology, frequency information, and collocations. Because Mayan linguistic epigraphy is in the early stages, a publicly available, editable corpus would be an invaluable resource in arriving at consensual readings.

Unfortunately, no such corpus currently exists. The purpose of this project is to present MayanWiki as a relational database of hieroglyphic transcriptions and transliterations with a wiki frontend that includes advanced search functionality that meets the aforementioned criteria. The principle behind the wiki is to accelerate the convergence of readings to the "truth". Once the database is fully populated by users, it will become a valuable tool allowing them to manipulate data in ways that will facilitate scientific discovery of new and interesting linguistic patterns.

ACKNOWLEDGMENTS

This project was funded in part by a BYU graduate research fellowship; without this funding, MayanWiki would probably not exist. I am very grateful to all of those at the department and university levels that selected my project over so many other very worthy proposals.

I would also like to thank my committee, Drs. John Robertson, Mark Davies, and Deryle Lonsdale, who so kindly and patiently helped guide and shape the project along the way. They were a constant source of encouragement and an invaluable source of information. I am also very thankful for their willingness to read drafts so quickly in order to accommodate an expedited schedule.

I am especially indebted to Dr. Robertson. I am still amazed that he was willing to convert a computer programmer into a linguist in order that I might participate in his research on Colonial Ch'olti'. This research assistantship provided so much more than monetary assistance: it was truly a once-in-a-lifetime opportunity to study the Mayan languages, do linguistic fieldwork in Jocotán, Guatemala, and translate the Colonial Ch'olti' liturgical documents. My participation in this project led me to where I am today. It was a joy to learn from such a brilliant mentor, and the wisdom he shared with us over the years will ever impact me. I am especially thankful for his patience with me, despite

my stubbornness; he always recognized my potential, even when my performance would indicate otherwise.

Dr. Stephen Houston also deserves special recognition. Although he was unable to continue as a committee member after his move to Brown University, it was he who taught me how to read the hieroglyphs and instilled in me a love for this beautiful writing system. He provided the initial encouragement that made me believe that this project could benefit the community. I also thank him for his advice in times of need.

I would also like to thank my friend and colleague, Danny Law. I will never forget the time spent in Dr. Robertson's office reading the Ch'olti' Arte, learning about the laws of analogy and historical linguistics, and translating the liturgy. His companionship in Guatemala helped alleviate the loneliness I sometimes felt as I thought about my pregnant wife back home. As the unofficial fourth committee member, his consultations provided valuable answers to questions. He is a true friend. He always learned so much faster than I, and he will undoubtedly be a very successful Mayanist.

My current Ph.D. advisor, Dr. Eric Ringger, deserves recognition as well. I am thankful that he, too, recognizes my potential rather than my weaknesses. He has also provided direct support for the project by allowing me to finish my special project this past summer at the expense of other productive research more in line with the goals of his lab. He is a great example to me of patience. I appreciate all that he has taught me during my first year in the program and I am looking forward to the upcoming years under his tutelage.

I am also thankful for those teachers and mentors who were patient while teaching linguistic principles to a computer scientist. In particular I am grateful for the

opportunities afforded me by Drs. Ray Graham and Deryle Lonsdale. Their support was instrumental in bringing me to where I am today.

My parents deserve recognition for their love and support. It would be impossible to enumerate the ways in which they have helped; would that everyone had parents as loving as mine.

I am also eternally indebted to my beautiful wife for her continual love and support. Meri lifted me when I was down and encouraged me to continue forward when things were difficult. She has done more than her fair share of household responsibilities and taking care of the boys and I am ever thankful for her support and patience as I pursue my dreams. I love her dearly!

Finally, I am thankful for my two sons, Alex and Jared. The distractions they provided (including their births) to my “tésis” were unforgettable. They have brought me more joy than I ever imagined. Never was a father more proud of his amazing boys!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 Introduction

Mystery spawns intrigue. Perhaps this is why the mysterious Maya civilization, with its grandiose cities and its once-cryptic writing system, has attracted the interest of so many and captivated great minds from even before the time that Stephens and Catherwood popularized the remarkable ruins the Mayans left behind. It was Stephens himself who, despite much criticism, believed that much of the mystery surrounding the Maya would disappear if the inscriptions would be deciphered (Coe, 1999). This belief inspired his challenge regarding the glyphs: “No Champollion has yet brought to them the energies of his inquiring mind. Who shall read them?” (Stephens, 1841, p. 160). That “Champollion” would not come for another hundred years when Yuri V. Knorosov discovered the true nature of the writing system. This Russian scholar was the first to recognize that the Mayan writing system consists of both logograms and syllabic symbols—much like the Japanese kanji and kana, respectively. In spite of fierce resistance from the influential Sir Eric Thompson that hindered the immediate acceptance of Knorosov’s convincing discovery, scholars now unanimously accept the true nature of the script as proposed by Knorosov.

Since that time, a high percentage of the glyphs has been deciphered. Nevertheless, many questions of grammar and spelling remain unanswered. In this regard, this chapter establishes four main points: First, the current focus of Mayan

epigraphy has shifted from decipherment to a thorough study of the language of the glyphs. Second, the currently available resources, while sufficient for decipherment, are helpful but insufficient for the type of linguistic study necessary for further progress in understanding the language of the glyphs. Next, criteria for a computerized database of transcriptions and transliterations are established. Finally, MayanWiki is presented as a resource that meets these criteria.

1.1 Current Direction of Mayan Epigraphic Linguistics

Since Knorosov's time, many important advances in the decipherment have been made. Decipherment of the glyphs proceeded at unprecedented rates between 1975 and 1995. David Stuart's article, "Ten Phonetic Syllables" (1987) played a very important role during this time, not only because of the important new decipherments proposed therein, but because of the methodology it established. During these years, the number of glyphs that were known jumped from several dozen to several hundred (Stuart, 2005a).

With the decipherment of such a large percentage of the known glyphs, it is natural to ask if all the real work has been done. The fact remains that there are still a number of glyphs whose phonetic or logographic values continue to elude epigraphers and this will probably always be the case, especially with the discovery of new sites and texts. However, even with these undeciphered glyphs, the majority of the corpus is readable, and understanding the language in which the glyphs are encoded has become a priority. Long ago Knorosov said, "As a result of decipherment, the study of texts becomes a branch of philology" (Knorosov, 1958). More recently, Wichmann (2004) adds, "If the mid-eighties represented the great boom in the phonetic decipherment of the

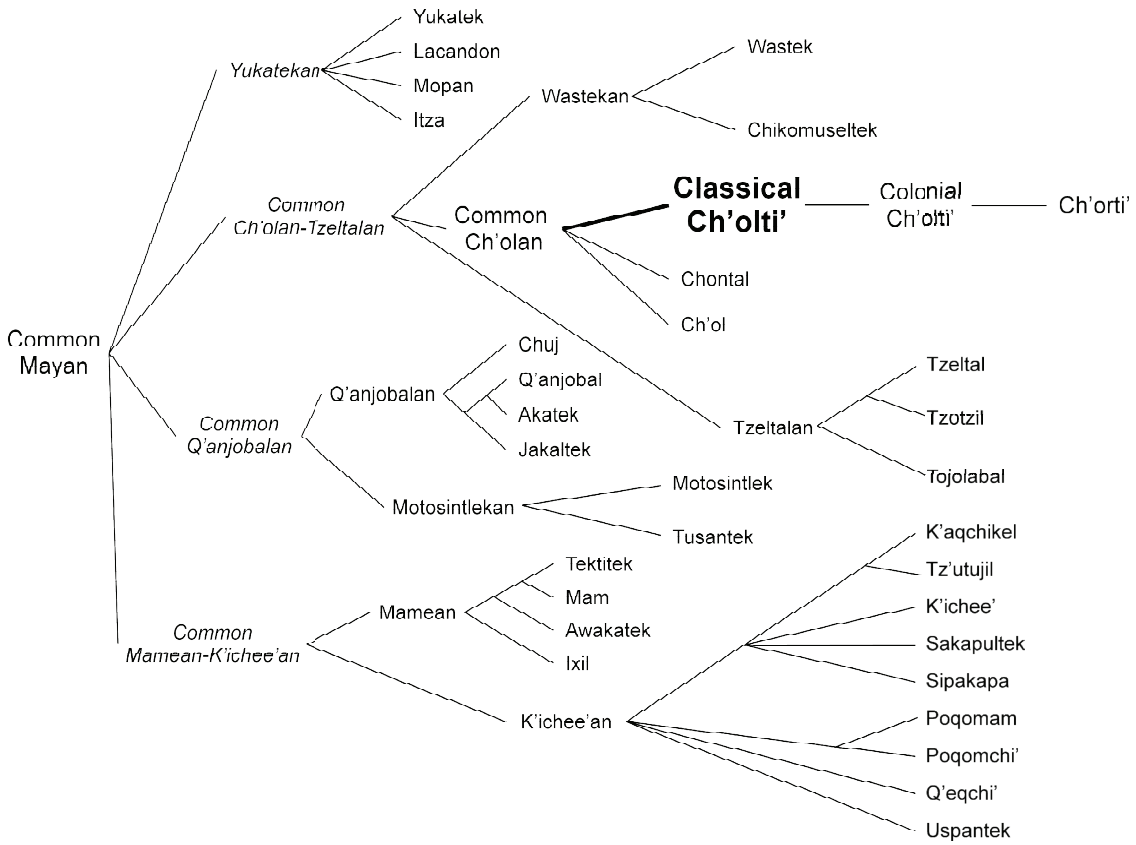


Figure 1-1 Mayan language families (after Law, 2006).

Maya script, with the high points being the publication of Justeson and Campbell (1984) and Stuart (1987), the late nineties and the turn of the century may be characterized as the culmination of its linguistic interpretation" (p. 1). Finally, Stuart (2005a) claims, "It is clear that learning Maya hieroglyphs and the language they recorded will become as essential a part of academic training in Maya studies as learning to read Latin is for historians of ancient Rome" (p. 5a). Clearly, any work in the field *must* be centered on linguistic study.

Many strides have already been made in this direction. Perhaps the most important start has been to investigate the language in which the hieroglyphs were written (see Figure 1-1 for a tree of the Mayan language family). There is very good evidence—

especially the presence of the verbal suffix –*wan* (e.g. MacLeod, 1984; MacLeod, 1987; Mathews & Justeson, 1984; Ringle, 1985)—that the script is based on a Ch'olan language. Houston, Stuart, and Robertson (2000) present additional linguistic evidence that the hieroglyphs of the Classic period represented a standardized, priestly language that they denominated Classical Ch'olti' as the direct ancestor of Colonial Ch'olti' and modern-day Ch'orti'; further support was added by Robertson, Houston, Law, and Haertel (in press). Another important step was the publishing of *The Linguistics of Maya Writing* (Wichmann, 2004) which presents initial research from many scholars on the language, phonology, and grammar of the hieroglyphs.

Despite these important advances, there is still much work to be done in the study of the linguistics of the hieroglyphs. For instance, Wald's (2007) recent dissertation proposes that the language of the hieroglyphs was Classic Ch'olan—a direct rejection of the Classical Ch'olti' proposal, despite continued evidence presented to the contrary (e.g. Robertson, Houston, & Law, in press). Moreover, many of the papers presented in Wichmann's book (2004) are proposals that are still being debated. In fact, within the volume itself, several papers present alternate views, e.g. regarding the status of tense and aspect (Robertson, Houston, & Stuart, 2004; Wald, 2004) and vowel disharmony (Houston, Stuart, & Robertson, 2004; Lacadena & Wichmann, 2004). Needless to say, much research remains in all areas of linguistics: phonology, morphology, syntax, discourse, etc.

1.2 Insufficiency of Currently Available Data

The object of the type of linguistic research currently being undertaken is to uncover the principles that govern all levels of the spoken and written forms of the hieroglyphic script through a systematic study of the available data. The most important data are obviously the physical glyphs themselves: any theory or hypothesis must ultimately be tested against the glyphs. However, the form of data most favorable to thorough linguistic study is the linguistic data derived from the glyphs, usually in the form of the so-called transliterations (see below), although transcriptions also contain important and useful information. Unfortunately, very little effort has been made thus far to publish transcriptions or transliterations of texts in their entirety, which may be indicative of the fact that the focus until recently has been on decipherment. One notable exception is Stuart's remarkable book (2005b) on Temple XIX at Palenque, which contains a transcription of all the texts discussed, and a transliteration and translation for the south and west faces of the platform. This is certainly a step in the right direction, but linguistic data are needed for more than just two texts. In short, there is a need for a corpus of transcribed and transliterated texts, without which linguistic research will be hindered.

1.2.1 Current Resources

That is not to say that other valuable resources do not exist. Indeed, progress continues to be made in the field due in large part to existing resources. These resources include photographs, line drawings, catalogs, syllabaries, lexicons, and dictionaries. Each of these resources is a level of abstraction from the physical inscriptions that successively move

towards linguistic interpretation. Each succeeding level of abstraction is increasingly more accessible and readily processed, by both human and machine.

The actual raw data are the physical glyphs found throughout the jungles of Central America or in museums and other collections. Photographs, though removed in time and space from the physical artifacts, preserve much of the same detail as the physical glyphs but are much more accessible. Furthermore, photographs essentially preserve the physical artifacts which are subject to erosion, looting, and other forms of destruction. Two of the more significant collections of photographs were painstakingly produced in early years by Maudslay (1889-1902) and Maler (1901); more recently Kerr and Kerr (1989-2001) have produced an important corpus of photographed vases using their rollout technique.

Like photographs, line drawings are also removed temporally and spatially from the physical data, but in addition, they abstract away unimportant physical detail such as surface erosion and depth in order to highlight the distinguishing features of each particular glyph instance (that is, the outline). This makes line drawings much easier to interpret than photographs and consequently is currently one of the most widely used and valuable resources. For this reason, most epigraphers maintain a large private collection of line drawings in part from their own work, but also reproductions of other's drawings as well. There are also a few publicly available collections of line drawings, both in print and electronic form (e.g. Graham, 1975-2006; Schele, 1998; Montgomery, 2000).

While no two instances of glyph "tokens" are exactly the same, certain glyphs are intended to represent the same abstract entity, which are termed *graphemes*. At this level of abstraction, minor details (such as inter- and intra-scribe variation and mode of

inscription) particular to each instance are ignored. Catalogs (e.g. Macri & Looper, 2003; Thompson, 1962) are an attempt to list all known graphemes. More importantly, each grapheme can be assigned a phonetic value; graphemes corresponding to a syllable are termed *syllabograms* while graphemes representing word roots are termed *logograms*. Each syllabogram or logogram may itself be one of many *allographs* for a particular syllable or logogram (also ambiguously called a *glyph*). For instance, the **u** syllabogram has a "fish-head" variant and a skull variant, among many others. In this case, **u** is the syllable and these variants are two of its allographs. The purpose of syllabaries and lexicons (e.g. Coe & Van Stone, 2005) is to list the various allographs of all graphemes for each syllable and word root based on their phonetic value. Dictionaries (e.g. Montgomery, 2002; Mathews & Bíró, 2006) are an extension to syllabaries and lexicons whose focus is on units of meaning and hence include strings of syllables and logograms.

Finally, the syllables and roots represented by actual occurrences of syllabograms and logograms in texts (i.e. the glyphic tokens) are combined to produce the words and morphemes represented by the hieroglyphic script. For convenience, this process is usually done in two parts. First, each glyphic token is romanized using its phonetic value. Then, based on a set of invertible spelling rules[1], this transcription is transliterated into standardized Mayan morphemes and words (using a phonemic alphabet consisting of Roman characters)[2]. If a Mayan scribe from Classical times were to learn to read this Romanized, alphabetic representation of his language, the transliteration should

---

[1] Spelling rules dictate how glyphs are to be used to represent the sounds of a language; these rules are "inverted" (whenever possible) to convert from a glyphic representation to a phonetic one.

[2] The reason this step is necessary is that the script is not entirely phonemic, as will be further explained in Chapter 2.

correspond to how he would read the hieroglyphs aloud. In addition to the transcriptions and transliterations, translations are often provided that mostly preserve the same meaning as the original Classical Ch'olti' text, but in English or any other language.

It is important to note that, although each of these levels is increasingly subject to human interpretation and thus more open to errors, the interpretive process is remarkably consistent and relatively few errors are actually committed. Indeed, our ability to understand glyphic texts depends on this interpretive process of transforming incisions in stone into phonetic representations and meaning in the mind.

1.2.2 Lack of Transcriptions, Transliterations, and Translations

Even though the glyphic data are available in these various forms, the principal forms pertinent to the study of the language of the hieroglyphs—the current focus in the field—are first the transliterations and second the transcriptions. Granted, when questionable, the transcriptions and transliterations need to be verified against line drawings, photographs, and dictionaries. Nevertheless, most work can be done using the transcriptions and transliterations alone. Unfortunately, as was previously mentioned, transcriptions, transliterations, and translations do not generally exist for texts—at least not publicly. Stuart (2005b) points out that one weakness of the field is that resources like these tend not to be made publicly available:

> Another motivation behind the precise treatment of the glyphs is to help do away with a small portion of the "grey literature" of unpublished readings and ideas that circulate among epigraphers, mainly by impermanent emails. (p. 15).

This is his motivation for including the transcriptions, transliterations, and translations in his book.

Although transcriptions, transliterations, and translations for full texts are scarce, authors are increasingly including transcriptions and even transliterations and translations of segments of texts in their work—a reflection of the change in emphasis in the field and level of decipherment. This practice is exemplified in the sourcebooks for the Maya Hieroglyph Forum (Schele, 1978-1988; Stuart, 2005-2007; Wanyerka, 1989-2004), but also found throughout the published literature (e.g. Mora-Marín, 2004). While these are certainly beneficial, a publicly available collection of all transcription would be more useful still.

1.3 Criteria for a Useful Corpus

Although there is an obvious need for a corpus of transcriptions and transliterations, not just any corpus will suffice. The first criterion and ultimate goal is that the corpus should be comprehensive and in electronic form. Furthermore, it is also necessary that the entire corpus be accessible from a single central location. Yet, a central corpus often introduces additional problems if privately owned, 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 don't 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 allow for meaningful study through the application of corpus linguistic principles. The latter three criteria are explained in further detail in the following sections.

1.3.1 Central Access

Not only is there a paucity of available linguistic data from the glyphs, but what little exists is 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 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. 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). It is true that a few accomplished epigraphers have committed the entire corpus to memory and others are familiar with a large portion of it. While this certainly speeds up the procedure, it is still possible to inadvertently overlook important data and certain information is not easily processed by the human mind.

It is important to note that making resources available electronically is not enough. If electronic resources are scattered across multiple web sites, or even fragmented within a single web site through multiple search engines or poor search facilities, the result is still a pseudo-procedure. For effective research the corpus must be available from a single central location, with a single, useful search engine.

1.3.2 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 (which I will refer to hereafter 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 will probably always be (and there currently are) 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 the database, which would undermine the purpose and existence of a central resource. Under these circumstances, little progress is made.

Even if we 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.

The final potential problem with a privately maintained database is that people would probably only be willing to submit 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 single place (i.e. program or web page), rather than requiring that users collect (linguistic) data across multiple databases or sites. However, anybody—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.

1.3.3 Principles of Corpus Linguistics

Surely, any corpus that is to be useful should allow the corpus to be searched 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 in a database is useless if the access software does not provide the ability to extract 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.

Although without access to a computerized corpus, Knorosov exemplified this process of using an appropriate database and good search methods for his remarkable breakthrough. Armed with the Dresden, Madrid, and Paris codices along with Bishop Diego de Landa's "alphabet"—all of which had been previously studied by others—and influenced by his unique background in Egyptology, Japanese literature, Arabic, and the Chinese and ancient Indian writing systems, Knorosov realized that Landa's "alphabet" was actually a syllabary, leading to the fundamental discovery that the hieroglyphs consist of syllabic symbols and logographs (Coe, 1999). Seeing the same data freely available to others, but in a new light, has allowed for the level of decipherment we now enjoy.

Within the context of linguistic corpora, and particularly computerized data, the three principal ways in which a corpus is re-ordered and manipulated is through the study of frequency, phraseology, and collocation. 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, monumental versus vessel inscriptions or early versus 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.

Since all study of the Mayan hieroglyphs implicitly includes some degree of corpus-based study—even in comparative-historical reconstructions—it is important to recognize the limitations of corpora (after Hunston, 2002):

- Corpora cannot identify what is possible or not in a language, simply what is frequent or not. In other words, a corpus alone is not sufficient to determine the grammaticality of phrases, but, for example, it can help identify the default (or most frequent) word order (at least in the priestly language).
- A corpus cannot show more than its contents. This is important since the type of language used on monuments and vases is restricted. In Hunston's (2002) words, "conclusions about language drawn from a corpus have to be treated as deductions, not as facts." (p. 23).
- A corpus can provide evidence of phenomena but this evidence must be interpreted by a human. The corpus allows the data to be analyzed, but

ultimately, interpretation and intuition from a creative and resourceful human mind are required.

- Corpora present language out of its context. Unlike English texts, many hieroglyphic texts also contain accompanying drawings and iconography that are not present in a textual transcription. This is why it is important that the texts be linked to a photograph or line drawings whenever possible. However, also note that intonation, kinesics, and other paralinguistic information cannot be learned from this corpus.

So long as these limitations are acknowledged, a corpus of hieroglyphic texts that allows for the type of study typical of corpus linguistics could do as much for the field of Mayan linguistic epigraphy as Stuart's (1987) "Ten Phonetic Syllables" did for decipherment.

1.4 MayanWiki

To summarize, the focus of Maya hieroglyphic studies has largely shifted away from decipherment to a study of the language itself. Such a study demands access to a central store containing the entire corpus of transcribed and transliterated texts. However, despite the fact that access to the database is central, ownership of the content should be decentralized such that anybody can be allowed to add, modify, and otherwise contribute to the database. Finally, the database should be based on corpus linguistic theory, allowing for a study of frequency, phraseology, and collocation. Unfortunately, no resource is currently available that meet these criteria. The purpose of this project is to introduce MayanWiki as a wiki-based, central corpus of hieroglyphic texts based on state-of-the-art corpus linguistic design that is openly editable by anyone. The goal is to

make data more accessible and manipulable in order to foster collaboration and encourage advances in the field, while still being flexible and adaptable.

MayanWiki is first and foremost a database; the heart of MayanWiki is a relational database that has been carefully designed and engineered to be able to handle glyphic data in the transcriptions and linguistic data from the transliterations. Surprisingly, custom relational databases are not typically used to store linguistic data for use in corpus linguistics, with the exception of the work done by (Davies, 2005; Davies, in press) and Christ (1994; Christ & Schulze, 1995). To my knowledge, this is the first time a custom relational database has been used to store linguistic data for an agglutinative, polysynthetic language (see further discussion in Chapter 3 for the typology of Classical Ch'olti') with the intent of corpus linguistic study. The schema presented herein can be adapted with little change for use with similar languages. It is important to note that the database schema for MayanWiki can exist entirely independently of the wiki frontend and constitute the single most important contribution of this project. The design of the database schema is presented in Chapter 6.

If the relational database is the heart of MayanWiki, then the wiki frontend is the face and senses through which data is viewed and entered. 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 anybody to contribute (although controls are available to avoid vandalism). When anybody can contribute, more data are made available, and existing data are readily correctable. Existing texts are readily 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 this project, 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 always 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.

The idea that anybody, 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, Wattenberg, & Dave, 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,

Wattenberg, & Dave, 2004). These same principles apply to MayanWiki, since it employs Wikipedia's software.

1.5 Outline of Remaining Chapters

This work draws from the theories and practice of several disciplines, including Mayan linguistics, corpus linguistics, and computer science (particularly, formal database design). For this reason, I have attempted to make as much of this special project accessible to readers of each of these disciplines, but my principle audience are linguists. Of course, it is not possible to make every section entirely understandable by all, but hopefully the references provided therein will aid the interested reader.

The remainder of this work is organized as follows. First, an overview of previous attempts to create computerized databases of the hieroglyphs is presented in Chapter 2. The subsequent chapters present the design and implementation of MayanWiki. Pertinent to the design of any database is a requirement analysis which seeks to answer two basic questions: (1) What are the data like that need to be stored? and (2) What questions does the database need to be able to answer? (Welling & Thomson, 2003, p. 30). The answers to these questions dictate *how* the data are modeled. Chapter 3 provides a summary of the language and the writing system to present a basic overview of the type of data that will be handled by MayanWiki. Chapter 4 addresses the first question by introducing the wiki frontend through which data are entered. The answer to question two is directly attended to in Chapter 5, which examines the search interface. The database design resulting from the answers to these questions is justified in Chapter 6. Additional useful features not

related to the database, but that are an integral part of the wiki frontend are presented in Chapter 7. Conclusions, discussion, and areas for future research are found in Chapter 8.

2 Previous Work

Although MayanWiki is the only publicly available, linguistic-centered corpus that is editable by the community, several previous and existing databases have similar goals. Although none of these databases has become widely-used resources in the field, all have important elements that have been incorporated into the design of MayanWiki.

The first computerized database of the glyphs was created by the Russians Evreinov, Kosarev and Ustinov (1961), which they used to produce their concordance of the codical signs. Due to the lack of information about their work and since it was limited to the codices, this database has little more than historic value.

The first comprehensive database of the hieroglyphic texts was started at the cusp of the era of the decipherment of the glyphs by Smith-Stark and Ringle (1981), who painstakingly encoded a large portion of the corpus using (updated) Thompson numbers. At the time, not enough glyphs had been deciphered to use syllables and logograms. Furthermore, the use of Thompson numbers would allow for a detailed distributional study of individual variants, unlike a phonetic-based transcription (although this database doesn't appear to have been used for any major decipherment). Unfortunately, by the time of the first publication resulting from this database (Ringle & Smith-Stark, 1996), the hieroglyphs had largely been deciphered. Surprisingly, however, the resulting concordance does not include phonetic readings. Coupled with the fact that the database

is not publicly available, this database is no longer useful, except perhaps to seed a database like MayanWiki through a proper conversion of Thompson numbers to phonetic symbols.

Bricker (1986) produced a database to study the syntax of the hieroglyphs. The database contained 1,000 clauses from 51 sites and the Dresden codex. The transcriptions were made using Thompson numbers, but were also meticulously annotated with part-of-speech information (verb, prepositional phrase, possessed noun, date, Emblem Glyph, etc.). It could return all occurrences of a glyph (presumably in context) as well as frequency information for a single glyph, category in each clausal position (1^{st}, 2^{nd}, etc.), or clause. This database was therefore capable of capturing all the major information stipulated by corpus linguistics theory. Regrettably, this database is not publicly available, does not appear to include phonetic readings, and unfortunately, contains many errors due to its early date.

The Maya Epigraphic Database (Alvarado, 1994) represents another 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." 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[3]. Perhaps the primary reason for this failure is its pre-maturity: it pre-dates 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. The encoding scheme, which is meant to be as objective as possible, is quite cumbersome and difficult[4]. Furthermore, it, like most others, is based on the obsolete Thompson numbers rather than phonetic values. Finally, the lack of a searchable interface within texts is an unsatisfactory oversight.

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 & Looper, 2003) is also directly available in the database, as is likely the case, then the

---

[3] 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.

[4] Chapter 4 briefly explains why a phonetic transcription is desirable, despite the fact that it is not as objective.

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[5]. 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 the public. Another drawback is that it relies on the non-standard, unused cataloging system created by the authors. 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 & Hernández, 2005). This database only encompasses the codices, and to date, only the Madrid codex is viewable and searchable on-line. It, like the MHD, 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 problems. For instance, using the advanced search wizard to find

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

occurrences of "deer" in the Madrid codex, approximately 25 results are returned that include the transliteration and a link to the actual text. Most results contain the word kéeh ('deer' in Yukatek), yet when following the link to the corresponding "frames" (which include a photograph of the text where the hit occurs), the corresponding glyphs are clearly **chi-ji** *chij*, ('deer' in the Ch'olti' family)! Interestingly, the transcription given for this frame is 'keh/chi-hi' (note the mis-transcription of the **ji** syllable). In other words, the transcriptions contain errors, and the transliterations appear to be mainly Yukatek-based (further confirmed by the predominance of Yukatek terms in the transliterations returned in the search results, including tonal information for vowels). Equally frustrating is the fact that a search for *chih* or *chij* returns no results—even though several of the results from a search for *deer* return instances of the latter (surely this could be remedied, however). Most importantly, using this interface, it is not possible to directly study other aspects of language, including 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[6]. 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.

MayanWiki seeks to incorporate the successful elements of past attempts while trying to avoid their pitfalls. Specifically, MayanWiki attempts to leverage the

[6] 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.

"networked scholarship" principle of the MED through its wiki, the public availability of the MHC, and the corpus linguistics theoretic approach of Bricker's (1986) database. If used, the content of MayanWiki is capable of growing to be as comprehensive as the MHD, and perhaps the MHD can even share its information with MayanWiki as an initial seed to the database so that it could eventually converge to more accepted, consensus-based readings.

3 Properties of the Script

Before attempting to model linguistic representations of the script, it is important to first characterize the salient features of the spoken and written language. These features dictate the structure of the database schema for internal storage of the data as well as the syntax for transcriptions and transliterations. Therefore, the properties identified here provide a backdrop against which specific design decisions are justified in subsequent chapters. This chapter does not provide a grammar of the glyphs because the focus of this project is to identify those glyphic and linguistic features that are important to entering, storing, and retrieving linguistic data. The interested reader is instead directed to (Coe & Van Stone, 2005) and the sourcebooks for the Maya Hieroglyph Forum (e.g. Stuart, 2005a); more advanced topics are covered in (Wichmann, 2004).

Typologically, Classical Ch'olti' was ergative-absolutive, VOA (Verb-Object-Agent), more agglutinative than fusional, and fairly polysynthetic[7]. The canonical root is a CVC syllable (occasionally CV syllables are also used and grouped with CVC roots), but multi-syllabic roots (usually labeled CVC+) are also frequent. In the language of the script, the vocalic system consists of five vowels: [i], [e], [a], [o], and [u]. Vowels can be

[7] Note, however, that the descendant languages are increasingly fusional, as evidenced by the fact that certain vowel sequences combine to produce a glottalized vowel at the expense of the original combination of vowels. This has the effect of eliminating morpheme boundaries, thus moving the language from an agglutinative to fusional typology.

simple, long, glottalized (i.e. V'), or subsequently aspirated (i.e. Vh). In addition, there are two semivowels, [w] and [j] (*y* in the traditional orthography[8]). The consonantal system consists of stops—plain and glottalized—fricatives, affricates, nasals, glides, and the lateral liquid /l/ but does not include voiced stops. Instead, ejective plosives and affricates characterize the Mayan languages. However, in place of a bilabial ejective, a voiced bilabial implosive is found. Table 3-1 provides an inventory of consonants in the traditional orthography for Mayan linguistics alongside the IPA equivalent; the remainder of this work employs the traditional orthography.

The writing system itself is quite old, with the oldest known inscription dating to 300 B.C., although the earliest writing is poorly understood (for a treatment of Early Classic writing see Law, 2006). The system was in continuous use up until shortly after the arrival of the Spaniards to the New World. However, most surviving texts date to the Classic Period, between 250 and 800 AD.

The Maya were excellent astronomers and careful record keepers. They were very

Table 3-1 Inventory of consonants in the script in the traditional orthography. IPA equivalents are given in brackets.

	Bilabial		**Alveolar**		**Palatal**		**Velar**		**Glottal**
	voiceless	**implosive**	**voiceless**	**glottalized**	**voiceless**	**glottalized**	**voiceless**	**glottalized**	**plain**
Stop	p [p]	b [ɓ]	t [t]	t' [t']			k [k]	k' [k']	' [ʔ]
Fricative			s [s]		x [ʃ]		j [x]		h [h]
Affricate			tz [ʦ]	tz' [ʦ']	ch [ʧ]	ch' [ʧ']			
Lateral Liquid			l [l]						
Nasal	m [m]		n [n]						
Glide					y [j]		w [w]		

[8] The traditional orthography originated with the *Diccionario Maya Cordemex* and was later adopted by the Academia de Lenguas Mayas de Guatemala (ALMG). In this project, the 'traditional' orthography is employed with the exception that the ' is omitted from *b'* since it is unambiguous.

altars, zoomorphs, stairways, facades, jambs, columns, panels, etc. A few examples of texts carved into wooden lintels and beams also exist. Additionally, texts can commonly be found beautifully painted or inscribed on ceramic vessels, although, due to the limited size of the medium, these texts tend to be shorter in length. And although there is ample evidence that a large number of plaster-coated bark paper books once existed (the Spanish record having burned every book they found), only four examples have escaped destruction: the *Dresden Codex*, *Madrid Codex*, *Paris Codex* and *Grolier Codex*. Other portable objects and natural settings are also host to the magnificent script: figurines, beads, shells, masks, bones, earspools, boulders, and cave walls, among other places. Indeed, "any durable surface seems to have been written on by scribes at one point or another" (Stuart, 2005a, p. 4).

Texts are typically arranged in a grid-like pattern, although other configurations are possible such as a T-shape, or the L-shaped pattern frequently found on pottery. The grid consists of rounded square or rectangular blocks called *glyph blocks*. In most cases, the glyph blocks are read in pairs of columns, from left-to-right, top-to-bottom (see Figure 3-1). Each block is assigned a coordinate within the grid; columns are labeled with letters while rows receive numbers. Each glyph block is in turn the host to one or more (typically two to four) *glyphs*—the basic unit of the script. Generally, the glyphs within a block constitute a syntactic or semantic unit such as a date, a (usually inflected) verb, a proper noun, etc. However, sometimes such units span multiple glyph blocks, and conversely, a single block can contain many units. Within a glyph block, glyphs can be organized in any number of different ways, often with a single, large, square or rectangular "main" sign that occupies most of the block, with other attached, oblong

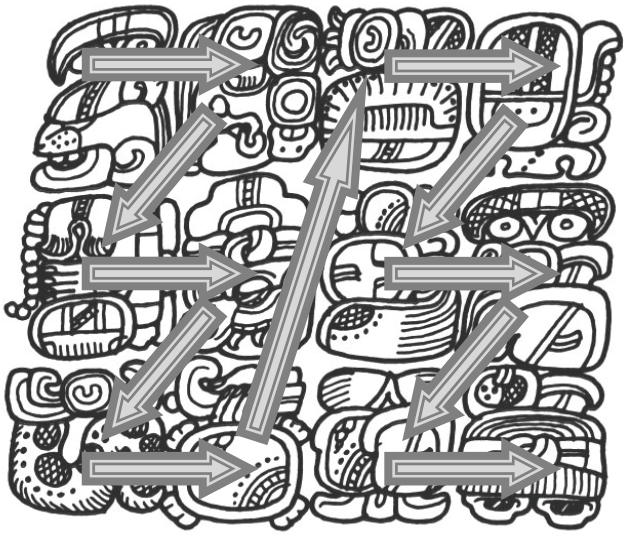


Figure 3-1 Canonical reading order of texts. Glyph blocks are read in columns of two, from left-to-right, top-to-bottom, as indicated by the grey arrows. Underlying drawing by Linda Schele, © David Schele, courtesy Foundation for the Advancement of Mesoamerican Studies, Inc., www.famsi.org.

"affixes". The scribe had some liberty with the order of the glyphs within the block, but generally, they are read from top-left to bottom-right. It should be noted that within a single text, several passages may exist where a passage is roughly equivalent to a paragraph. Dates often mark the beginning of a passage.

As mentioned previously in the introductory chapter, there are two types of glyphs: *logograms* (sometimes called *logographs*) and CV *syllabograms* (signs for each of the plain vowels also exist). A logogram (transcribed with uppercase letters) represents a word root, usually a verb or a noun, which can be inflected by the affixation of additional glyphs. Syllabograms (transcribed using lowercase letters) can also combine together to form word roots. However, since most roots are CVC, or otherwise consonant final, a syllabic spelling has an "extra", unpronounced vowel. More often than not, this vowel matches the internal vowel of the root in a process known as *synharmony*. For example, Knorosov showed that the syllables **ku** and **tzu** combined to form *kutz* 'turkey'

(see Figure 3-2). However, since the syllabic system is unable to represent complex vowels (long, glottalized, or aspirated), *disharmony* can be used to signal vowel complexity, although it can represent simple vowels as well (Robertson, Houston, & Stuart, 2004). Thus, the word *muut* 'bird' is never spelled ***mu-tu**. Instead, it is spelled **mu-ti** in order to indicate the long vowel (Stuart, 2005a) (see Figure 3-2). It should be noted that, on occasion, some syllabically spelled words are actually underspelled, meaning that they are missing their last syllable entirely, and a competent reader was expected to fill in the missing syllable (Zender, 1999).

Syllables are also frequently used to complement logograms, a fact that has greatly aided decipherment of many logograms. *Complementation* consists of affixing a syllable to a logogram that duplicates either the first or last syllable of the logogram to help mark the pronunciation of logograms. For instance, the logogram **CHAN** 'sky, snake' is often followed by the syllable **na** to form **CHAN-na** and **WINIK** 'man' is known to be spelled **wi-WINIK-ki**.

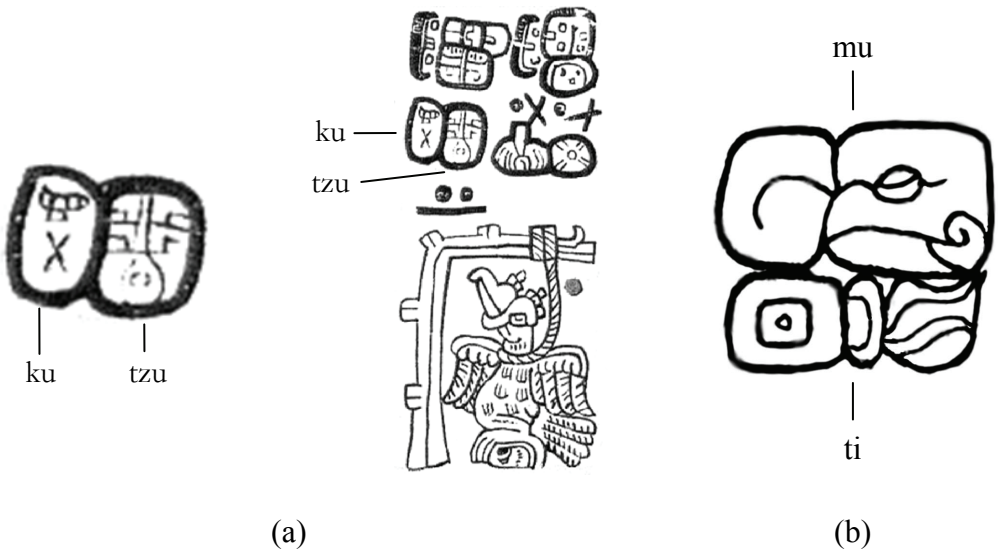


Figure 3-2 Synharmony and disharmony in the script. (a) Synharmony as found in the Madrid codex © Foundation for the Advancement of Mesoamerican Studies, Inc., www.famsi.org (FAMSI). The syllables ku **and** tzu **combine to form *kutz* 'turkey'; the turkey depicted below the text helped Knorosov make this decipherment. (b) An example of disharmony. The syllables** mu **and** ti **combine to form *muut* 'bird' with a long vowel.**

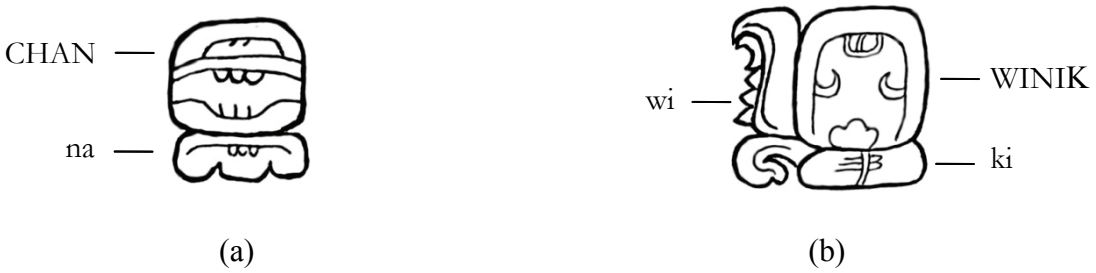


Figure 3-3 Complementation. (a) The syllabic complement na **is appended to the logogram CHAN. (b)** wi **is prepended and *ki* is appended to the logogram WINIK.**

There is another important class of signs called morphosyllables that has been proposed (Houston, Robertson, & Stuart, 2001). Morphosyllables are word-final syllables that represent grammatical meaning in the form of morphemes, and in this sense are simultaneously logographic (hence their usual capitalization). Interestingly, the vowel in a morphosyllable, unlike other word-final syllables, is in fact actually pronounced, although these syllables can be 'reversed', i.e. pronounced as VC. However, the written vowel does not necessarily correspond to the spoken vowel; for this reason, the concept

of disharmony usually does not apply when morphosyllables are used. A frequent example of this is **AJAW-IL**, ajaw-il, 'king-ness'. Notice that in this example, the syllable **li** is actually pronounced [il]; the vowel is known through careful linguistic reconstruction (Houston, Robertson, & Stuart, 2001). In the case of the passive marker -**AJ**, the vowel is always *a* as in **jo-ch'o-AJ**, *joch'-aj*, 'it is drilled'. Two of the morphosyllables (–**WA** and –**YI**) are "regular", and the underspecified vowel simply matches the root vowel. For instance, the common phrase **U-CHOK-WA** was pronounced *u-chok-ow* 'he scatters it'. For the rest, a competent reader must "fill in" the vowel. Although linguistic reconstructions can be helpful, the reflexes of many of these morphemes in the daughter languages differ greatly and are somewhat unpredictable, making it difficult or even impossible, to determine a definitive parent form. For this reason, Houston, Robertson, and Stuart (2001) wisely recommend that these forms be transliterated using a generic *V* for the vowel. A list of morphosyllables and their meaning is included in Table 3-2.

Scribes employed several other processes that were necessitated by the limited amount of space within a given glyph block. At times, one glyph is reduced in size and then infixed inside of another. *Conflation* is a similar process in which two (or more) glyphs are combined together into a single glyph, but each component maintains its same

Table 3-2 Morphosyllables and their function.

Morphosyllable	Source Syllable	Function
-AW	wa	Declarative mood (CVC transitives)
-IY	yi	Medio-passive (CVC root transitives)
-IL	li	Abstractive; marks possessed nouns
-IB	bi	Instrumental
-IS	si	Nominalizer
-AJ	ja	Passive

relative size, and the distinctive features from both glyphs are present in the conflated version. *Superimposition* occurs when one sign occludes a portion of another glyph—as if one glyph were physically on top of the other—usually creating the appearance that part of the occluded glyph is "affixed" to the occluding sign. In fact, some glyphs are almost always occluded by others. Lastly, the Mayan scribes employed a "repeat" symbol consisting of two dots that indicated that the attached syllable was to be doubled. This is the only diacritic in the script (Zender, 1999).

The script is further supplemented by the use of sign substitution (also known as polyvalence). There are many words and syllables that can be represented by more than one logogram or syllabogram, which enabled scribes to express their creativity through their writing. However, not only do logograms and syllables have multiple, distinct possible signs, but some signs function simultaneously as syllabograms and logograms[9], e.g. **ku** and **TUUN**. The exact usage can usually be determined in context or by other means (for instance, **TUUN** is almost always followed by the phonetic complement –**ni**). Note that while a logogram can have multiple pronunciations and even double as a syllable, a syllabogram can only represent one CV pronunciation (Zender, 1999).

This short treatment of the hieroglyphic script hardly does justice to its beauty and complexity. Nevertheless, the principles introduced here will be useful in modeling them in MayanWiki.

[9] Some syllables not categorized as morphosyllables are inherently logographic because they are simultaneously morphemes, such as the dependent ergative pronoun *u-* and the deictic *i*.

4 Data Entry

This chapter discusses the process of entering new data through MayanWiki's interface. The type of data the database backend will need to be capable of handling determines the way that users enter data. The process itself is relatively simple: first a page is declared as containing a *hieroglyphic text*, (i.e. the transcription of a single text), and then metadata about this text (such as site, name, date, and medium) are added. Next, the text is divided into cohesive passages, which are optionally assigned a name. Finally, the transcription, transliteration, and translation of each passage are provided. Each of these steps is explained in turn below; a basic knowledge of using the MediaWiki software is assumed (see MediaWiki Handbook, 2007 for documentation).

4.1 Adding a Hieroglyphic Text to a Page

Since users are free to add any content they desire to MayanWiki, a page can contain information treating almost any topic—from an individual glyph, to the biography of a prominent researcher. However, some pages will contain the transcriptions, transliterations, and translations of texts. MayanWiki makes the simplifying assumption that a single page can contain the transcription of at most one hieroglyphic text. Furthermore, it assumes that a single text will also be transcribed on a single page rather than across multiple pages. In general, this enhances the organization of the content on

Palenque Temple XIX Alfarda Tablet

Medium:	Tablet
Date:	734 AD

Image

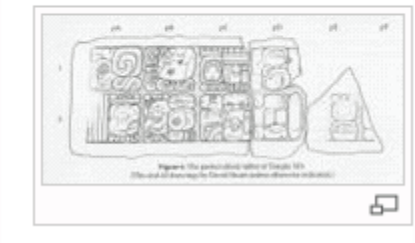


Figure 4-1 Display of metadata resulting from the TextInfo template.

MayanWiki, which in turn simplifies the process of locating information. Any page can contain a hieroglyphic text, although it is strongly recommended that hieroglyphic texts be limited to pages dedicated to information regarding the text itself.

Once a page that is to contain the hieroglyphic text exists, the page is ready for normal editing—by clicking the "edit" tab at the top of the web page. Inside of the edit box (where all edits described in this chapter are performed), a reference to the *TextInfo* template[10] is added at the appropriate place (the user is free to choose the location, but the top of the page will be most appropriate in most cases). The template also specifies the metadata for the text: the site where the text was found (if applicable), the name of the text (for ceramic vessels, this may be the Kerr number), its medium, the date of the inscription (if known), and a reference to an image of the text. This template serves two purposes. First, it indicates to MayanWiki that the page contains a hieroglyphic text. Second, the template produces a "sidebar" (also known as an infobox) when the page is

[10] Templates are used in MediaWiki to produce consistent output among related pages. They typically take arguments that are used to fill in certain parts of the template. In this case, the template produces an "infobox" with the metadata for the site (see Figure 4-1).

viewed; Figure 4-1 contains an example infobox. In the current implementation, the medium can be anything the user pleases; however, it is recommended that users typically choose from the same list employed in (Graham, 1975-2006). The date can be given as either a long count or a Gregorian date. In either case, "?" or "ca." can be used to indicate that the date is uncertain. Finally, the inclusion of a reference or line drawings in the template is particularly helpful where data need to be verified.

4.2 Passages

As explained earlier, many texts consist of smaller cohesive passages that are analogous to paragraphs. A passage is declared on a page containing the *TextInfo* template explained above by using the XML-like syntax shown in Figure 4-2. Each passage is allowed an optional name that should be unique within the text itself, but need not be unique across texts. Of course, passages should appear in the wikitext in the order that they occur in the hieroglyphic text. All transcriptions, transliterations, and translations that are to be added to the database *must* be contained in a passage[11]—even if the particular text contains only a single passage. It is possible, and often useful, for transcriptions to be posted that aren't added to the database. For instance, a user may want to propose an alternative transcription on his or her personal page. Or perhaps small sections of texts are to be transcribed on the discussion page for a particular dispute or proposal. This text is not intended to be added to the corpus as a replacement for the consensus-based version of

[11] Content not contained in a passage will still display on the page and is searchable through the wiki interface, but not from the advanced linguistic search described in Chapter 5.

the text, but obviously it should still be available like any other wikitext. In this case, the text should be added outside of a passage, like all other wikitext.

4.3 Transcriptions

The process of transcription consists of changing the complex graphical form of each glyph to its equivalent phonetic, numeric, or diacritic value. Adding a transcription to a passage is relatively easy: the syntax presented below is followed to present a Romanized transcription of the graphical text within the passage. The exact system used in MayanWiki is inferred from Stuart's (2005b) own transcriptions, which are themselves mostly based on common practice. Invalid input is not accepted by the system and users are warned of syntax errors. The rules are as follows[12] (all examples from Stuart, 2005b):

- Each glyph block is transcribed on its own line
- Each line begins with the coordinate of the block followed by a colon, e.g. **P1: 6-AJAW**
- Glyphs within a block are separated by hyphens, e.g. **8-CHAK-SIHOOM-ma**
- Sub-blocks, when present, are indicated by a space between glyphs of adjacent sub-blocks, e.g. **ba-ch'o-ko ?-NAL-la**
- Logograms are capitalized, e.g. **OTOT**
- Syllables are all lowercase, e.g. **ya**

[12] The syntax described in this section belongs to the class of languages known in computational theory as regular languages (see Sipser, 1997), which is a nice theoretical property that allows for easy parsing. Ignoring some of the finer details (including case), language transcriptions can essentially be recognized by the regular expression of the form (^[a-z]+[0-9]+: [a-z]+ (-[a-z]+)*([a-z]+ (-[a-z]+)*)*$)+. Although not proven here, the part of the syntax pertaining to multi-glyph reconstructions is still regular.

- Morphosyllables are represented in all capital letters, but with the vowel first, e.g. **AJ**
- Numbers are transcribed with Hindu-Arabic numerals, e.g. **5**
- The repeat diacritic is indicated by appending "^2" (in the spirit of the mathematical notation for squared) to the glyph to be repeated; this is regardless of the location of attachment of the diacritic to the glyph, e.g. **3-jo^2-lo**
- Logograms whose phonetic values are uncertain are followed by "(?)", e.g. **SIH(?)**
- Logograms known by nicknames are transcribed in quotes, e.g. **"CHIKCHAN"**
- When a particular instance of a glyph is obtained through a reasonable but uncertain guess, it is followed by a "?", e.g. **CHOK?**
- When an instance is entirely unreadable a sole "?" is used
- Reconstructed data of missing data whose content can be derived is enclosed in square braces (this could include more than one glyph), e.g. **[1-?-?-?]**
- When a glyph block is completely missing from a text, but is known to have existed (for example, if half of the glyph block has been destroyed), it is reconstructed as "[…]", e.g. **12-[…]**

Stuart (2005b) also opts to omit word-initial glottal stops and the apostrophe from the implosive *b'*, while recommending *ts* and *ts'* for the traditional *tz* and *tz'*. MayanWiki adopts the former two conventions, but adheres to the traditional orthography for *tz* and *tz'*.

Nevertheless, there are two other principal deviances from (Stuart, 2005b) worth noting. First, Stuart uses superscript notation to indicate the repeat diacritic and appears to prefer placement of the diacritic relative to its actual occurrence (i.e. either before or after the glyph to which it is attached)[13]. Requiring that the diacritic always follow the glyph in the transcription implies that studies of the distribution of the diacritic cannot be undertaken (i.e. before, after, above or below the attached glyph), but this does not aide the study of the language itself and has already been adequately studied previously, e.g. (Zender, 1999).

The second exception to (Stuart, 2005b) concerns the use of a parenthesized question mark to indicate a logogram with uncertain phonetic value. This notation allows for the disambiguation of a question mark following a logogram which traditionally indicates both an uncertain phonetic value and uncertainty whether a particular token is indeed the transcribed logogram. While it is relatively easy for a trained epigrapher to disambiguate the two uses, it is impossible for a computer to distinguish them without additional information about which logograms are known. To illustrate such ambiguity, consider the Palenque Temple XIX stone panel. All that remains from the text on this panel are three fragments. Since the breaks from the fragments cut across glyph blocks, some glyph blocks are partially missing. Such is the case for the top third of the block at coordinate P6. Nevertheless, it is reasonably clear that the "main" glyph is the familiar **CHOK** 'scatter' logogram. However, there is some possibility that it is a different glyph; in order to indicate that there is some uncertainty pertaining to the reading, P6 is transcribed as **U-CHOK?-ji** (part of the **u** glyph is also missing but that reading is more

[13] Since there is only one occurrence of the diacritic in his book, this is only an educated guess.

certain). On the other hand, I4 and I6 on the south side of the Palenque Temple XIX platform are very legibly the "birth" glyph. This glyph is thought to have been pronounced /sih/, but this has not yet been decisively shown, so it is transcribed as **SIH(?)**. If there were an instance of the "birth" glyph that was partially eroded or otherwise uncertain, it would be transcribed as **SIH(?)?**.

Note that glyph types of totally unknown phonetic value (of which there are very few) can be labeled using their Thompson number (or, for that matter, label from any other catalog) rather than a simple "?". Doing so may allow these glyphs to be studied more rigorously within MayanWiki and lead to their eventual decipherment. However, this would add to the complexity required to contribute to texts (most epigraphers no longer know the T-numbers by memory) and hence discourage contributions. Furthermore, the language itself can be studied equally well whether a T-number or "?" is used to mark these instances. Therefore, MayanWiki does not require anything more than a simple "?".

The system of transcriptions employed in MayanWiki does not encode all available information from the glyphs. For instance, the relative position of each glyph block gives way to the reading order stipulated by the transcriber. The processes of conflation, infixation, and superimposition are all "undone" when expanded into the transcription. The transcription used in MayanWiki also ignores polyvalency: no notion of *which* glyph (syllabogram or logogram) of a particular grapheme is ever indicated. At first, this may seem like a disadvantage in comparison with previous databases which, for the most part, retain a good deal of this information. However, while this extra information may be used in decipherment, it is unnecessary for the study of the language

itself. Furthermore, transcription is less subjective than it may seem—there is seldom any disagreement among trained epigraphers on transcriptions, except for the phonetic values of some logograms. The simpler it is to transcribe texts, the more people will be willing to participate. Indeed, the success of MayanWiki depends on circumspect simplicity.

4.4 Transliterations

The purpose of a transliteration is to convert the glyph-by-glyph transcription into words and morphemes so that the result faithfully represents how the written form would have been pronounced if read aloud. Thus, a transliteration inherently depends on a transcription—even if this transcription is not explicitly written out. For this reason, MayanWiki disallows the existence of transliterations without their corresponding transcription. In MayanWiki, a transliteration is added to a passage simply by adding a blank line after the transcription; MayanWiki automatically interprets what follows to be the transliteration. Although the transliterations are optional, they are highly encouraged since they allow for a more direct study of the language than the glyphic data alone.

The rules of syntax that dictate the entry of transliterations are as follows[14] (once again inferred from (Stuart, 2005b), from which the examples are also extracted):

- One clause is transcribed per line; it consists of a predicate with its associated modifiers (informally, a complete thought)

[14] If the distinction is made between the syntax used to enter transliterations (with their character level tokens) and the syntax of the language itself, then the syntax of the former is also a regular language, although the latter may not be. In essence, it can be recognized by a regular expression of the form (^[a-z]+ (-[a-z]+)*([a-z]+ (-[a-z]+)*)*$)+.

- Case is insignificant; thus, the first letter of each line and/or proper nouns may be capitalized if desired
- Words are separated with spaces, e.g. *Bolon Ik'*
- Morphemes are separated with hyphens, e.g. *Jun-haab-iiy*
- Numbers are spelled out, e.g. *Lajchan*
- Long vowels are represented by doubling the vowel, e.g. *haab*
- Known morphemes with uncertain pronunciation (usually from glyphs of the same nature) are followed by "(?)", e.g. *Wayhaab(?)*
- When a particular transcription contains elements obtained through a reasonable guess, the transliteration should also be followed by a "?", although it may be preferable to treat such elements as entirely unknown
- "...?.." is used when a pronunciation is entirely unknown
- Content from reconstructed glyphs whose content can be derived is presented between square braces in the transliteration
- Content from glyphs that are missing and unknowable is transliterated as "[..?..]"

These rules are similar to those presented for transcriptions—which has the added benefit of simplifying the code used to parse them. It is important to note that the natural unit for transliterations is the clause, but the correspondence between glyph blocks and clauses is not one-to-one. Therefore, it is not always possible to exactly match a clause to its constituent glyphs, and a transliteration will usually contain fewer "lines" than a transcription. However, because clauses represent cohesive units of language, data entry is simplified and user contribution is thus encouraged.

```
{{TextInfo|
|site=Palenque
|name=Temple XIX Platform South Side
|medium=Platform
|date = 734 AD
}}

<passage name="S-6">

K4: 2-6-WINIKI-ji-ya
L4: 15-HAAB-ya
K5: 1-WINIKHAAB?-ya
L5: 9-IK'
K6: CHUM-SAK-SIHOOM-ma
L6: u-NAAH-TAL-la
M1: AJAW-?-ya-ni
N1: ?-NAL-IXIM?
M2: ?-MUWAAN-ni-MAT
N2: K'UHUL-MAT-la-AJAW

cha'-[..?..]-wak-winik-ij-iiy ho'lajun-haab-iiy jun-..?..-iiy
bolon Ik' chum Saksihoom
u-naah-tal ajaw-yan Akan(?)-nal Ixim ..?..-Muwaan-Mat K'uhul-Matwil-Ajaw

Two days, six winals, fifteen years and one-score years later
It is Nine Ik', the Seating of Saksihoom.
It is the first becoming a lord of Akan?-nal Ixim ? Muwaan Mat, the Holy Lord of
Matwil.

</passage>
```

Figure 4-2 Example markup for passage S-6 of the main text of the south side of the platform at Temple XIX at Palenque.

4.5 Translations

Once a phrase has been transcribed and transliterated, it can readily be translated. For now, MayanWiki only accepts English translations. It is important to note that there are some elements not transliterable, but which are translatable when the "meaning" of a lexical item is known but its linguistic reconstruction is not (e.g. the birth glyph). In such cases, the translation complements transliteration by providing useful additional information that is otherwise unknown in transliteration.

As with transliterations, an empty line separates translations from transliterations. Each line of the translation should correspond to a single clause of the transliteration; that is, there should be the same number of lines in the translation as there are clauses in the

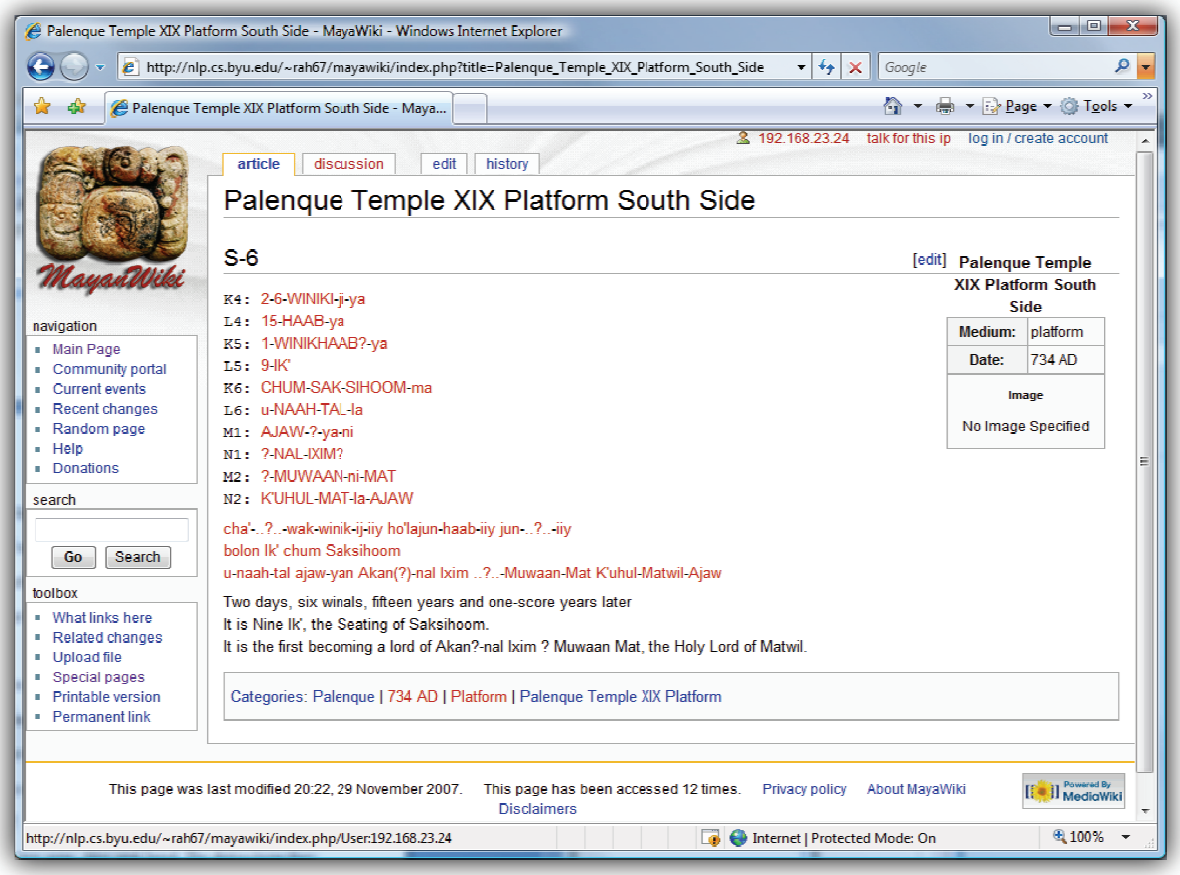


Figure 4-3 Output resulting from markup shown in Figure 4-2.

transliteration. Thus, by definition, a clause should be translatable in such a way that the result is cohesive without having to cross phrase boundaries. Otherwise, translations are unrestricted as to their structure and the content they can hold.

4.6 Example

To illustrate the data entry process, an example taken from (Stuart, 2005b) will suffice. In this example, passage "S-6" from the main text of the south side of the platform at Temple XIX at Palenque is transcribed, transliterated, and translated. The markup is found in Figure 4-2; the output is shown in Figure 4-3.

4.7 Summary

The process of transcribing, transliterating, and translating a text using MayanWiki is relatively easy and is based on familiar syntax. One important outcome of this chapter is the standardization of the transcription and transliteration syntax. As Stuart (2005b) notes, "publications in Maya epigraphy are highly inconsistent when it comes to transcribing hieroglyphic signs." (p. 8). The syntax presented here should help alleviate this problem. The simplicity and familiarity of the syntax should furthermore encourage more users to contribute texts to MayanWiki. In fact, since the success of MayanWiki depends on its simplicity, the consistent syntax is important to MayanWiki's success.

5 Search Engine

Serious study of language is best accommodated with linguistic data that are readily accessible and easily manipulable. Once data are present in MayanWiki, its search engine allows for quick access and meaningful manipulation of the data. Even though the database itself has a perfect memory, it is merely a tool capable of doing only what a human agent instructs it. Data analysis begins by determining *what* should be studied and then *how* that data should be arranged to facilitate analysis. Observations derived from the MayanWiki database will allow for significant new understanding of the hieroglyphic corpus.

The MayanWiki search engine is a state-of-the-art tool with an AJAX-enabled user interface. Since MayanWiki, like any other database of the hieroglyphs, is a corpus, its search engine has been designed around the methodology of corpus linguistics. As explained in the introductory chapter, corpora are typically used to study three basic aspects of language: phraseology, frequency, and collocation. This chapter discusses the types of searches that make study of these aspects of language possible in MayanWiki; each is presented in turn.

5.1 Phraseology

The basic mechanism for studying phraseology is the use of concordance lines that show particular search terms in the context they occur (Hunston, 2002). Although the study of phraseology would seem to be more applicable to the linguistic data contained in the transliterations, concordances of the glyphic data are useful as well. MayanWiki allows for searches to be performed in both areas.

The most basic search is locating all occurrences of a particular glyph or lexeme. For instance, one may want to find all occurrences of the **–AJ** morphosyllable or the lexeme *pakal* 'shield'. Note that searching for logograms will not always produce the same results as searching for the corresponding lexeme. Often times, lexemes had both syllabic and logographic spellings, resulting in multiple representations of the same word. This is analogous to English which permits the use of the logograph '1' or the phonetic 'one' to represent the same quantity and which are pronounced exactly the same. To exemplify, imagine a search intended to identify all occurrences of the word *pakal* 'shield'. A search for **PAKAL** in the transcriptions will unfortunately omit all occurrences of **pa-ka-la**; however, searching for *pakal* in the transliterations will return all results. Thus, it is usually preferable to search transliterations for lexemes.

Searches are not limited to single glyphs or lexemes/morphemes, however; any number may be strung together[15]. For instance, it may be of linguistic interest to investigate the suffix **ji-ya**, or perhaps the proper name and title, "K'uhul Matwil ajaw".

[15] MySQL, the backend for MayanWiki, has some limitations that effectively limit the number of contiguous glyphs/lexemes, but this limit exceeds the length of practical searches.

In both cases, proper study would require finding all instance of the corresponding phrase.

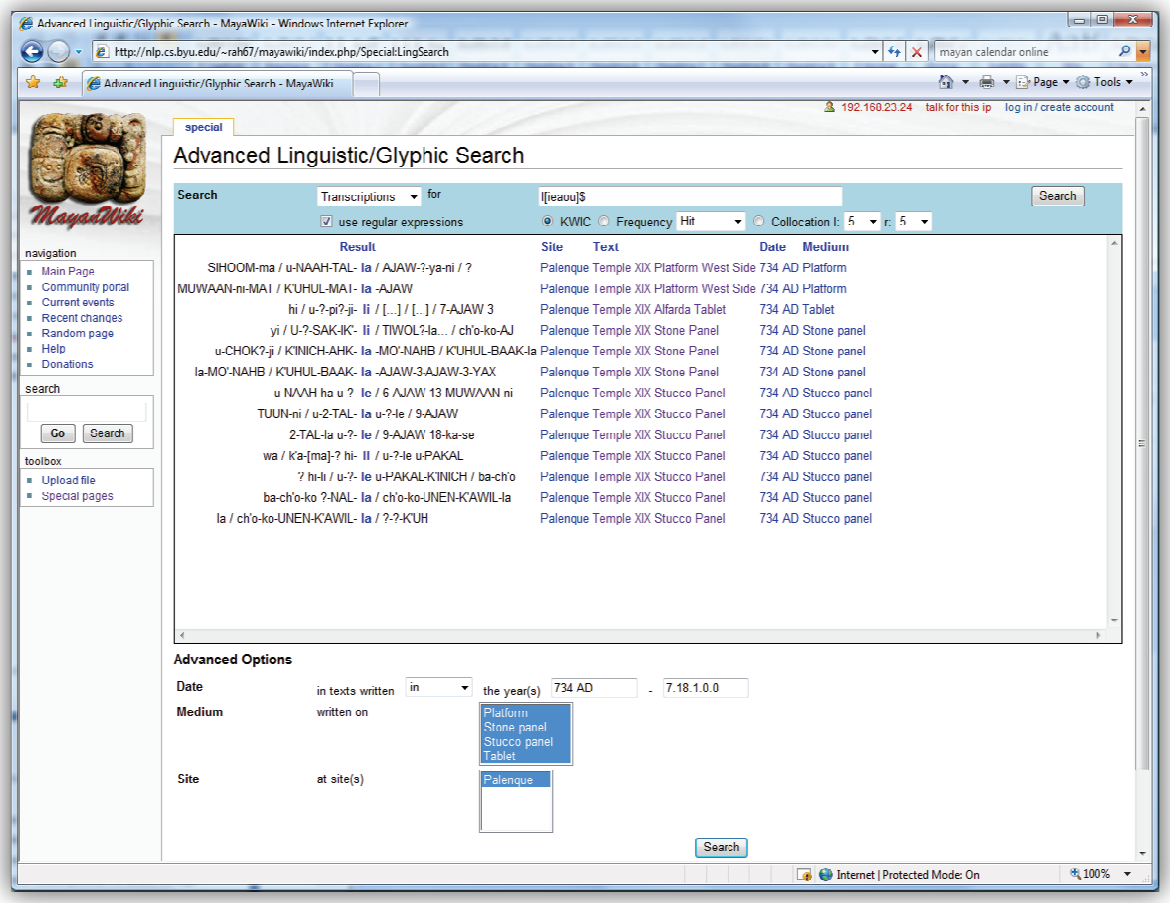
Searches can be more powerful still: any valid POSIX regular expression can be used in place of a glyph/lexeme. Although the usages of this flexible feature are nearly limitless, a few examples will demonstrate some practical uses of regular expressions. The simplest use of regular expressions is wildcard expressions. For instance, one might be interested in studying the occurrences of all "l" syllables (**la**, **le**, **li**, **lo**, **lu**). In this case, a search for `l.` would return the desired results since the '.' character stands in the place of any character. Another use of wildcards might be to study all of the syllables ending in "a", in which case `.*a` would work (`.+a` would ensure there was at least one character before the *a*). Character classes are even more useful. For instance, we might be interested in finding all occurrences of CVC roots, which could be approximated with a `.[ieaou].` pattern. Logograms can be found by searching for all uppercase glyphs, e.g. `[A-Z']+`; syllables could be found by searching for all lowercase glyphs: `[a-z']+`.

Another use of regular expressions is the specification of alternate search strings. This may be particularly useful in accounting for differences in spelling rules. Although there tends to be little disagreement for transcriptions, some differences exist in the way lexemes are transliterated. Within the wiki, these differences will eventually straighten themselves out to a general consensus. In the meantime, however, these types of differences can be accounted for by specifying known alternative spellings, e.g. `took'|to'k'`. It is also possible to specify that any character or parenthesized groups of characters are optional by adding a question mark, e.g. `took'|to'o?k'`.

All searches can be limited by date, medium, or site. Dates can be specified as an exact date (e.g. 734 AD), or as a range (e.g. 250-900 AD). They can also be expressed as Gregorian dates or in Long Count notation to accommodate the aforementioned need to use both. It is also possible to specify more than one medium and/or site as search criteria. These criteria allow for a thorough study of diatopic and diachronic language variation. For example, the acquisition of data by Hruby and Child (2004) for their remarkable study on the introduction of *–wan* from Chontal to Classical Ch'olti' could have been done in a matter of minutes in MayanWiki (assuming, of course, that it was fully populated); likewise for other studies (e.g. Hruby & Robertson, 2001).

The search results are returned in a table that includes the "hit" within a context of

Figure 5-1 Example search results.



several glyphs/morphemes to the left and right, the text on which the hit occurs, the site of origin of the text, the date that the text was inscribed, and the medium that the text is inscribed on (see Figure 5-1). Conveniently, the results can also be sorted by these same columns: alphabetically by the text of the hit, site name, date, or medium. The data can be sorted ascending or descending and by multiple columns (e.g. ascending by hit and descending by date). This allows the data to be organized in ways more conducive to finding interesting patterns that can lead to insightful deductions about the language.

5.2 Frequency

The frequency with which words or phrases occur can contain useful information about language. Indeed, most successful speech and natural language processing tools such as speech recognizers, internet search engines, part-of-speech taggers, and machine translation tools, use statistics based on the frequency of words. Although absolute frequency can contain useful information about language (for example, the most frequent words tend to be the most irregular), study of the relative frequency of words between two parts of a corpus (e.g. two time periods or two geographic regions) is generally more informative. Using this technique, the distribution of words can be studied over time, for instance, to compare Pre-Classic, Classic, and Post-Classic texts. Differences in media or between sites can also be analyzed.

In MayanWiki, any search that can be performed as explained in the previous section can be analyzed by frequency. After selecting the option for returning frequency information, the user is then allowed to select which fields to group frequency

information by; that is, whether to compare frequencies by hit, site, medium, or by date: tun, k'atun, or bak'tun (analogous to year, decade, and century, respectively).

The table of results used to display frequency information differs from the one used to display concordance lines. The first column contains the data for the group for which frequency information has been requested. For instance, if the search was grouped by media, then the first column would be "media" and would contain one row for each possible medium, e.g. "Stone tablet", "Stela", "Vase", etc. The next column contains the raw count for the number of times the search term was found for the group by field.

Although this number can be useful, it can also be misleading. For instance, suppose we would like to study the distribution of the **u** glyph across different sites. This glyph will appear many more times at a site like Palenque than it will at the much smaller Pomoná. That does not mean, however, that the scribes of Pomoná used the **u** glyph less frequently than the scribes at Palenque. Instead, this is merely an indication that there are less *total* glyphs at Pomoná than Palenque. For this reason, the third column consists of the frequency of the search term relative to the total number of tokens for the group by column (e.g. the total number of glyphs at Pomoná). These normalized values are typically more interesting than the absolute counts. In the example given above, we would expect to find approximately the same proportion of **u** glyphs at both sites.

5.3 Collocation

Collocation is the study of the setting in which a given word occurs, which is a significant factor in the study of language. In English, for example, the word *head* tends to co-occur most frequently with *office*, *department*, and *state*—even more so than body

parts like *hands*, *shoulders*, and *eyes*—suggesting that the metaphorical use of *head* is more common in English than the physical use (at least in the British National Corpus). MayanWiki provides collocational information by analyzing a specified number of words to the right and left of the search term. Any terms that can be used in a regular search can also be used to find collocational information. However, while it is possible to perform searches based on collocational information over the transcription data, this type of search is much more useful with the linguistic data contained in the transliterations.

The information returned by a collocate search includes a list of individual words found nearby the search term, the number of times each one was found in this context, the total number of times the collocate was found in the corpus, the percentage of times that the collocate appears in this context, and the pointwise mutual information[16] for this collocation. Pointwise mutual information helps differentiate collocations that are "accidental", occurring because one or both terms appear with relatively high frequently and thus co-occur out of chance rather than significance[17].

5.4 Summary

MayanWiki has a very powerful search engine that allows for flexible searches based on phraseology, frequency, or collocations. Results can be sorted in diverse ways to help the researcher discover patterns that may otherwise be difficult to discern. When used correctly, regular expressions can enhance searches performed in MayanWiki. All of

---

[16] Mathematically, pointwise mutual information is defined as $\log \frac{p(x,y)}{p(x)p(y)}$.

[17] Several authors have noted the shortcomings of this measure (Manning & Schütze, 1999; see Church & Gale, 1991) and instead recommend Pearson's χ^2 test and/or likelihood ratios; Manning & Schütze (1999) suggest that the latter most befits the study of language.

these features enable the data to be searched and manipulated faster than ever before, which in turn will allow the data to be seen as never before. Even the most capable epigrapher cannot possibly manipulate the data in his or her head as quickly as the computer can nor even in the same ways as the computer. Clearly, then, MayanWiki has the potential to accelerate progress in the field as never before.

6 Database Schema

The use of relational databases to store linguistic data is an emerging approach for storing corpora—particularly very large corpora. This approach has been successfully leveraged by Davies (2005; in press) to allow large corpora to be searched nearly instantaneously using flexible criteria including such items as date, register, part-of-speech, lemmas, word stems, synonyms and more. This architecture has been successfully applied to the British National Corpus, Corpus of Contemporary American English, TIME magazine, Corpus del Español, Corpus do Português, Oxford English Dictionary, Early English Books Online, and Literature Online, all of which are available from http://corpus.byu.edu. Although there are similarities between Davies' databases and MayanWiki's, there are also fundamental differences. One important difference is that the languages represented by the aforementioned corpora are more fusional or analytic than polysynthetic and/or agglutinative. Hence, the basic unit of these corpora is the (fully inflected) word whereas the morpheme/lexeme is more appropriate for more polysynthetic or agglutinative languages. Another significant difference is that the data stored in MayanWiki are subject to user modification.

This chapter details the design of MayanWiki's database based on the requirements outlined in the previous chapters. First, an overview of some principles of sound database design is presented and the need for the separation of the conceptual,

logical, and physical design is motivated. Each of these levels of design is then presented in turn; the corresponding diagrams can be found in Appendices A-C.

6.1 Database Design Principles

"Those who are enamored of practice without theory are like a pilot who goes into a ship without rudder or compass and never has any certainty where he is going. Practice should always be based upon a sound knowledge of theory."[18]

Leonardo da Vinci (1452-1519)

A surprising number of databases, both public and private, are poorly designed. The designers are often unaware of the cost of their decisions, excusing their lack of foresight and planning with the affirmation, "it works, doesn't it?" If indeed it is possible to produce a working database without the overhead of a structured design process, what, then, are the motivations for so doing?

Simply stated, good design saves time and frustration in both the short and long terms. The principle source of problems in a poorly engineered database (which usually results from lack of design) is redundancy. Besides wasting space and other resources, redundancy opens the door to insert, update, and delete anomalies (see Welling & Thomson, 2003). If data are unnecessarily repeated in the database then information must also be manually and needlessly repeated during inserts or updates. Even the slightest mistake during these operations could result in inconsistent data, rendering it virtually useless. A delete anomaly occurs when the existence of one entity (e.g., the branch of a bank) inadvertently depends on the existence of other entities (e.g., loans). If all of the loans for a particular branch get removed (e.g. once they are paid off), the branch effectively ceases to exist in the database—usually an undesirable property. In all of these

[18] This quote was taken from (Date, 2005).

cases, data redundancy unnecessarily burdens the application programmer with the responsibility of maintaining consistency. These problems require additional initial programming time and make the application more error prone; once problems do appear, they can be very difficult to track down. Furthermore, when application requirements change, it may be difficult to integrate the changes into a poorly designed database. On the other hand, a well-designed database with no inconsistencies minimizes initial programming time and maintenance and is more amenable to future modifications. Clearly, the benefits of good design more than compensate for the initial effort required and those who fail to design their database properly will eventually pay the (often much higher) price.

A well-designed database is often created through a process consisting of four important stages: data requirements analysis, conceptual design, logical design, and physical design[19]. One purpose of previous chapters has been to outline the requirements of the data used in MayanWiki; this chapter presents details relating to the other three phases. Conceptual design identifies those processes necessary for a particular application, what data are necessary for performing these processes, and the relationships among—and constraints on—the data. This is done in an abstract manner that is independent of the technology that will be used to implement the design, whether as a relational database, object relational database, XML, etc. The purpose of logical design is to make decisions related to the type of database used to represent the conceptual design,

---

[19] There is significant disagreement about the exact division of the conceptual, logical, and physical designs. In this chapter, I consider the conceptual design to be all information that can be represented independent of the type of database, the logical design to contain all information dependent on the type of database (e.g. relational database) but independent of vendor, and the physical design to include all information that could vary depending on the specific choice of database (e.g. MySQL).

but without committing to a specific vendor. For instance, a relational database can be chosen as the technology without committing to Oracle, MySQL, or SQL Server—each of which exhibit various differences—in which case the conceptual schema would subsequently be converted to relational tables. Finally, the physical design process involves selecting a particular vendor and then making decisions related to final implementation. For a relational database, this involves generating SQL statements to create tables, determining which columns to index, choosing which indexes should be clustered, etc.

This chapter presents the conceptual, logical, and physical design based on the data requirements outlined in the previous chapters. Besides leading to a better overall design, this process should make the design of MayanWiki's database more transparent and easier to understand. Certainly, understanding the entities and their relationships as presented in the conceptual model is much easier than to understand their corresponding roles and functions in the various `CREATE TABLE` SQL statements. In addition, the clarity created by this design process will make it simpler for similar projects to adapt and implement relevant parts of the database schema. This includes not only projects that involve storage of hieroglyphic data (although use of MayanWiki itself is encouraged in many cases), but also similar data in other scripts or languages—especially those that are more polysynthetic and/or agglutinative morphologically. Even though modifications may be necessary, the task is made simpler with information from all three phases of the design process. The information in this chapter should also make it possible to re-implement the ideas, whether in whole or in part, using a technology other than a relational database (e.g. an XML or object-relational database).

6.2 Overview of the Entity-Relationship Model and Diagram

Probably the most common conceptual model is the entity-relationship (E-R) model introduced by Chen (1976), which has since been elaborated; the overview presented here closely follows (Silberschatz, Korth, & Sudrashan, 2002). In an E-R model, an *entity* is simply an object, usually a person, place, or thing to be modeled from the real world. For instance, in a banking system, an individual customer, a loan, and a specific branch are all entities. The collection of all entities of a given type (e.g. all customers, or all branches) is called an *entity set*. Each entity can have any number of *attributes*, or properties, that define it; all entities in an entity set share the same attributes, although each entity can have different values for the attributes. Thus, each customer could have a social security number, an address, and phone-number; a loan might have a loan number and an amount due; and a branch would possibly have a name and a city. Some customers may share the same phone-number (e.g. husband and wife) and some loans may have the same amount due, but usually these are different.

An attribute or set of attributes that uniquely identifies a single entity within an entity set is called a *key*. Since multiple keys can exist, the *primary key* is the key that is chosen by the database designer to be used to uniquely identify entities. To illustrate, the social security number can be used as the primary key for the employee entity set and the combination of the branch name and city could be the primary key for the branch entity set, if branches in different cities are allowed to have the same names (otherwise, the branch name should be the primary key). All entities but *weak entities* are required to have a primary key. A weak entity is one whose identity is only unique when combined with the primary key of the *owner entity set*. In the banking example, if we were to

consider payment as an additional entity with the attributes "payment number" and "payment amount", it is useful to consider "payment number" as a type of unique identifier, yet it would be unique only for the loan with which it is associated. Hence, payment is a weak entity set that depends on the loan entity set, where "payment number" is the *discriminator*.

Entities are related to one another through *relationships*, which may also have attributes. The set of relationships between more than one entity set is called a *relationship set*. In the case of the bank, customers "have" loans and are "members of" a particular branch. The relationship between a weak entity and its *owner* is called an *identifying relationship. Cardinality ratios* specify how many entities from an entity set are involved in the relationship to another entity set. For instance, a customer may have many loans and a loan may be held by multiple customers (for instance, by spouses); this is an example of a many-to-many relationship. Other relationships include one-to-one, one-to-many, and many-to-one (depending on the direction of the relationship). A *participation constraint* specifies whether all entities in an entity set are required to participate in a given relationship. When all entities are required to participate, this is called *total participation*; otherwise, the participation is *partial*. To exemplify, while every loan at the bank must be associated with a customer, not every customer need have a loan in order to be a customer. Thus, the customer entity set has partial participation in the relationship, while that of the loan entity set is total.

Each of these concepts can be visualized in an E-R diagram using the following components, adapted from (Silberschatz, Korth, & Sudrashan, 2002):

- **Rectangles** depict entity sets

- **Diamonds** are used to represent relationship sets
- **Ellipses** denote attributes of an entity set or relationship set
- Attributes are **underlined** to indicate they form part of the primary key
- Discriminators of weak entities are **doubly underlined**
- **Double rectangles** indicate weak entities
- **Double diamonds** represent the identifying relationship of a weak entity
- An **arrowhead** on a connecting line between an entity set and a relationship set indicates the "one" side of a one-to-one or many-to-one relationship. The arrowhead is always adjacent the entity set.
- A **bold** connecting line designates a relationship as total

More advanced concepts and components exist, but this basic set will suffice for the current project; more details can be found in (Silberschatz, Korth, & Sudrashan, 2002).

6.3 Conceptual Model: Entity-Relationship Schema

This section presents the entities and relationships pertinent to MayanWiki's design. Each sub-section presents a specific entity along with its relationship to other entities in the database. In the process, simplifications and other modeling decisions are justified based on information contained in the previous chapters. The full E-R diagram is presented in Appendix A.

6.3.1 HieroglyphicText

At the core of the database is the *HieroglyphicText* entity. Each *HieroglyphicText* has a unique *Name* (e.g. "Copan Stela J"). Similarly, because Maya writing was in use over

such a wide span of time, many interesting temporal differences also exist; for an example of a study based on temporal differences, see (Hruby & Child, Chontal linguistic influence in ancient Maya writing, 2004). For this reason, the composite attribute *InscriptionDate* is posited for the *HieroglyphicText* entity consisting of a *JulianDay* field and a Boolean marker, *isApprox*, that indicates whether the date is approximate (those obtained from the texts themselves are seldom approximate). The use of a Julian day is motivated by the fact that it is sometimes convenient to work with Gregorian dates, but more culturally appropriate—and common—to use the Mayan Long Count. Storing the Julian day allows for efficient comparison between dates in queries and simple conversion between the Long Count and the Gregorian date. Since not all dates are exact, e.g. those derived from carbon dating or guessed from the calendar round, dates should be allowed to be flagged as approximate. This allows searches to specifically include or exclude such dates.

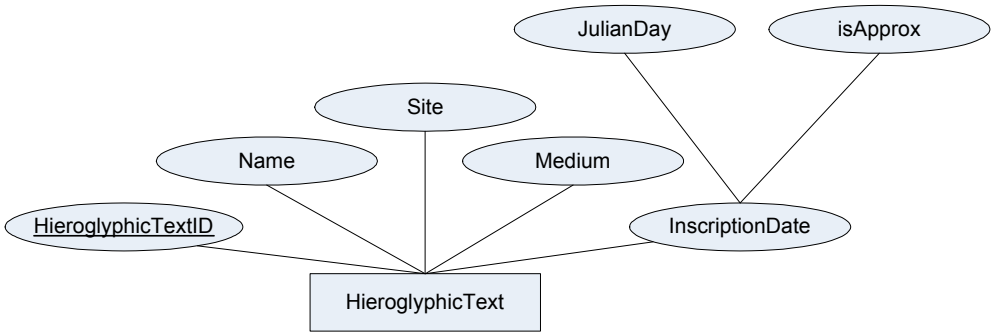


Figure 6-1 HieroglyphicText entity.

6.3.2 Site

As has been previously discussed, it may be interesting to study diatopic language variation within the script[20]. Naturally, every text was written in some geographic location by a scribe who himself originates from (a possibly distinct) geographic region. It is reasonable to assume that texts originating from nearby geographic regions may reflect interesting localized linguistic trends or patterns (though the actual grammar has been shown to be remarkably consistent; see footnote 20). Some differences are known (Houston, Robertson, & Stuart, 2001), but others may only be discernible through careful study using corpus linguistic methodology, such as that offered by MayanWiki. Since most texts are found during excavations of a particular archaeological site, it is reasonable to track texts by their site of origin, when known.

This gives rise to the *Site* entity set, which contains the names of all the possible sites where texts could have originated. Each text is presumed to have originated from one site. Since the origin of some texts—most notably looted material—is unknown, and

[20] It is worth noting here that there is remarkable consistency across the lowlands in the grammar of the glyphs, even in locations where distinct languages were spoken (e.g. Copan vs. Chichén Itzá; see Houston, Robertson, & Stuart, 2001). Nevertheless, some minor variation does exist which may prove interesting.

it is possible for a site to exist in the database without texts (see below), the participation of both *Site* and *HieroglyphicText* in their relationship with each other is partial.

In the current implementation, the *Site* entity set may seem trivial, consisting of only one attribute: the name of the site (*Name*). Indeed, rather than creating a separate entity, it would have instead been possible to track the name of the site as an attribute of *HieroglyphicText*. However, creating a separate entity set for *Site* has several advantages. First, it allows sites to be added to the database even before texts are added for that site. This would require users to add texts only for sites that already exist in the database; users would first need to add a page for the site, then to add the text. Although this is currently not the case, it would help reduce possible errors, and consequently, increase the total content of MayanWiki. For instance, a misspelled site name in a text would be flagged as an error rather than associating the text with a new site with the misspelled name, as is currently the case. Another advantage is that, in the future, new attributes and relationships could be added to the *Site* entity set. For example, a new attribute could be added to *Site* that specifies the Classical Ch'olti' name for each site. Furthermore, sites could be grouped by broad geographic relationships, allowing studies about broader relationships; political relationships among sites could be tracked as well. Finally, creation of a new entity avoids a delete anomaly wherein the deletion (or absence) of all texts for a site would create a situation in which that site is completely unknown to the database. For these reasons, it seems preferable to treat a *Site* as a separate entity.

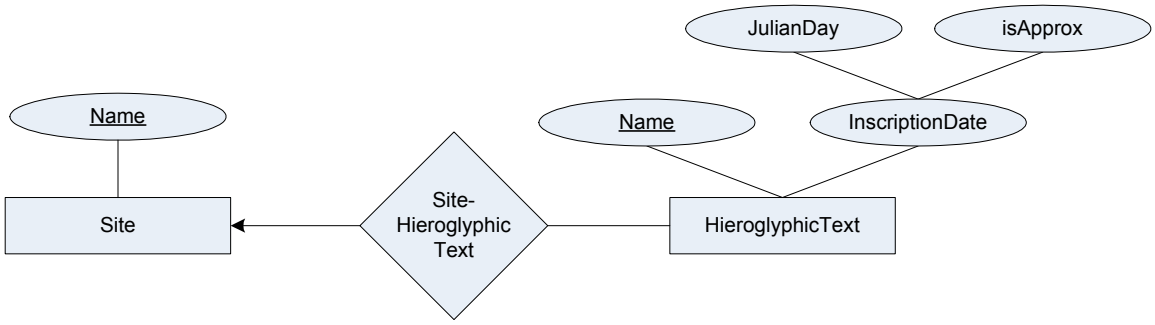


Figure 6-2 Site entity.

6.3.3 Medium

Another variable of interest that has been mentioned is the media on which texts are written, which corresponds roughly to the concept of "register". For this purpose, the *Medium* entity set was created with the sole attribute *Name*. It is presumed that the medium for all texts is known, although, at least while the database is being initially populated, not all media will necessarily be used. Thus, the participation of *HieroglyphicText* in the relation with *Medium* is total, while *Medium*'s participation is partial.

While, like *Site*, it could be possible to create an attribute on *HieroglyphicText* to record this information, for all of the same reasons as explained with *Site*, it is preferable to create a separate entity. For instance, it may be desirable to predefine a set of media that users must use. Not only does this reduce errors, but it helps limit the domain that can be used, simplifying not only the user's task, but searches performed based on media as well. Another potential use of *Medium* as an entity is to create a hierarchy of media that could allow searches to be performed at broader (or narrower) levels with relative ease.

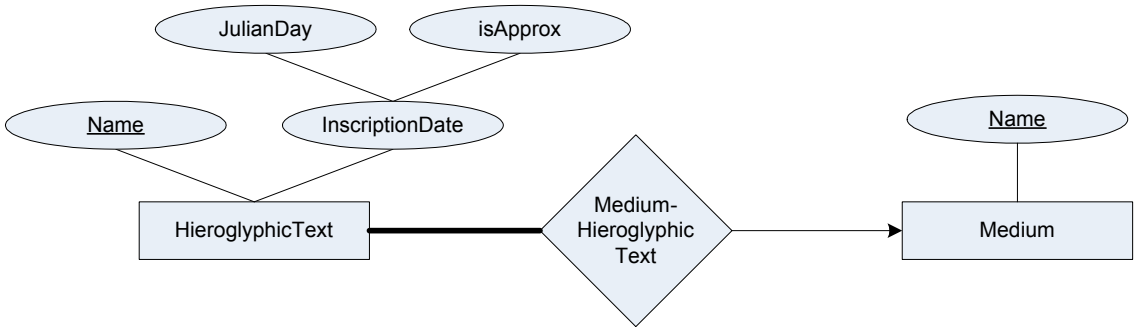


Figure 6-3 Medium entity.

6.3.4 Page

Each page in the wiki, regardless of whether or not it contains a hieroglyphic text, is an entity (*Page*) with a unique *Title*. Furthermore, each *Page* can host only one *HieroglyphicText* because it is much easier to locate texts when each *HieroglyphicText* is on its own *Page*. Furthermore, using the current mechanisms described earlier, it is easier to add texts to the database when the entire text is on a single page. Moreover, allowing more than one *HieroglyphicText* per *Page* could result in long, difficult to read pages.

Additionally, MayanWiki requires each *HieroglyphicText* to be transcribed only once, on a single *Page*. Without this restriction, a text could be transcribed in several different places. At first glance, the ability to house multiple transcriptions of the same text may seem desirable. For example, if one user disagrees with a particular transcription, they could just add their own; other users of the wiki would be free to compare the various versions of the transcriptions for a particular text. However, this presents some serious challenges. A major problem is that multiple transcriptions will complicate searches. This would result in searches cluttered with different versions, making it difficult to sort through the data. Furthermore, frequency and collocational information would also be skewed if texts could be repeated multiple times. The biggest

problem, however, is that multiple versions of a text undermines the wiki principle. If it is easier to produce alternative transcriptions than to resolve differences through careful, thorough discussion, then differences of opinion will tend to diverge rather than converge—a direct contradiction of the original goals of the wiki. After all, Wikipedia does not allow three or four versions of a controversial article to float around.

Therefore, the relationship between *Page* and *HieroglyphicText* is one-to-one; *HieroglyphicText*'s participation is total, while *Page*'s is partial (every text must be on a page, but not every page need host a text).

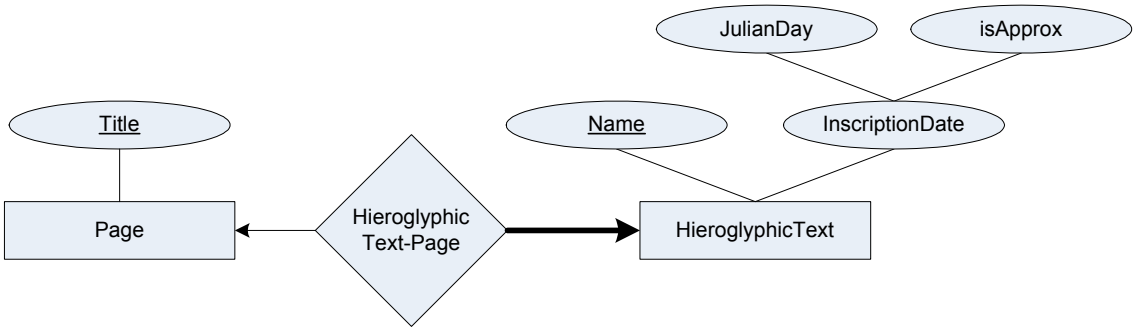


Figure 6-4 Page entity.

6.3.5 Passage

Texts are divided into one or more passages. A *Passage* has a *PassageNum* that identifies its order within a text, in addition to an optional *Name*. However, since every text will have a "first" *Passage*, *PassageNum* is not sufficient to uniquely identify a *Passage*. For this reason *Passage* is modeled as a weak identity owned by *HieroglyphicText*. Note that it is possible to declare that a text exists without necessarily adding passages to it. Thus, its participation in the relationship is partial (the participation of all weak entities is necessarily total).

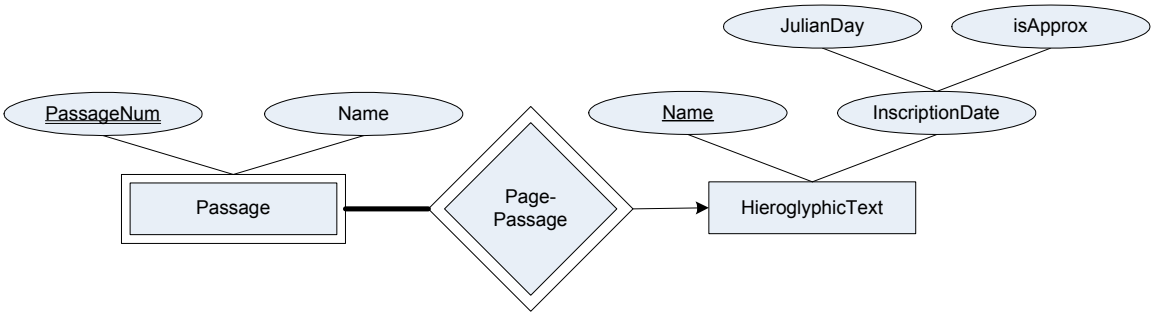


Figure 6-5 Passage weak entity.

6.3.6 GlyphBlock and Line

Nearly every *Passage* will be transcribed using one or more *GlyphBlocks*, although in rare cases in which a passage is known to exist but is missing or obliterated, a *Passage* may exist as a "placeholder" without a *GlyphBlock* (a one-to-one and partial relationship); the reverse direction of the relationship is clearly many-to-one and total. A *GlyphBlock* is a weak entity with a *GlyphNum* as the discriminator that serves as an index for each *GlyphBlock* within the *Passage*. Every glyph block additionally has an associated *Coordinate* that can be used to locate the glyph block on a pictorial representation of the text.

A transcription is transliterated using at least one *Line* of words and morphemes. Like a *GlyphBlock,* a *Line* can be treated as a weak entity that holds sequential *LineNumbers* within each *Passage* (each *Passage* starts on line 1). In addition, each Classical Ch'olti' *Line* can also have a corresponding English *Translation*. The process of transliterating a text inherently requires that the text be transcribed, whether implicitly "in the head" of the transliterator, or explicitly as an intermediate step. Since the goal of MayanWiki is to collect all known transcriptions and transliterations, it is reasonable to require that the translation used to create the transliteration also be present. Hence, the

relationship between *Passage* and *Line* is one-to-many; while the participation of *Passage* is partial, that of *Line* is total.

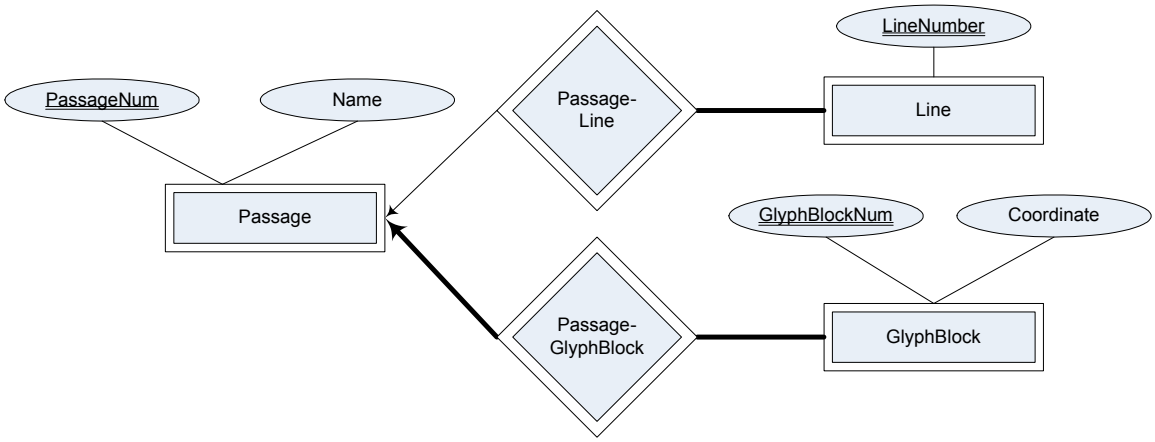


Figure 6-6 GlyphBlock and Line weak entities.

6.3.7 SubBlock and WordToken

Recall that a glyph block may actually contain more than one sub-block. Hence the need for the *SubBlock* weak entity set with the discriminator *SubBlockNum* which provides the index of the S*ubBlock* within each *GlyphBlock*. Similarly, a *Line* can contain multiple *Word*s, each of which can be indexed by their position in a *Line* via the discriminator *WordNum*. It should be noted that neither of these entities contain actual word or sub-block types. That is, these entities are *not* dictionaries of words or sub-blocks. They simply serve to mark the word number or sub-block number of each (morphemic/glyphic) token. In polysynthetic languages, dictionaries are seldom created at the "word" level since the high degree of inflection leads to an enormous number of possible word types[21]. Furthermore, the inflections do not alter the core meaning of the verb. Just as inflected

[21] Since the vast majority of inscriptions are in third person, and the style of discourse is fairly restricted, the actual number of "word" types is relatively low for the hieroglyphs, but this is a matter of principle: there is no need to store word types and unnecessarily complicate the database design.

verb forms in Spanish are not found in a dictionary, neither is it necessary that they be modeled here; similar arguments apply at a graphemic level with *SubBlock*s.

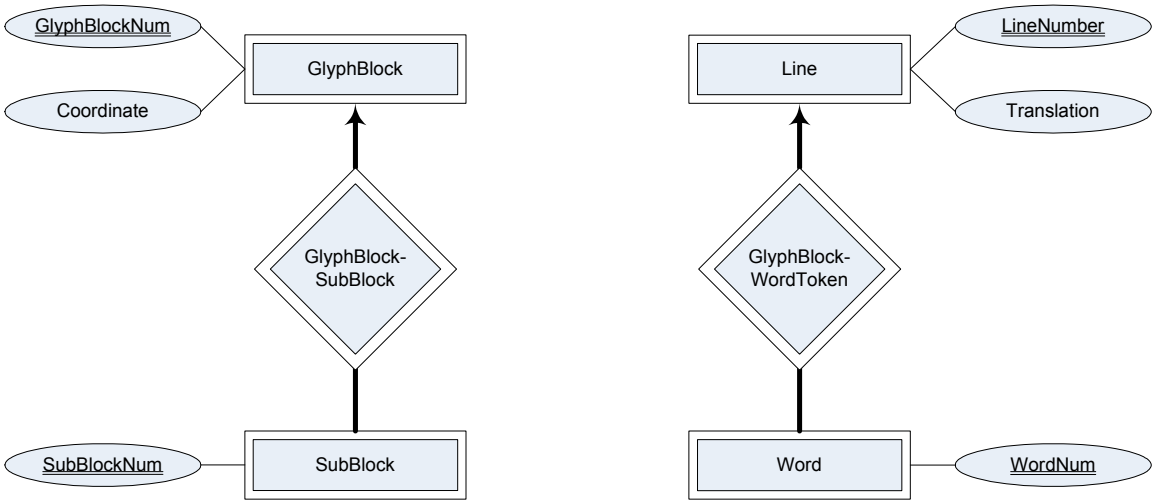


Figure 6-7 SubBlock and Word weak entities.

6.3.8 GlyphToken and MorphemeToken

Each *SubBlock* consists of one or more *GlyphToken*s; each *Word* of multiple *MorphemeTokens*. These two weak entities are indexed by their position within a *SubBlock* or *Word* by their discriminators, *GlyphNum* and *MorphemeNum*, respectively.

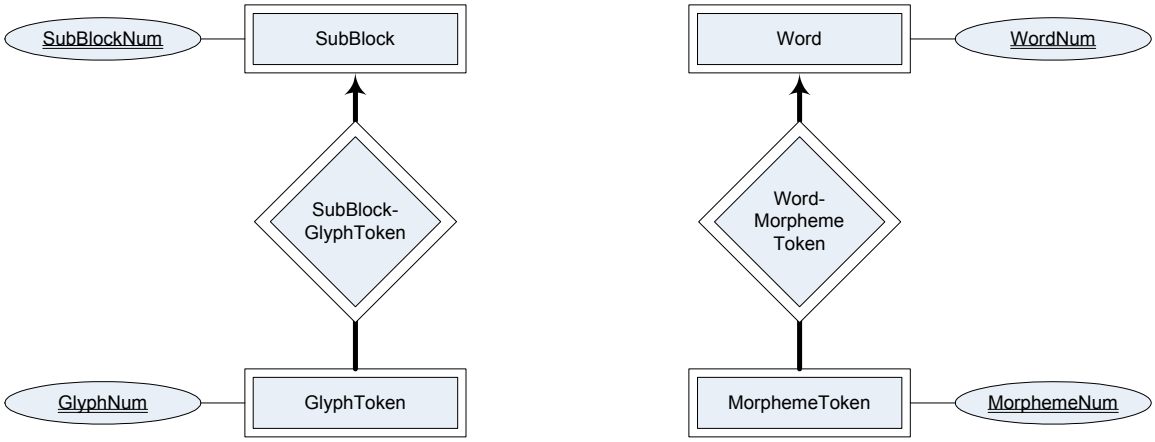


Figure 6-8 GlyphToken and MorphemeToken weak entities.

6.3.9 Glyph and Morpheme

The basic unit of transcription and transliteration are the glyph and the morpheme, respectively. The corresponding entity sets, *Glyph* and *Morpheme,* can be thought of as being analogous to word lists or dictionaries of their respective domain. For the time being, we only track the transcription and spelling of *Glyph*s and *Morpheme*s; it is not difficult to imagine additional properties that could be added to these entities in the future. It should also be noted that special characters introduced in Chapter 4 such as "…" and "?" are entities in both the *Glyph* and *Morpheme* entity sets. In future versions of MayanWiki, these may serve as wildcards when matched against search strings. On the other hand, instead of storing the repeat diacritic as an entity, MayanWiki literally reduplicates the glyph to be repeated in the database. This means that a search for "ka-ka-wa" will find both instances in which it is spelled out literally and instances that employ the diacritic.

Obviously, each *GlyphToken* is instantiated with a single *Glyph* type, although a single *Glyph* type can clearly be instantiated with many tokens; likewise for *MorphemeToken* and *Morpheme*. At times, erosion or other circumstances requires that

the transcriber make an educated guess as to the exact value of a particular glyph token; other times, portions of texts (usually dates) are reconstructable. This requires the inclusion of the Boolean attributes *isGuess*, *isReconstruction*, and *hasReconBound* (short for "has reconstructed boundary") in the relationship that represents the instantiation of a particular *Token* (*GlyphInstance* and *MorphemeInstance*). The function of the former two attributes should be clear from their names while the last attribute serves to disambiguate a rare, but possible, situation exemplified by the following transcriptions: [1]-[?] vs. [1-?]. In the first case, the boundary is known, even though the second glyph is not—perhaps the face of the glyph has eroded beyond recognition, but there is still evidence of a glyph boundary. In the second case, even the boundary must be reconstructed. Because *Glyph* and *Morpheme* are lists that can exist independent of their instantiation, their participation in their respective relationship is partial.

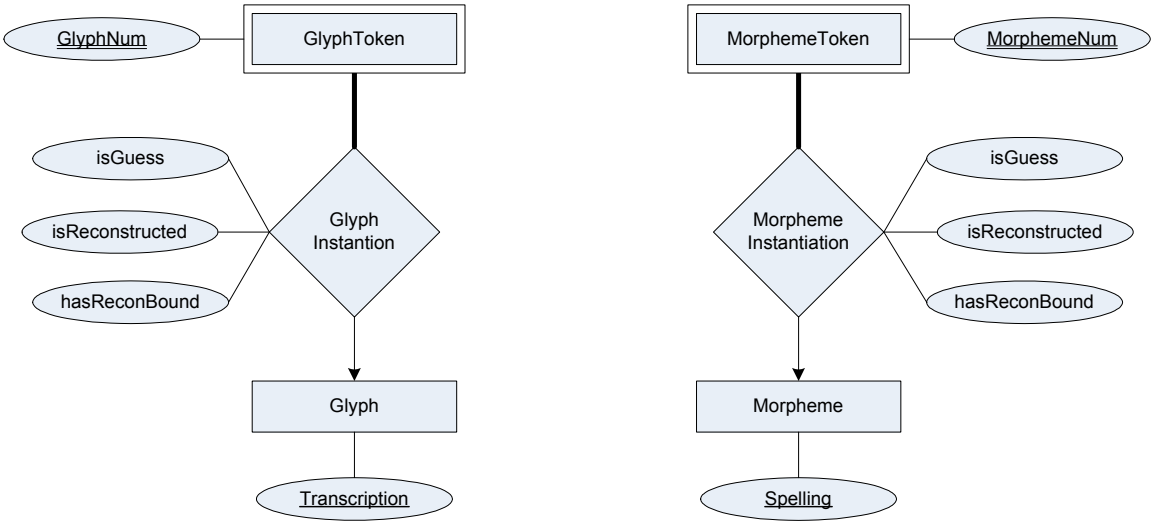


Figure 6-9 Glyph and Morpheme entities.

6.4 Logical Model: Relational Schema

Since E-R schemas are closely related to relational databases, the reduction of an E-R schema to a relational schema is straightforward:

- Each entity becomes a table and its attributes become the table's columns.
- A complex attribute is typically crated with separate columns for each of the "sub"-attributes.
- A many-to-many relationship set also becomes a table. In addition to its own attributes, columns are created for each member of the primary keys of the participating tables. The primary key of this table is the union of the attributes of the primary keys of the entities participating in the relationship.
- A one-to-one relationship set where both the participation of both entities is total merges both entities into a single table, which includes all attributes from both entities and the relationship set. The primary key from either entity can be chosen as the primary key for the resultant table.

- A one-to-one relationship set where the participation of one of the entities is partial can be represented by adding the attributes of the relationship and the primary key of the entity with partial participation to the table for the entity with total participation; the primary key from either entity may be used in the resultant table.
- A many-to-one relationship set where the participation of the entity on the many side is total causes the addition of the attributes of the relationship and the primary key of the entity on the "one" side to the table for the entity on the "many" side.
- A many-to-one relationship set where the participation of the entity on the many side is partial can be treated as if it were total, but the additional column must be "nullable". Otherwise, a separate table is created for the relationship that includes the primary keys from both entities and the attributes from the relationship; the primary key from the entity on the "one" side is used as the primary key.
- A table is created for each weak entity set consisting of all of the attributes of the weak entity and the relationship in addition to the primary key of the owning entity set, which combines with the discriminator to form the primary key of the resultant table.

This algorithm can be applied in a straightforward fashion to most of the entities listed above; the reader is spared the exact details of the conversion of every entity and

relationship set[22]. However, a diagram of the resulting relational model is provided in Appendix B. A few exceptional cases warrant further explanation.

First, the *Page* entity is defined in MediaWiki. In MediaWiki's implementation, the table is assigned a surrogate key, *page_id* and the column name for title is actually *page_title*. The nature of the relationship between *HieroglyphicText* and *Page* allows the primary key from either entity to be used as the primary key for the *HieroglyphicText*. I have opted to use *page_id* (which I rename to *PageID*), in part for reasons of efficiency. Normally, this would result in the creation of a unique key for the *Name* column of *HierolgyphicText*. However, now that *Name* is not the primary key, it is possible to instead define a unique key over both *Site* and *Name*. This allows the name of the text to *not* reduplicate the name of the site. For instance instead of storing "Copan Stela J" as the name of the text and "Copan" as the name of the site, the name simply becomes "Stela J". This is particularly useful for monumental texts, but not as beneficial for vessels and portable objects. Note that it is not possible to represent this feature at the conceptual level.

Moreover, the basic algorithm does not adequately explain how to convert the chain of weak entities from *Passage* to *GlyphToken* and *WordToken* into tables. Applying the algorithm above would produce a *Passage* table with the following columns: *HieroglyphicTextName*, *PassageNum*, and *PassageName*[23] (the primary key is

---

[22] Incidentally, some attributes have been renamed in the conversion to table columns for clarity or to avoid name clashes—most of these should be obvious.

[23] For illustrative purposes, I have chosen to show this example as if the *Name* column of *HieroglyphicText* had been used as the primary key for that table, contrary to what was described in the previous paragraph. In this case, it is necessary to rename the *Name* columns of *HieroglyphicText* and *Passage* so that they don't conflict with one another.

underlined). But how should the next entity, *GlyphBlock* be handled? According to the algorithm, we take the primary key of the "strong entity", which in this case should be interpreted as the primary key of the previously created *Passage* table, i.e. the combination of the true strong entity and all previous weak entities up until that point. Thus, the *GlyphBlock* table has the columns: *HieroglyphicTextName*, *PassageNum*, *GlyphBlockNum*, and *Coordinate*. This algorithm is applied to all of the successive weak entities in the chain. Interestingly, after all of the entities in the database are converted to tables, the table for *GlyphToken* has an intuitive structure: *HieroglyphicTextName*, *PassageNum*, *GlyphBlockNum, SubBlockNum, GlyphNum, Transcription*, *isGuess, isReconstructed,* and *hasReconBound*. This means that each row of the *GlyphToken* table stores the location of a particular glyph instance (including the text, the passage number, and so on), the glyph itself, and some information about the token's status as a guess or reconstruction. This is very much like how someone might build a spreadsheet to store the corpus (as opposed to a word processing document, which would instead contain the syntagmatic layout). Of course, the tables related to transliterations are created in exactly the same fashion.

Only four fields in the database are optional, i.e. "nullable". The algorithm provided above specifies that the column added to *HieroglyphicText* as a foreign key to *Site* is to be "nullable". In the same table, the *JulianDay* field is likewise optional since there are cases when the date is unknown and even too difficult to surmise. Note that when this is the case, the *isApprox* field (which is renamed to *DateIsApprox*) is irrelevant; thus, it needn't be nullable. In previous chapters, it has been explained that the

Name field of *Passage* and the *Translation* field of *Line* are also optional. All other fields in the database are required.

Although tables that contain the primary keys of other entities are created with foreign key constraints when converting to a logical design, it is still necessary to specify the behavior of these constraints when the key of referent table is updated or deleted in the logical design phase. In MayanWiki, all updates are automatically cascaded, although updates to primary keys should be rare. On the other hand, deletes are sometimes restricted and other times cascaded. In tables that refer to *Site*, *Medium*, *Glyph*, or *Morpheme*, deletes are disallowed. As a consequence, no site, medium, glyph, or morpheme can be removed from the database if it is "used" by any text. On the other hand, when a user requests that a page be deleted from MayanWiki, it is convenient to automatically remove the corresponding *HieroglyphicText*, and all of its data; otherwise the MayanWiki extensions would be required to perform these deletes. Thus, these particular constraints are cascaded. One final change is made to the logical design: a surrogate key is added to the *GlyphToken* table called *GlyphInstanceID;* although not as necessary, we similarly add a *MorphInstanceID* to the *MorphemeToken* table for consistency. To motivate this decision, consider the following example: a researcher desires to find all occurrences of **u-ts'i-bi** 'his writing, painting'. At first, one might assume that these three glyphs should reside within the same glyph block. However, as mentioned in Chapter 2, there are times when this is not the case—when what would otherwise be a simple "unit" such as this inflected verb can cross the glyph block boundary. Thus, the query must find all instances where the *GlyphNum* of **u** is one less than the *GlyphNum* of **ts'i** in the same *GlyphBlock*, as well as the case where **u** is the last

glyph in its block and **ts'i** is the first glyph in the next block. Needless to say, this complicates queries appreciably. If it were possible to guarantee through some means that the *GlyphInstanceID*s for every glyph are assigned in increasing order, regardless of passage, block, and sub-block boundaries, then queries would be greatly simplified[24].

The experienced database administrator may question the lack of surrogate keys for the majority of tables. After all, many—if not most—administrators make insistent use of surrogate keys, assigning them to nearly every table created. Although surrogate keys can be beneficial in many situations, they are also considered a premature optimization (Surrogate Key, 2007). Before automatically assigning surrogate keys to tables, it is best to consider the most common usages of the database and to use profiling techniques in order to determine what optimizations will make the biggest impact. In the case of MayanWiki's database, the only type of query currently being done on the database is the searches; the data displayed on the wiki are stored separately by MediaWiki. Adding surrogate keys to each of the tables created from the weak entities will result in some reduction of redundancy. To see this, imagine the spreadsheet described above. If there are thirty glyphs per text on average, the *HieroglyphicTextName* will be repeated thirty times in a row! However, it is important to point out that this type of redundancy (which results from multi-column primary keys) is not covered under any of the normal forms. Furthermore, if all tables had surrogate keys, then *each* token in a query would require a join between *Passage*, *GlyphBlock*, *SubBlock*, *GlyphToken*, and *Glyph* (or the analogous columns for transliterations); without surrogate keys, none of

[24] This is not possible, or at least feasible, using mechanisms provided by the database alone. This issue is addressed further in the next section.

these joins are necessary. The lack of joins could have modest performance benefits, although future work should scrutinize this issue further, especially in the presence of larger amounts of data. In addition, requiring fewer joins allows for larger strings of tokens to be searched before the (vendor-dependent) maximum number of tables allowed per query is reached. The important point is that performance issues are often unknowable until tested in production environments with real data and most optimizations should be deferred until performance can be profiled.

Other tables are affected by the lack of surrogate keys as well, specifically, *Site*, *Medium*, *Glyph*, and *Word*. Because these entities currently lack other attributes, there is no anticipated benefit to using a surrogate key. In the case of searches done on these values, all of which are strings, efficiency related to comparisons is not an issue because indexes are created in the foreign key columns of the referring tables. If anything, efficiency should be slightly increased *without* surrogate keys because no join is necessary when querying the referring tables. It is true, however, that a cascaded update must do more work without a surrogate key. Fortunately, no updates are currently being done in MayanWiki. Instead edits are done by deleting previously entries and inserting the updated data afresh. Again, it is important to note that optimizations should be carefully thought out based on what is known about the database a priori rather than implemented blindly out of habit or even based on *perceived* enhancements rather than those profiled from real use of the database.

6.5 Physical Design

The first decision in the physical design of the database is to choose a database management system (DBMS). MySQL was selected for MayanWiki, mostly because MediaWiki works best with this DBMS. Another consideration of the physical design concerns the choice of the MySQL-specific engine type for each of the tables. All tables are implemented with the InnoDB engine in order to take advantage of transactions and foreign key constraints—a must in order to maintain consistent data. A consequence of using InnoDB tables is that the clustered index must be the primary key and all indexes must be B-trees (for more information on B-trees, see Silberschatz, Korth, & Sudrashan, 2002). Furthermore, MySQL automatically creates indexes for all of the foreign keys and unique constraints mentioned in the previous section.

The next consideration for the physical design of the database are the (sometimes MySQL-specific) data types chosen for each of the columns from the logical design. The string fields in the database have been implemented using a VARCHAR column type in the latin1 character set. Unicode is unnecessary since the only non-ASCII characters are some accents in certain site names that are included in the latin1 character set. The main reason for choosing VARCHAR over CHAR is that VARCHARs occupy less space and allow for more data to be held in the index. This in turn reduces the amount of I/O needed for searches and hence will result in faster searches. Since the database does not require updates (recall that changes in the database are accomplished by deleting the old information and inserting the new data), there is no performance degradation associated with updating VARCHAR entries. Choosing an appropriate length is difficult, but a maximum length of 50 appears to be sufficient for all fields except for two. The

Translation field needs to accommodate strings the length of a typical clause; 1024 characters appears to be sufficient[25]. The second exception is the *Coordinate* field of *GlyphBlock*; most coordinates will be two or three characters in length, but there are exceptions, most notably blocks that span more than one row or column which could occupy more. Thus, the chosen length is 15 characters. The default collation for character fields in MySQL is case insensitive which is desirable for all textual fields except for *Transcription* of the *Glyph* entity. Here, conflicts can arise if case is ignored. For instance, the glyph **u** when used as the 3rd person ergative pronoun is often transcribed as a logogram, i.e. **U**. However, there are other, less common situations in which **u** can occur and it is desirable that these cases be separate. Therefore, this field is specified with the binary collation, which is case-sensitive and extremely efficient.

The numeric fields are less complicated. The *JulianDay* field is best implemented using an unsigned, 32-bit MEDIUMINT, which allows for dates in the range of 4,713 BC to 41,222 AD—clearly within the range of any possible inscription. The discriminator column of the tables for each of the weak entities is implemented using an unsigned, 8-bit TINYINT which allows for up to 255 passages per text, blocks or lines per passage, sub-blocks/words per glyph block/line, or glyphs/morphemes per sub-block/word; in each case, this is more than sufficient. Finally, the *GlyphInstanceID* and *MorphInstanceID* columns should fit in a 32-bit INT[26].

---

[25] For reference, the average number of characters per sentence in this paragraph is under 150.

[26] Because the size of the number needed for this value is related to the number of modifications made to texts in the wiki, it is difficult to estimate how big of a number is reasonable. This number should be sufficient for at least the first few years, if not forever. If it ever does become a problem, the keys can be reordered to fill in missing values, or the size of the number can be extended at that point (64-bit machines should be the norm by then, in any case).

A special table is needed to facilitate the mechanism described in the previous section concerning the desirable property that subsequent glyphs and morphemes in a text are guaranteed to be assigned an identifier that is exactly one greater than the previous glyph or morpheme in the text. Under normal circumstances, the database would allow the data for more than one text to be inserted concurrently, which would cause the identifiers to be interleaved between the two texts rather than creating sequential identifiers within each text. One option would be to use an auto-increment field and lock the table before any bulk insertion of data. However, this can cause unnecessary delays when two texts are being simultaneously added or modified in the database. Instead, MayanWiki employs a counter table with one row columns that stores the current value of the identifier for both the *GlyphInstanceID* and the *MorphInstanceID* fields in the database. Before any data are inserted, this *counter* table is locked, the next available value for the appropriate identifier is read and subsequently updated based on the number of glyphs or morphemes to be inserted, and then the counter table is unlocked. This is efficient because the counter table will be locked for much less time than if the entire *GlyphToken* or *MorphemeToken* table were locked while data were inserted. This is particularly true because the *Counters* table is stored completely in memory (using the memory engine). Nonetheless, the identifiers are still guaranteed to be sequential within any given text.

Appendix C contains the data definition statements for the physical model.

6.6 Summary

This chapter has detailed the conceptual, logical, and physical design of MayanWiki's database by justifying decisions based on the assumptions that were made about the data according to how it would be used in MayanWiki; care was also taken to avoid making premature optimizations. Because all of the assumptions inherent to the final design are made explicit in this chapter, the reader is free to evaluate the assumptions, and easily adapt them to other circumstances. In addition, the information from all three levels of design should make the function and role of each table and each column in the final database clearer for those trying to understand the structure of MayanWiki's database. This is particularly useful for those trying to adapt the concepts presented here to other scripts or languages.

By following this somewhat rigorous design process, earlier ideas about the database have been significantly refined and improved, resulting in a much simpler database that is both efficient and easy to understand. The end result is a highly normalized database in Boyce Codd-normal form[27] (and, trivially, in fourth and fifth normal forms)[28], that performs well in practice. Although this is reason enough to have followed proper design principles, this chapter has presented additional reasons for carefully following the process.

---

[27] Some may consider that the separation of logograms from syllables by case within the same column is a violation of first normal form. However, given the ambiguous nature of the definition of a "domain", this claim is difficult to support. For instance, even a character string is arguably divisible into smaller parts and one column should be made for each character. That said, future work could consider formally separating the different types of glyphs.

[28] MayanWiki's database does not contain any relations that could violate fourth or fifth normal form.

As users contribute more content to MayanWiki, it may become necessary to re-evaluate some of the decisions made in this chapter in order to optimize performance. Although the corpus of all hieroglyphic texts is relatively small, it may be necessary to denormalize the database in order to increase query performance. For most studies of language, it is necessary to study strings consisting of more than one lexeme (or glyph). That is, typical queries are syntactic in nature, but the data are stored in a more paradigmatic fashion. Typically, the largest degradation in performance is due to long query strings that require the token table (*GlyphToken* or *MorphemeToken*) to be joined with itself once for every token in the query in order to convert from the paradigmatic storage of the data to a more syntagmatic view. Since the join is based on a function of columns (the position of the one token must immediately precede that of the next token of the query) rather than an indexed column, this type of query is expensive and hard to optimize. However, the use of a functional index or its equivalent could mitigate this problem. Another potential solution is to add a table that holds windows of sequential text, that is, a table which has one column for each sequential glyph or morpheme within a given window size; each "window" (i.e. row in the table) overlaps the previous one by exactly one glyph/morpheme. Unfortunately, this introduces a large amount of redundancy: each token is repeated *n* times, where *n* is the chosen size of the window! However, all of the problems associated with redundant data—except of course, size—are eliminated with the implementation of the table as a materialized view. Since the corpus of the hieroglyphs is relatively small and disk space is cheap, the cost in terms of space is negligible. However, storing the data more syntagmatically could considerably enhance the performance of queries. Particularly, the more syntagmatic the query (i.e. the

longer the search string), the more appropriate it is to use a syntagmatic table and hence, the greater the benefits.

7 Wiki Features

All of the features of MayanWiki explained up to this point have been implemented as extensions to MediaWiki, the open-source software created for and used by Wikipedia. In fact, MayanWiki is actually nothing more than an implementation of MediaWiki that includes extensions to parse hieroglyphic texts, add them to the database, and perform advanced searches. One advantage to this approach is that MayanWiki immediately inherits all of the functionality of the MediaWiki software. Furthermore, because no changes were made to the source code for MediaWiki, the MayanWiki extensions should continue to run on future versions of MediaWiki that maintain backwards-compatibility with the extensions for the latest version available from the subversion repository (version 1.11), thus inheriting any new functionality in these versions as well.

Due to its high profile as the software for Wikipedia, MediaWiki is very mature, feature-filled, and customizable. It is outside the scope of this project to detail every feature of MediaWiki and how such features might be used to aid research in Mayan studies. However, there are several features that deserve special notice in this chapter.

7.1 Built-in Search

The reader may have noticed that, despite the fact that users can add English translations to hieroglyphic texts, these translations cannot be searched by the advanced search engine

described in Chapter 5. The reason for this is that translations do not, and should not, be analyzed in the same way that the source language data is. In most cases, a "Google-like" search—where a list of texts containing the English search term is returned—is sufficient. In fact, MediaWiki already incorporates such a search engine, which is one reason that this functionality was not reduplicated in MayanWiki. However, MediaWiki's built-in search engine ignores words shorter than four letters by default[29], effectively discarding most uninteresting English words, but also disposing of most interesting Mayan glyphs and lexemes (recall that the canonical root is of the form CVC). Nevertheless, the search engine found in the "toolbox" on the side panel of every page is satisfactory for finding keywords in translations and even keywords in the transcriptions and transliterations that are long enough. The search engine is also effective for finding pages about a particular text or site as well as information contained in other articles.

7.2 Hierarchical Categories

The concept of a category is familiar to most Wikipedia users. Articles are typically assigned several categories that reflect the content of the page—not unlike "keywords" assigned articles in online academic journals. This effectively groups similar articles together and can be helpful if a user is browsing for articles on a particular topic and would like to read several pages from this topic.

One of the novel features of MediaWiki is that it treats categories as hierarchical. That is, categories can themselves have categories. Thus, the category "Lexicography"

[29] This is actually a MySQL setting that can be altered. However, it is not recommended that this be changed as it could create very large indexes.

might be categorized as belonging to the "Linguistics" category which in turn might belong to the "Social Sciences" category and so on. These hierarchical categories can be exploited in MayanWiki for several purposes. For instance, one might group together all of the texts from a particular temple at a particular site, such as those found at Palenque Temple XIX. Each of the temples at a site might in turn be grouped together by that site, several sites by geographic region, etc. Another practice could include categorizing articles by date, particularly certain significant periods within the long count, so that all texts written in the same time period are grouped together. MayanWiki automatically adds some preliminary categories such as these.

Another attractive use of categories is to build a catalog that correlates the numbering system of previous catalogs and ultimately links them to their phonetic value. To illustrate, suppose a page is added to the wiki treating the **u** glyph. Such a page might include information regarding the identification of each of the allographs, including references to the published literature and a gallery of drawings. This page could then be categorized with the Thompson numbers (and/or other system) of each of the variants of the **u** glyph, e.g. T0001, T0738c, etc. as well as HE6, AA4, and so on (according to Macri's classification). The Thompson numbers could then be further categorized as "main signs", "affixes", etc. and Macri's signs can be grouped into their categories (e.g. animals, hands, etc.) and "subcategories" (e.g. monkey, fish, etc.). This could be done for *any* classification system and the result would be a browsable, correlated catalog of catalogs.

This powerful feature could be employed in many more ways than those described here. This feature will help create new organizational systems and improve the usability of the content of MayanWiki.

7.3 Public discussion

Each and every page created in MayanWiki has an associated "Discussion" page. The purpose of this page is to allow users to discuss changes to an article that could be seen as controversial or major. Rather than editing the page itself, it is sometimes wisest to discuss and resolves issues in a way that does not disrupt the current content of the page. This is particularly beneficial to pages containing hieroglyphic texts, since users can work out issues such as "spelling" rules, translations, etc. These discussions should in fact increase communication among researchers—especially those of different schools of thought. The current process for resolving differences is painstakingly slow (it took almost thirty years for the field to universally accept the phonetic nature of the script!). One reason for such delays is the large amount of time required for ideas to see the light of publication. Although MayanWiki is not a substitute for published material, core arguments that will eventually be published can receive rapid feedback thereby improving the quality of published material and the content of the wiki. In essence, MayanWiki can help increase the speed of the flow of information in the field and close the "knowledge gap" (see Lih, 2004).

Unfortunately, one of the major hindrances in the field of Mayan studies is the preponderance of what Stuart (2005b) terms "grey material": unpublished manuscripts, impermanent email exchanges, etc. Another function of discussion pages is to replace this

"grey material" with publicly available discussions. In addition, otherwise unpublishable material could be added to the wiki so that it is available to a much wider audience than is currently the case. This increase in the availability of information should foment new ideas and speed up advancement in the field.

7.4 User Pages

Every registered user in MediaWiki has his or her own user space. This user space includes a main page which can be used for various purposes. For example, it could consist of information about the user, including interests and possibly a curriculum vita. It is also supposed to help organize information regarding the articles that the user is editing (Wikipedia:User Page, 2007).

Users can create as many sub-pages as they need. One possible use for a sub-page is to propose alternative transcriptions, transliterations, or translations to the generally accepted versions that may be too lengthy to include on a discussion page; the discussion page will simply include a link to the sub-page. Note, however, that these transcriptions are *not* added to the corpus as are the generally-accepted counterparts (that is, they do not show up in the results obtained through the advanced linguistic search described in Chapter 5)[30]. If everybody were allowed to add their own version of transcription to the corpus, then there would be little reason to carefully and thoroughly discuss differences in opinion on the discussion page. The net result would be that there would be little

[30] The reader is reminded that in a wiki, anybody can add new transcriptions or edit them. There is no appointed editor or owner of any of the content and changes occur immediately; this is true even of the "generally-accepted" transcriptions. The community as a whole is responsible for the contents of the database (also see the section on vandalism protection below).

consensus and little progress. Since the *raison d'être* of the wiki part of MayanWiki is to make it possible for the corpus to eventually converge to the truth, this situation undermines the purpose of the wiki.

7.5 Image Pages

MediaWiki allows images and other content (such as sound files, documents, etc.) to be uploaded to the wiki. Each image is treated as its own page and can therefore contain any amount of useful, searchable metadata. Categories can also be assigned to images. Those that are familiar with the databases contained on FAMSI's web site will recognize the potential that a resource like this could have: imagine a searchable database with everybody's personal drawings, not just Linda Schele's and John Montgomery's. With cooperation from FAMSI, it may be possible to seed MayanWiki with their valuable resources. MayanWiki can potentially grow to hold every useful line drawing and photograph in the public domain (perhaps one day including Maudslay's (1889-1902) and Maler's (1901) pioneering work).

Researchers may initially be reluctant to upload their personal collection of line drawings and photographs because of the open nature of the wiki. With good reason, calligraphers, artists, and photographers may wish to retain the rights to their creative work. Fortunately, artistic works such as these can explicitly be licensed under the Creative Commons licensing system (other licenses are possible). Typically, a Creative Commons license allows others to use the work only when including the original artists' name, although other terms are possible. In general, this increases awareness of an artist and is typically mutually beneficial to the public and the artist.

7.6 Other Articles

As has been mentioned on several occasions, MayanWiki is not limited to content containing hieroglyphic data; articles can be created on any topic, although writers are encouraged to stick to the general theme of Mayan studies. MayanWiki will eventually contain a full syllabary and dictionary, with pages for each of the syllables and logograms that explain their origin and the history of their decipherment. In due course, each site will have its own page with a brief description of its location, origin, history, and dynastic succession. Such pages could also hold maps of the site, a gallery of photographs, the emblem glyph for that site, and links to all the texts coming from the site. Each temple or area within a site will likely also have its own page with similar content. It is not unreasonable to assume that articles will be added related to the prominent (usually deceased) researchers in the field like Yuri Knorosov and Eric Thompson. However, it is important to note that some content is more appropriate in other resources, such as Wikipedia, and care should be taken to avoid duplicate information.

7.7 Vandalism Protection

One weakness of using a wiki to host important linguistic data is the potential for vandalism: changes to the content made with the intent to ruin the data. The main defense against vandalism is to make it much harder to vandalize a page than to undo vandalism (Viegas, Wattenberg, & Dave, 2004). In MediaWiki, this is accomplished in several ways. First, every page keeps a complete history of every change made to the page; from this page, changes can be reverted to a previous version with only two clicks of the mouse. Furthermore, users can opt to “watch” any page. A list of changes to pages on a

user's watch list can be accessed at anytime. In fact, users can optionally be notified by email whenever a page they are watching changes and users can choose to automatically add pages they edit to their watch list. These features make it so that vandalism can be removed quickly and easily. In rare cases, it may also be necessary to "protect" a page. Depending on the level of protection, protected pages are not allowed to be edited by anonymous users or non-administrators[31].

Another related issue arises from the distributed, asynchronous nature of the internet. It is entirely possible for more than one person to edit a page at the same time and conflicting edits may arise when the second person attempts to save their edits. In this case, the second user is shown an editable copy of the page submitted by the first person, a list of differences, and their own version before the conflict. This feature is necessary to avoid one person inadvertently overwriting the work of another.

7.8 Namespaces

Namespaces are a mechanism for keeping unrelated content separate from each other, *at a project level*. The main namespace is the default, most common namespace where all articles belong. User pages, on the other hand, are kept in their own separate namespace. Using MediaWiki's default search engine, it is possible to limit searches to specific namespaces. For instance, it would be possible to search for words in English translations found on User's private pages only, or, conversely, only in the mainstream articles.

[31] The exact system of roles and privileges is of low importance in a wiki and thus is not treated in this special project. The main purpose of the administrative role is to prevent vandalism and is not to delegate ownership of texts or other content to certain users. The exact system used to grant administrative privileges will depend on the final resting place of MayanWiki and hence is outside the scope of this special project.

Another use is to limit searches to images. This could be useful to find images with metadata that could otherwise bring up a large number of hits in the main namespace. If a significant amount of content not containing transcriptions eventually resides in MayanWiki, it may become desirable to move all pages containing texts to a separate namespace so that they may be searched separately from the rest of the wiki. Certainly other uses of namespaces are possible as well.

7.9 Stubs

"Stubs" are technically not a feature of MediaWiki but rather a technique commonly used in Wikipedia. "Stubs" are articles created as placeholders for real content. They typically contain very little content—often a few sentences about the subject that is to be represented on the page along with an invitation to expand the article. Stubs are often created through automatic means using external data sources in order to add a large number of articles to the wiki at the same time. This technique might be used to automatically create a page for every known site and every known text based on data contained in some other database. Even though these pages will not initially contain much information, it is advantageous that they at least exist in the database, in part because it is less intimidating to edit an existing page than to create a new one.

7.10 Summary

The maturity of MediaWiki brings with it a host of features that can enhance MayanWiki as a central resource of not only linguistic data from hieroglyphic texts, but a repository of line drawings and photographs as well as information about all things Mayan. It is

hoped that the features described in this chapter can help make MayanWiki an important resource for Mayan studies that hosts critical data, fosters collaboration, and encourages contributions.

8 Conclusion

Since the large-scale decipherment of the glyphs in the late 1980s, focus in the study of Mayan hieroglyphic texts has shifted to trying to fully understand the language of the glyphs. Unfortunately, the data needed to carry out proper studies are scattered across many different sources and, consequently, much work has been left to the memory of the modern epigrapher. While there have certainly been many important and significant advances in these circumstances, the advantages of an electronic repository of hieroglyphic texts are obvious and many stand to benefit from such a corpus. This special project has presented MayanWiki as a viable solution to this problem. MayanWiki is unique because it is based on corpus linguistic principles, stores linguistic and glyphic data in a relational database, and relies on user contributions so that the texts can reflect current research and eventually converge to the truth through consensus. A further benefit of this work has been to propose an unambiguous standardization of the syntax of transcriptions and transliterations. As a result, MayanWiki will provide invaluable data to Mayan linguists and epigraphers. In addition, students, archeologists, anthropologists, historians, hobbyists, and even scholars will be able to do research and access valuable data without the need to commit the entire corpus to memory.

Despite these benefits, epigraphers can be finicky about their source of data and not every repository will fulfill the needs of epigraphers in such a way that it can

sufficiently aid research. The requirements of a successful corpus of hieroglyphic texts that is to aid research are threefold. First, the data must be as comprehensive as possible and be stored in a central location that is easily accessible and publicly available. Second, the search engine must allow the data to be readily available and manipulable in ways that are relevant to linguistic study. In particular, it should be possible to study phraseology, frequency, and collocations. Lastly, the community must be able to contribute new data and modify existing data. This is the most efficient and effective way to ensure that the data converge to the truth as quickly as possible.

Each of these goals presents unique challenges and up until this point, no existing or proposed database fulfills all three requirements; for this reason MayanWiki was created. The type of texts that MayanWiki stores are phonetic transcriptions along with their accompanying transliterations and English translations. All of the data contained in MayanWiki are publicly available over the internet and MayanWiki's search engine allows for powerful, linguistic-oriented searches to be performed using concordances, frequency tables, and collocational statistics. This powerful search tool can help researchers form new hypotheses and support them with relevant data. Since MayanWiki is in fact a wiki, users can contribute new texts and modify existing ones based on a thorough discussion of evidence in favor of the change. Most of these needs are accommodated by the carefully designed relational database backend.

Even though the three aforementioned requirements are necessary for success, they do not automatically guarantee it. There are three strategies that can further improve the chances of MayanWiki's success. First, the database must be populated as quickly and extensively as possible. Next, a policy of "conservative transcriptions, innovative

explanations" must be firmly established. And finally, users must be encouraged to contribute as often as possible. Each of these strategies is discussed in turn.

The major weakness of MayanWiki in its current state is that it presently contains only a handful of texts. This of course violates the first requirement of a useful database since the database is far from comprehensive. However, because users are allowed—even encouraged—to submit texts, 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. In fact, if any of the more comprehensive databases were willing to contribute their data (with proper attribution, of course), MayanWiki could be populated with data for most texts overnight. If even only a handful of knowledgeable students or researchers were to continuously contribute data, much progress will still be made. It may even be possible for students to add content as part of the learning experience in an introductory course to hieroglyphic writing. Because it will eventually be a comprehensive corpus, it fulfills the first requirement for usefulness. However, the process of populating the database is the first key to success.

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. However, if users are encouraged to transcribe texts as conservatively as possible—that is, based on accepted, published decipherments, etc.—then 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. 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[32]. This can be accomplished primarily through feedback on discussion pages and reverting changes deemed non-conservative.

Finally, even when MayanWiki is comprehensive and enjoys a relatively large and diverse user base, it is important that all users contribute as often as possible. Otherwise, readings will stay relatively the same and convergence to the truth will be somewhat retarded. This of course is difficult to enforce. However, by employing "stubs" that solicit content and by constantly reminding users that their help is needed and to make corrections where needed (e.g. as an "advertisement" at the top of search results), users can be encouraged to participate as much as possible and the quality of the data available on MayanWiki will increase.

---

[32] The main contributors to the project are those who actively submit new content or edit old content. It is assumed that these contributors will be familiar with the policies set forth by MayanWiki.

Other enhancements to MayanWiki could increase its effectiveness. One area of focus should be on further simplifying the data entry and editing process. Although the process is fairly simple and straightforward, some improvements can still be made. For instance, a what-you-see-is-what-you-mean editor would be less intimidating than requiring that users know wikitext (note, however, that no wikitext is needed for modifying or adding transcriptions). The edit page could be altered so that it is more amenable to aligning transcriptions, transliterations, and translations such that each is visible while editing the others. Finally, the display of the data could be made more appealing.

Certain processes can be automated to various degrees in order to simplify the data entry process. For instance, optical character recognition techniques could be used to initially transcribe a text and link each glyph of the text to a region of the scanned image. The edit distance algorithm (Wagner & Fischer, 1974) could be used to automatically map morphemes in the transliteration to their source glyphs in the transcription. Natural language processing techniques could further be used to suggest readings for obliterated or unknown glyphs, to cluster similar texts together, to automatically find topics in texts, and to locate and annotate proper names, among other things[33]. Spelling rules could also be inferred from the data that allow transliterations to automatically (or semi-automatically) be created from the transcriptions. In all of these applications, the output from the machine would need to be reviewed by a human, and a wiki is the perfect medium for this. For instance, after submitting a line drawing, the machine could attempt

[33] The quality of these automatic processes depends in large part on the amount of data available. In terms of linguistic corpora, the Mayan corpus is relatively small. Even so, correcting the automatic output of a computer is often much easier than starting from scratch.

to automatically transcribe it and allow the user to make changes as necessary. Then, after transcribing a text, the user would press a button that automatically creates a transliteration with a mapping from each glyph block to each clause of the transliteration. The submitter will correct the output for any mistakes and then save the text. Subsequent users, for example, those who have that particular page on their watch list, will further review the output and make corrections as necessary.

The search engine could also be enhanced. For one, keywords could be introduced as "syntactic sugar" for common searches, e.g. "C" for consonants and "V" for vowels. Options can also be added that allow guesses and unknown glyphs/lexemes in texts to act as wildcards in searches. More significantly, perhaps, would be the ability to compare frequency and collocational information for two different time periods or regions side-by-side in order to make a more direct comparison than is currently possible.

One of the major advantages to using relational databases to store linguistic data is that an essentially unlimited amount of annotation could be added (see Davies, 2005). This means that information regarding part-of-speech, semantic roles, lemma, etc. can be added for each token. It is also possible to annotate each type, for instance, to add etymological information or root type. In fact, the database could be expanded to hold content of the most important dictionaries from various languages and make them searchable in convenient ways. Even though these annotations could be useful, it is important to note that each of these require additional input from a human annotator. If these annotations were to complicate the data entry process or interfere with the interpretation of texts in any way, they would discourage use of the system. The lack of structure inherent in wikis and web pages in general, makes it easy to add and edit

content and encourages participation. On the other hand, annotations add structure and make contributions more difficult. Therefore a balance should be maintained between imposing too much structure on data through annotations and increasing the ability to find new and interesting patterns using this additional structure.

One result of this special project has been to carefully layout the design of the database that underlies MayanWiki. As has been mentioned several times, this schema can be adapted to hold linguistic data from similar languages, perhaps as a means of preservation or otherwise intended for linguistic study. The more a language tends towards polysynthesis, the more appropriate it is to store data morpheme-by-morpheme and ignore word tokens altogether. Languages that are more isolating are best handled with words as the most basic unit of the database.

Despite the fact that most of the glyphs have been deciphered, much mystery still surrounds the language of the hieroglyphs and the people that wrote them. MayanWiki represents a significant advancement in the field of Mayan linguistic epigraphy that can help uncover many of these mysteries faster than ever before. But ultimately, MayanWiki is nothing more than a tool to be used by the next Champollion.

9 References

Abercrombie, D. (1965). *Studies in phonetics and linguistics.* London: Oxford University Press.

Alvarado, R. C. (1994). *The Mayan Epigraphic Database Project*. Retrieved August 2007, from http://www3.iath.virginia.edu/med/

Bricker, V. R. (1986). *A grammar of Mayan hieroglyphs.* New Orleans: The Middle American Research Institute.

Chen, P. P. (1976). The entity-relationship model: Toward a unified view of data. *ACM Transactions on Database Systems , 1* (1), 9-36.

Christ, O. (1994). *The IMS Corpus Workbench technical manual.* Stuttgart: Institüt für maschinelle Sprachverarbeitung, Universität Stuttgart.

Christ, O., & Schulze, B. M. (1995). Ein flexibles und modulares anfragesystem für textcorpora. *Tagungsbericht des Arbeitstreffen Lexikon + Text.* Niemeyer, Tübingen.

Church, K. W., & Gale, W. A. (1991). Concordances for parallel text. *Proceedings of the Seventh Annual Conference of the UW Centre for the New OED and Text Research*, (pp. 40-62). Oxford.

Coe, M. D. (1999). *Breaking the Maya code.* New York: Thames & Hudson.

Coe, M. D., & Van Stone, M. (2005). *Reading the Maya glyphs* (2nd Edition ed.). New York: Thames & Hudson.

Date, C. (2005). *Database in depth.* O'Reilly.

Davies, M. (in press). Relational databases as a robust architecture for the analysis of word frequency. In D. Archer (Ed.), *AHRC ICT Methods Network: Expert Seminar on Linguistics: Word Frequency and Keyword Extraction.* Brookfield, VT: Ashgate Publishing Company.

Davies, M. (2005). The advantage of using relational databases for large corpora: Speed, advanced queries, and unlimited annotation. *International Journal of Corpus Linguistics, 10* (3), 307-334.

Evreinov, E. V., Kosarev, Y. G., & Ustinov, V. A. (1961). *Primenenie èlektronnyx vyčislitel'nyx mašin v issledovanii pis'mennosti drevnix majja [The application of computers for the decipherment of the ancient Maya script]*. Novosibirsk: Akademija Nauk SSR - Sibirskoje otdelenie.

FAMSI. (n.d.). *The Madrid Codex.* Retrieved August 7, 2007, from FAMSI: http://www.famsi.org/mayawriting/codices/pdf/madrid_rosny_bb.pdf

Giles, J. (2005). Internet encyclopaedias go head to head. *Nature , 438* (7070), 900-901.

Graham, I. (1975-2006). *Corpus of Maya hieroglyphic inscriptions.* Cambridge, MA: Peabody Museum of Archaeology and Ethnology, Harvard University.

Houston, S., Robertson, J., & Stuart, D. (2001). Quality and quantity in glyphic nouns and adjectives. *Research Reports on Ancient Maya Writing 47*. Washington, D.C.: Center for Maya Research.

Houston, S., Robertson, J., & Stuart, D. (2000). The language of Classic Maya inscriptions. *Current Anthropology, 41*, 321-356.

Houston, S., Stuart, D., & Robertson, J. (2004). Disharmony in Maya hieroglyphic writing: Linguistic change and continuity in Classic society. In S. Wichmann, *The linguistics of Maya writing* (pp. 83-101). Salt Lake City: The University of Utah Press.

Hruby, Z. X., & Child, M. B. (2004). Chontal linguistic influence in ancient Maya writing. In S. Wichmann, *The linguistics of Maya writing* (pp. 13-26). Salt Lake City: The University of Utah Press.

Hruby, Z. X., & Robertson, J. S. (2001). Evidence for language change in ancient Maya writing: A case study of the verb tzutz. *Research Reports on Ancient Maya Writing 50* . Washington, D.C.: Center for Maya Research.

Hunston, S. (2002). *Corpora in applied linguistics.* Cambridge, UK: Cambridge University Press.

Justeson, J. S., & Campbell, L. (Eds.). (1984). Phoneticism in Mayan hieroglyphic writing. *Institute for Mesoamerican Studies Publication no. 9*. Albany: State University of New York.

Kerr, J., & Kerr, B. (1989-2001). *The Maya Vase Book: A Corpus of Rollout Photographs of Maya Vases* (Vols. I-VI). New York: Kerr Associates.

Knorosov, Y. V. (1958). The problem of the study of the Maya hieroglyphic writing. *American Antiquity, 23* (3), 248-291.

Lacadena, A., & Wichmann, S. (2004). On the representation of the glottal stop in Maya writing. In S. Wichmann, *The linguistics of Maya writing* (pp. 103-162). Salt Lake City: The University of Utah Press.

Law, D. A. (2006). *A grammatical description of the Early Classic Maya hieroglyphic inscriptions (Master's Thesis, Brigham Young University, 2006)*. Retrieved from http://contentdm.lib.byu.edu/ETD/image/etd1264.pdf

Lih, A. (2004, April 16-17). Wikipedia as participatory journalism: Reliable Sources? Metrics for evaluating collaborative media as a news resource. *5th International Symposium on Online Journalism* . University of Texas at Austin. Retrieved August 2007 from http://online.journalism.utexas.edu/2004/papers/Edwards.pdf.

MacLeod, B. (1987). An epigrapher's annotated index to Cholan and Yucatecan verb morphology. *University of Missouri Monographs in Anthropology no. 9*. Columbia: University of Missouri.

MacLeod, B. (1984). Cholan and Yucatecan morphology and glyphic verbal affixes in the inscriptions. *Phoneticism in Mayan hieroglyphic writing. Institute for Mesoamerican Studies Publication no. 9*, 233-63. (J. S. Justeson, & L. Campbell, Eds.) Albany: State University of New York.

Macri, M. J. (2001). *Maya Hieroglyphic Database Project*. Retrieved from http://nas.ucdavis.edu/NALC/mhdhome.html

Macri, M. J., & Looper, M. G. (2003). *The new catalog of Maya hieroglyphs, Volume One: The Classic period inscriptions.* Norman, OK: University of Oklahoma Press.

Maler, T. (1901). Researches in the central portion of the Usumacinta Valley. *Memoirs of the Peabody Museum of Archaeology and Ethnology, 2* (1).

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing.* Cambridge, MA: MIT Press.

Mathews, P., & Bíró, P. (2006). *Maya hieroglyph dictionary*. (Foundation for the Advancement of Mesoamerican Studies, Inc.) Retrieved August 2007, from FAMSI: http://research.famsi.org/mdp/mdp_index.php

Mathews, P., & Justeson, J. S. (1984). Patterns of sign substitution in Mayan hieroglyphic writing: The "affix cluster". *Phoneticism in Mayan hieroglyphic writing. Institute for Mesoamerican Studies Publication no. 9* , 185-231. (J. S. Justeson, & L. Campbell, Eds.) Albany: State University of New York.

Maudslay, A. P. (1889-1902). Archaeology. *Biologia Centrali-Americana; or, contributions to the knowledge of the fauna and flora of Mexico and Central America , I-VI* . (F. D. Godman, & O. Salvin, Eds.) London: R. H. Porter and Dulau and Co.

MediaWiki Handbook. (2007, September 2). Retrieved September 2007, from http://meta.wikimedia.org/wiki/Help:Contents

Montgomery, J. (2002). *Dictionary of Maya hieroglyphs.* New York: Hippocrene Books.

Montgomery, J. (2000). *The John Montgomery drawing collection*. Retrieved August 2007, from FAMSI: http://www.famsi.org/research/montgomery/

Mora-Marín, D. F. (2004). The preferred argument structure of Classic Lowland Mayan texts. In S. Wichmann, *The linguistics of Maya writing* (pp. 339-361). Salt Lake City: The University of Utah Press.

Ringle, W. M. (1985). Notes on two tablets of unknown provenance. In V. M. Fields (Ed.), *Fifth palenque round table, 1983: Palenque Round Table Series, vol. 7.* (pp. 151-58). San Francisco: Pre-Columbian Art Research Institute.

Ringle, W. M., & Smith-Stark, T. C. (1996). *A concordance to the inscriptions of Palenque, Chiapas, Mexico: Middle American Research Institute, publication 62.* New Orleans: Middle American Research Institute, Tulane University.

Robertson, J., Houston, S., & Law, D. H. (in press). Most Maya glyphs are in Classic Ch'olti'. In B. Metz, C. McNeil, & K. Hull (Eds.), *The Ch'orti' area, past and present.* Gainesville: University Press of Florida.

Robertson, J., Houston, S., & Stuart, D. (2004). Tense and aspect in Maya hieroglyphic script. In S. Wichmann, *The linguistics of Maya writing* (pp. 259-289). Salt Lake City: The University of Utah Press.

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

Schele, L. (1978-1988). *Notebook for the Maya Hieroglyphic Writing Workshop at Texas.* Austin, Texas: Institute of Latin American Studies, The University of Texas at Austin.

Schele, L. (1998). *The Linda Schele drawing collection*. Retrieved August 2007, from FAMSI: http://www.famsi.org/research/schele/

Silberschatz, A., Korth, H. F., & Sudrashan, S. (2002). *Database system concepts* (4th ed.). Boston: McGraw-Hill.

Sipser, M. (1997). *Introduction to the theory of computation.* Boston: PWS Publishing Company.

Smith-Stark, T., & Ringle, W. M. (1981). *Tinkered pans, charlock, coulter and share: Revising Thompson's catalog.* Unpublished manuscript.

Stephens, J. L. (1841). *Incidents of travel in Central America, Chiapas, and Yucatan.* London: W. Clowes and Sons.

Stuart, D. (1987). Ten phonetic syllables. *Research Reports on Ancient Maya Writing 14*. Washington, D.C.: Center for Maya Research.

Stuart, D. (2005a). *Sourcebook for the 29th Maya Hieroglyph Forum.* Austin, Texas: Department of Art and Art History, The University of Texas at Austin.

Stuart, D. (2005-2007). *Sourcebook for the Maya Hieroglyph Forum.* Austin, Texas: Department of Art and Art History, The University of Texas at Austin.

Stuart, D. (2005b). *The inscriptions from Temple XIX at Palenque.* San Francisco: The Pre-Columbian Art Research Institute.

Surrogate Key. (2007, August 30). Retrieved August 2007, from Wikipedia, The Free Encyclopedia: http://en.wikipedia.org/wiki/Surrogate_Key

Thompson, J. E. (1962). *A catalog of Maya hieroglyphs*. Norman, OK: University of Oklahoma Press.

Vail, G., & Hernández, C. (2005). *The Maya Hieroglyphic Codices, Version 2.0*. Retrieved from a website and database available online at: www.doaks.org/pc_research_projects.html (supersedes 2002, version 1.0 on November 1, 2005). Also available at www.mayacodices.org.

Viegas, F. B., Wattenberg, M., & Dave, K. (2004). Studying cooperation and conflict between authors with history flow visualizations. *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 575-582). Vienna, Austria: ACM Press.

Wagner, R. A., & Fischer, M. J. (1974). The string-to-string correction problem. *Journal of the Association for Computing Machinery , 21* (1), 168-173.

Wald, R. F. (2004). Telling time in Classic-Ch'olan and Acalan-Chontal narrative: The linguistic basis of some temporal discourse patterns in Maya hieroglyphic and Acalan-Chontal texts. In S. Wichmann, *The linguistics of Maya writing* (pp. 211-257). Salt Lake City: The University of Utah Press.

Wald, R. F. (2007). *The verbal complex in Classic-period Maya hieroglyphic inscriptions: Its implications for language identification and change (Ph.D. Dissertation, University of Texas at Austin, 2007)*. Retrieved from http://www.lib.utexas.edu/etd/d/2007/waldr55205/waldr55205.pdf

Wanyerka, P. (Ed.). (1989-2004). *The Proceedings of the Maya Hieroglyphic Workshop.* Austin, Texas: Art Department, University of Texas at Austin.

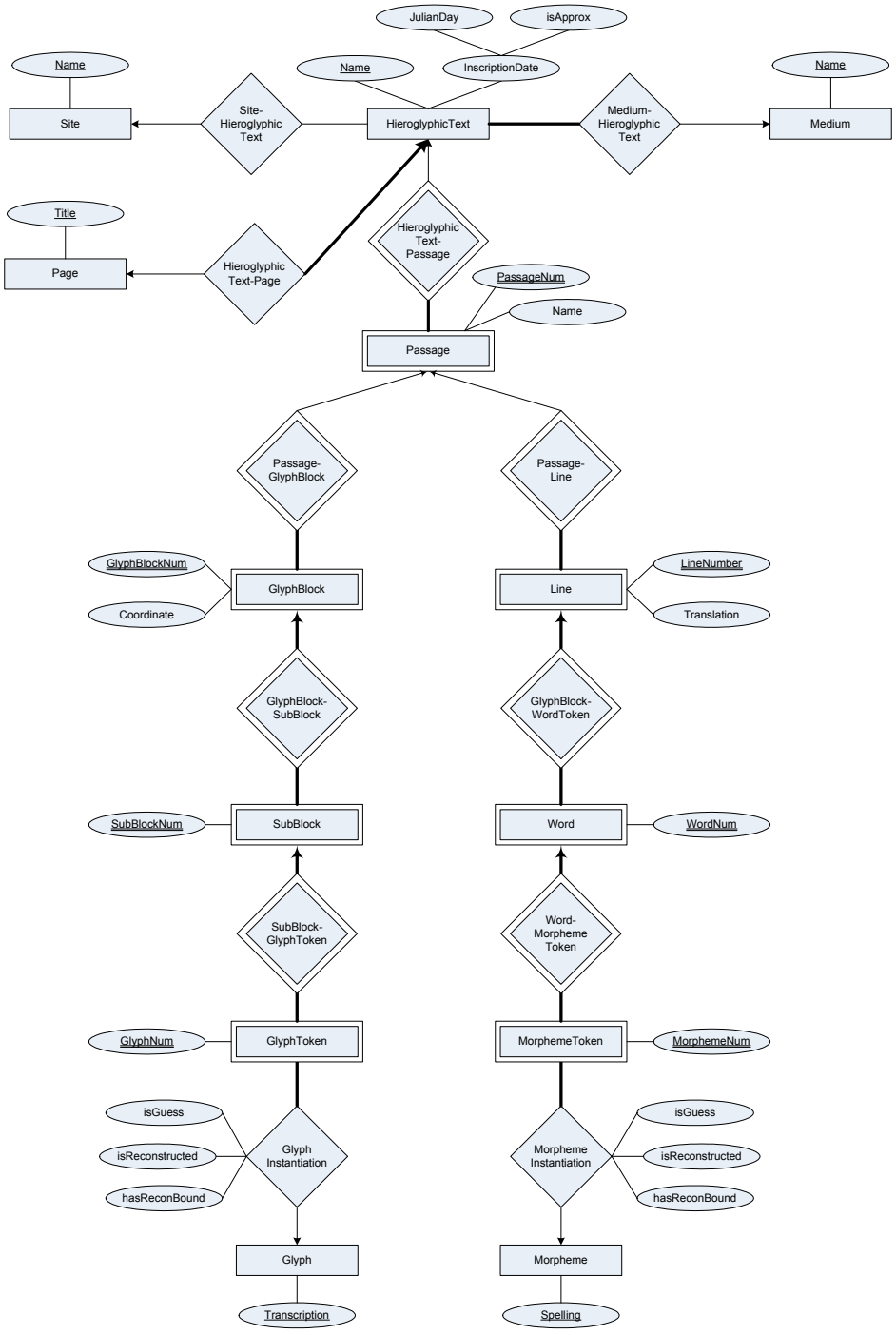
Welling, L., & Thomson, L. (2003). *MySQL tutorial.* Indianapolis, IN: MySQL Press.

Wichmann, S. (2004). The linguistic epigraphy of Maya writing: Recent advances and questions for future research. In S. Wichmann, *The linguistics of Maya writing* (pp. 1-10). Salt Lake City: University of Utah Press.

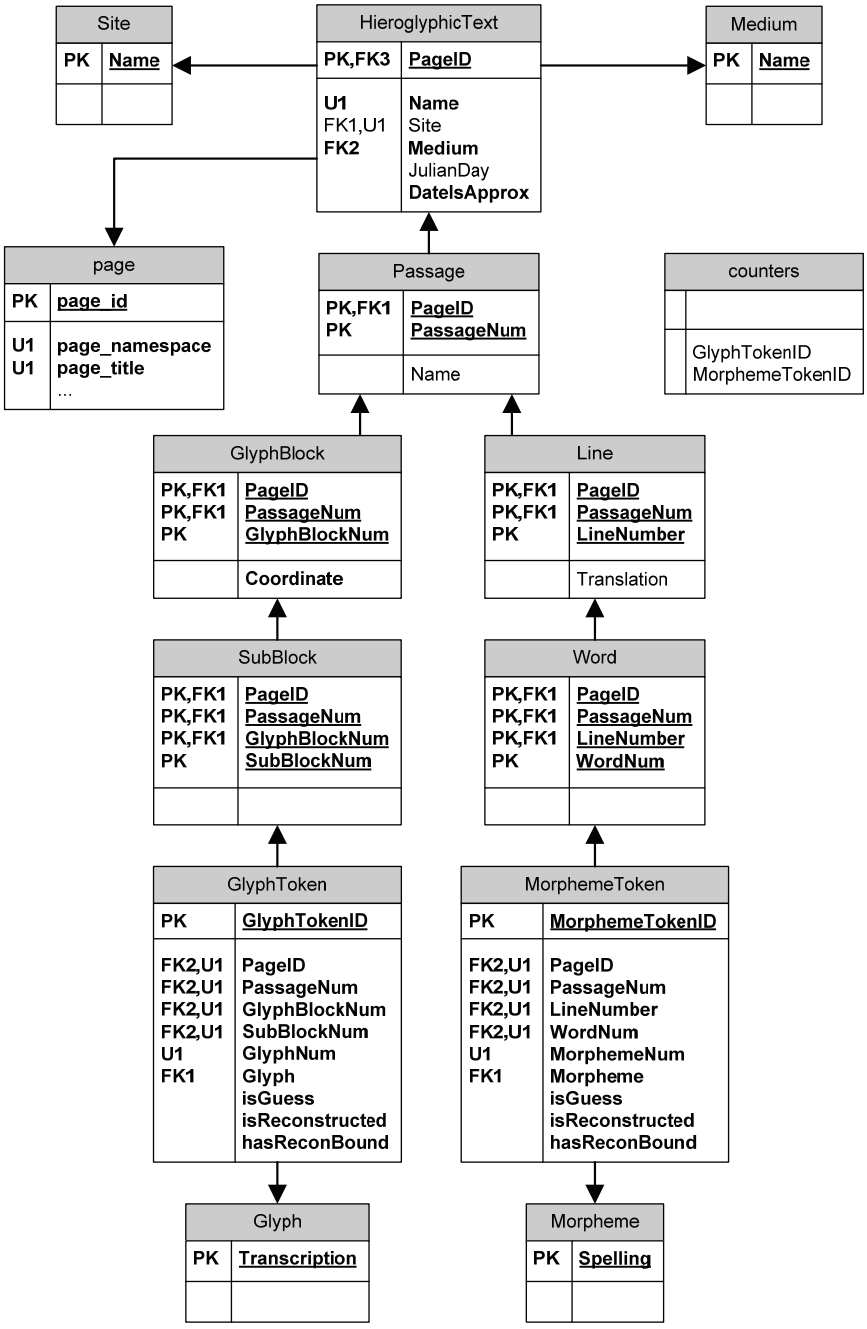
Wikipedia:User Page. (2007, September 6). Retrieved September 6, 2007, from http://en.wikipedia.org/wiki/Wikipedia:User_page

Zender, M. U. (1999). *Diacritical marks and underspelling in the classic Maya script: Implications for decipherment (Master's Thesis, Department of Archaeology, University of Calgary, 1999).* Retrieved from http://hdl.handle.net/1880/25407

Appendix A. Entity Relationship Diagram



Appendix B. Relational Model



Appendix C. MySQL Data Definition Statements

```
CREATE TABLE Site (
  Name            VARCHAR(50) CHARSET latin1 NOT NULL PRIMARY KEY
) ENGINE = InnoDB;

CREATE TABLE Medium (
  Name            VARCHAR(50) CHARSET latin1 NOT NULL PRIMARY KEY
) ENGINE = InnoDB;

CREATE TABLE HieroglyphicText (
  PageID          INTEGER UNSIGNED NOT NULL,
  Name            VARCHAR(50) CHARSET latin1 NOT NULL,
  Site            VARCHAR(50) CHARSET latin1,
  Medium          VARCHAR(50) CHARSET latin1 NOT NULL,
  JulianDay       INTEGER UNSIGNED,
  DateIsApprox    BOOLEAN NOT NULL DEFAULT FALSE,

  PRIMARY KEY (PageID),
  UNIQUE KEY (Site, Name),
  FOREIGN KEY (PageID) REFERENCES page(page_id) ON DELETE CASCADE ON
      UPDATE CASCADE,
  FOREIGN KEY (Site) REFERENCES Site(Name) ON DELETE RESTRICT ON UPDATE
      CASCADE,
  FOREIGN KEY (Medium) REFERENCES Medium(Name) ON DELETE RESTRICT ON
      UPDATE CASCADE

) ENGINE = InnoDB;

CREATE TABLE Passage (
  PageID          INTEGER UNSIGNED NOT NULL,
  PassageNum      TINYINT UNSIGNED NOT NULL,
  Name            VARCHAR(50) CHARSET latin1,

  PRIMARY KEY (PageID, PassageNum),
  FOREIGN KEY (PageID) REFERENCES HieroglyphicText(PageID) ON DELETE
      CASCADE ON UPDATE CASCADE

) ENGINE = InnoDB;
```

```
CREATE TABLE GlyphBlock (
  PageID          INTEGER UNSIGNED NOT NULL,
  PassageNum      TINYINT UNSIGNED NOT NULL,
  GlyphBlockNum   TINYINT UNSIGNED NOT NULL,
  Coordinate      VARCHAR(15) CHARSET latin1 NOT NULL,

  PRIMARY KEY (PageID, PassageNum, GlyphBlockNum),
  FOREIGN KEY (PageID, PassageNum) REFERENCES Passage(PageID,
      PassageNum) ON DELETE CASCADE ON UPDATE CASCADE

) ENGINE = InnoDB;

CREATE TABLE Line (
  PageID          INTEGER UNSIGNED NOT NULL,
  PassageNum      TINYINT UNSIGNED NOT NULL,
  LineNumber      TINYINT UNSIGNED NOT NULL,
  Translation     VARCHAR(1024) CHARSET latin1,

  PRIMARY KEY (PageID, PassageNum, LineNumber),
  FOREIGN KEY (PageID, PassageNum) REFERENCES Passage(PageID,
      PassageNum) ON DELETE CASCADE ON UPDATE CASCADE

) ENGINE = InnoDB;

CREATE TABLE SubBlock (
  PageID          INTEGER UNSIGNED NOT NULL,
  PassageNum      TINYINT UNSIGNED NOT NULL,
  GlyphBlockNum   TINYINT UNSIGNED NOT NULL,
  SubBlockNum     TINYINT UNSIGNED NOT NULL,

  PRIMARY KEY (PageID, PassageNum, GlyphBlockNum, SubBlockNum),
  FOREIGN KEY (PageID, PassageNum, GlyphBlockNum) REFERENCES
      GlyphBlock(PageID, PassageNum, GlyphBlockNum) ON DELETE CASCADE
      ON UPDATE CASCADE

) ENGINE = InnoDB;

CREATE TABLE Word (
  PageID          INTEGER UNSIGNED NOT NULL,
  PassageNum      TINYINT UNSIGNED NOT NULL,
  LineNumber      TINYINT UNSIGNED NOT NULL,
  WordNum         TINYINT UNSIGNED NOT NULL,

  PRIMARY KEY (PageID, PassageNum, LineNumber, WordNum),
  FOREIGN KEY (PageID, PassageNum, LineNumber) REFERENCES Line(PageID,
      PassageNum, LineNumber) ON DELETE CASCADE ON UPDATE CASCADE

) ENGINE = InnoDB;

CREATE TABLE Glyph (
  Transcription    VARCHAR(50) CHARSET latin1 COLLATE latin1_bin NOT
      NULL PRIMARY KEY
) ENGINE = InnoDB;
```

```
CREATE TABLE Morpheme (
  Spelling        VARCHAR(50) CHARSET latin1 NOT NULL PRIMARY KEY
) ENGINE = InnoDB;

CREATE TABLE GlyphToken (
  GlyphTokenID    INTEGER UNSIGNED NOT NULL,
  PageID          INTEGER UNSIGNED NOT NULL,
  PassageNum      TINYINT UNSIGNED NOT NULL,
  GlyphBlockNum   TINYINT UNSIGNED NOT NULL,
  SubBlockNum     TINYINT UNSIGNED NOT NULL,
  GlyphNum        TINYINT UNSIGNED NOT NULL,
  Glyph           VARCHAR(50) CHARSET latin1 COLLATE latin1_bin NOT
      NULL,
  isGuess         BOOLEAN NOT NULL,
  isReconstructed BOOLEAN NOT NULL,
  hasReconBound   BOOLEAN NOT NULL,

  PRIMARY KEY (GlyphTokenID),
  UNIQUE (PageID, PassageNum, GlyphBlockNum, SubBlockNum, GlyphNum),
  FOREIGN KEY (PageID, PassageNum, GlyphBlockNum, SubBlockNum)
      REFERENCES SubBlock(PageID, PassageNum, GlyphBlockNum,
      SubBlockNum) ON DELETE CASCADE ON UPDATE CASCADE,
  FOREIGN KEY (Glyph) REFERENCES Glyph(Transcription)

) ENGINE = InnoDB;

CREATE TABLE MorphemeToken (
  MorphemeTokenID INTEGER UNSIGNED NOT NULL,
  PageID          INTEGER UNSIGNED NOT NULL,
  PassageNum      TINYINT UNSIGNED NOT NULL,
  LineNumber      TINYINT UNSIGNED NOT NULL,
  WordNum         TINYINT UNSIGNED NOT NULL,
  MorphemeNum     TINYINT UNSIGNED NOT NULL,
  Morpheme        VARCHAR(50) CHARSET latin1 NOT NULL,
  isGuess         BOOLEAN NOT NULL,
  isReconstructed BOOLEAN NOT NULL,
  hasReconBound   BOOLEAN NOT NULL,

  PRIMARY KEY (MorphemeTokenID),
  UNIQUE (PageID, PassageNum, LineNumber, WordNum, MorphemeNum),
  FOREIGN KEY (PageID, PassageNum, LineNumber, WordNum) REFERENCES
      Word(PageID, PassageNum, LineNumber, WordNum) ON DELETE CASCADE
      ON UPDATE CASCADE,
  FOREIGN KEY (Morpheme) REFERENCES Morpheme(Spelling)

) ENGINE = InnoDB;

CREATE TABLE counters ENGINE=MEMORY
SELECT IFNULL(MAX(GlyphTokenID),1) AS GlyphCounter,
 IFNULL(MAX(MorphemeTokenID),1) AS MorphemeCounter
FROM GlyphToken, MorphemeToken;
```