1. A VIRTUAL RESEARCH ENVIRONMENT TO DOCUMENT AND ANALYZE NON-ALPHABETIC WRITING SYSTEMS

A Case Study for Maya Writing

Abstract

No existing digital work environment can sufficiently represent the traditional epigraphic workflow 'documentation, analysis, interpretation, publication' for a non-alphabetic writing system. Using the Maya hieroglyphic script, this workflow will be transferred to a digital epigraphy. Digital methods and tools will be developed and reused in a Virtual Research Environment to create a freely accessible, annotated textual corpus, including metadata on cultural and object history and references.

Keywords

1.1. Digitalizing the Epigraphic Workflow

In order to develop a digital work environment that represents the traditional epigraphic workflow for documenting and studying inscriptions in a non-alphabetic script, the individual steps of this workflow need to be technically implemented using methods and tools from the Digital Humanities. The workflow begins with the documentation of the text carriers and compilation of descriptive data; proceeds to epigraphic analysis, including sign classification and transliteration and transcription of the texts; continues with morphological segmentation and linguistic interpretation; and concludes with translation and digital publication of the inscription [Diederichs 2015].

The Virtual Research Environment (VRE) TextGrid initially provides the necessary framework for executing the workflow that technically facilitates documentation and digital registration of the text carriers in the form of images and drawings, as well as annotation of the objects with metadata. TextGrid offers input forms, annotation tools, and a Text-Image-Link-Editor, and provides long-term storage, and free access to the data in an open web repository [Neuroth & Lohmeier 2011].

Although more detailed steps in an epigraphic workflow, such as detailed linguistic and epigraphic analysis of primary texts, can be technically carried out within a VRE such as TextGrid, they still necessitate creation and adaptation of their own XML-based metadata model, as well as annotation schemas for object and textual contexts [Prager 2015]. Using this approach, the Maya hieroglyphic script will for the first time be able to fulfill a central requirement of corpus linguistics, namely machine readability [McEnery & Wilson 2001].

Figure 1.1. provides a schematic illustration of the following points that represents in great detail the steps in epigraphically analyzing Maya inscriptions, a process which can also be used in modified form to study other non-alphabetic writing systems. The modulated VRE is thus applicable not only to the study of Maya texts, but also to researching texts composed in other hieroglyphic, cuneiform, or linear writing systems.

1. Identification of hieroglyph blocks by alphanumeric classification
2. Original spelling
3. Reading order of individual signs
4. Number of signs within one block
5. Identification and isolation of signs in one block
6. Classification of signs based on the sign catalog of Eric Thompson [Thompson 1962]
7. Classification of signs in one block [Thompson 1962]

8. Transliteration of individual signs
9. Description of sign function (syllables, logographs)
10. Transliteration
11. Broad transcription
12. Morphological segmentation
13. Morphological analysis
14. Determining congruence between block- and word-boundaries
15. Determining syntactic function
16. Determining sentence constituents
17. Translation of single words
18. Literal translation
19. Rough translation

1	D13			E13					D14		
2											
3	type A			type B					type A		
4	3			5					3		
5											
6⁶	T264	T21	T575	T1	T257	T1	T624	T178	T74	T712°504	T178
7	T264:21:575			T1:257.1:624:178					T74:712°504:178		
8	ju	bu	yi	U	TOK'	U	PAK	la	ma	CH'AB° AK'	la
9	syallable	syallable	syallable	logograph	logograph	logograph	logograph	syllable	syllable	logograph	syllable
10	ju-bu-yi			U-TOK' U-PAK-la					MA' CH'AB°AK'-la		
11	jubuy			utok' upakal					ma' ch'ab [ma'] ak'al		
12	jub-uy-ø			u-tok' u-pak-al					ma' ch'ab ma' ak'-al		
13	to come down-THM-3SA			3SE-flint 3SE-shield					NEG-creation-[NEG]-night		
14	1:1			1:2					1:4		
15	verb			subject							
16	verb			patient							
17	verbal phrase			nominal phrase							
18	"came down"			"his flint, his shield"					"no creation, no night"		
19	"it came down his flint, his shield, he who is without creation, without night"										
20	"defeated was the army of the captive"										

Fig. 1.1. Visualization of the important steps in the workflow of epigraphic analysis, drawings by Christian Prager, 2015

1.1.1. Conceptual Questions

When developing and technically implementing the digital epigraphic research environment, the major focus is on pursuing the following objectives: 1) creation of a markup schema to represent these steps (which in practice are interconnected and may also occur parallel to each other in the form of alternative interpretations);

2) defining of the precise requirements that a data structure must fulfill in order to sufficiently represent complex i.e. non-alphabetic textual data for epigraphic analysis and research.

2.1. Prerequisites for Making a Non-Alphabetic Writing System Digitally Accessible

In order to make syllabic or logo-syllabic writing systems accessible for corpus linguistics, for instance to carry out a corpus analysis as part of the basic lexicography required to compile a dictionary of Classic Mayan, the phonology, morphology, syntax, semantics, and pragmatics of the language in question must be captured, marked up, and saved in the corpus. As such, the basic characteristics that distinguish the graphematic lexicon of syllabic and logo-syllabic scripts from those of alphabetic writing systems must be taken into account. Thus, methodological and technical prerequisites must be fulfilled in order to be able to digitally represent and analyze a non-alphabetic script, such as Classic Mayan, for the purposes of epigraphic analysis.

2.1.1. Registering Object Data

When designing the digital epigraphic work environment, one must also account for the current state of decipherment of the script and language that are to be documented. Thus, Classic Mayan presents great challenges, as at least one third of the known graphemes have not yet been completely deciphered. Even when a lexeme can be read, its etymology or semantic domain is often unknown, and in some cases not even its lexical class can be determined, which causes difficulties when attempting to read and understand a text. In order to decode an unknown sign, word, or sign sequence in its respective context of use, for instance, semantic fields are generally investigated by means of applied substitution. From a structuralist perspective, this method investigates the paradigmatic relation between two linguistic entities as indicated by whether or not they appear in the same context [Bußmann 2002]. This step yields results when sufficient quantities of data can be digitally compiled, marked up and investigated using corpus linguistics.

The VRE currently being planned is the laboratory for this process, where new decipherments will be obtained and existing readings will be tested against primary source material while also considering all of the contextual data. The context of the contents that is referenced during decipherment may refer to the text carrier itself. For this reason, registering data on the object on which the text is written is just as important as documenting the text itself.

The decoding of Ugaritic cuneiform at the beginning of the twentieth century provides an example of a successful decipherment that resulted from studying the relationship between object and textual data. Scholars suspected that the word "axe" occurred in repeating, brief sign sequences on inscribed bronze axes. Building on this hypothesis, they determined the phonetic values of individual signs using data on the Canaantic language and completely deciphered the script [Day 2002]. Such investigations enable initial association of signs with both content and function in the relevant language so that the signs may later be grammatically, morphologically, semantically, etc., defined and eventually deciphered. This work also facilitates more precise typological classification of a writing system [Gronemeyer 2015].

Like the relationship between writing and the object, the relationship between text and image must also be coded. Associated depictions can potentially illuminate the contents of an inscription. Working from this analytical basis, a writing system can be registered in its respective contexts of use in order to derive semantic meanings and, ideally, linguistic decipherments. As such, contextual information concerning the text carrier as an object must be recorded, because its physical features may influence the text's contents and arrangement. In this respect, the size of the object, for example, may be just as influential as the material itself.

For example, the text carrier may influence scribal economy. The recent discovery of an inscription on very hard jade in Nim Li Punit (Belize) illustrates how the use of this particular writing surface can lead to a reduced spelling, meaning that the scribe omitted final syllables or used very simple sign variants that were almost geometric.

The form of a text carrier and its function also influence the text. Different lexemes may thus be determined for describing round or quadrilateral altars, with the result that conclusions may be drawn concerning the semantics and possible linguistic decipherments. Physical description of the text carrier as an object and its linguistic, historical, and cultural contextualization are indispensable for deciphering a previously unknown text or analyzing its arrangement. These metadata must be consulted in order to grammatologically and linguistically analyze the text's contents.

Modelling an object biography and culturally contextualizing the texts must be initially addressed with information technology in order to facilitate epigraphic analysis of the text's contents in its extra-linguistic, cultural context. Thus, it is necessary to create an object metadata schema that captures

such information concerning the language's cultural area and makes it available for use.

For these purposes, an epigraphic object metadata schema that compiles and represents extra-linguistic information about the inscriptions in an ontological structure must be created, in addition to an XML-based analytical metadata schema for linguistics.

The CIDOC-CRM standard for documenting cultural heritage offers a broad foundation that can be supplemented with other standards, such as Dublin Core or the SKOS vocabulary. In addition to taking advantage of the comprehensive understandability of CIDOC-CRM, the scheme also requires to reuse data and to connect them with data from other research projects in the spirit of Linked Open Data [Diederichs 2015 and Prager 2014].

2.1.2. Making Linguistic Data Accessible

An interdisciplinary and widely accepted standard that takes into account both generalized and special characteristics of writing systems beyond those of a single script must be developed in order to be able to satisfactorily annotate records from various writing traditions that use syllabic signs and logograms to represent language.

In most cases, the scripts annotated and encoded in TEI and also those prepared for epigraphic analysis in EpiDoc are alphabetic, and most of them are linear and arranged in rows. Non-alphabetic writing systems, such as Egyptian, various cuneiform systems, or Hieroglyphic Luwian, arrange their graphemes principally in groups or blocks, and only secondarily in rows or columns, with a high rate of metathesis [Lacau 1903] in order to optimize the use of space. Furthermore, many of these writing systems demonstrate a high degree of allography, which is no longer apparent in a pure transliteration. Various desiderata become apparent when an epigraphic project attempts to use the standards provided by TEI and EpiDoc to edit texts composed in these scripts.

Annotating texts composed in such writing systems that use only alphabetic transcription is insufficient; instead, the original spelling of a lemma should be represented. This approach facilitates studies of paleography or sign usage across time and space. In addition, it can reveal preferred sign arrangements within a block or preferences for particular signs that correspond to the material used or to the subject of the text, for instance.

Thus, when creating a standard for non-alphabetic scripts, object metadata and annotations for orthographic and linguistic analysis must be taken into

account, and the creative process should also promote discussion of terminological and typographic conventions for these annotations [Sachse & Dürr 2015]. The goal is to create a generic markup model that can be implemented regardless of the particular script or language in question. The inscriptions of the Maya hieroglyphic writing system alone feature attestations of various vernacular languages [Lacadena & Wichmann 2002], a phenomenon which raises several preliminary questions: what exactly are the demands, and what are the goals of an epigraphic and linguistic annotation? Where do possibilities and limitations exist for annotating these writing systems using established standards, such as TEI or EpiDoc? Existing standards have to be modified to some extend and possibly new standards have to be created, to overcome these limitations.

2.1.2.1. Representing the Primary Text Source

When defining a text, acknowledging an interpretation of the source text itself is unavoidable. The meaning of the "original text" that resides in the "source text" is an important point of discussion in epigraphy. Thus, the annotation does not contain a guiding text in the conventional sense, because it is merely reconstructed using all given text information and one's own conclusions and interpretations, and then represented with the aid of alphabetic transliteration and transcription. In this respect, one should always leave open the possibility of separating primary data (in the case of hieroglyphic texts, photographs of the text carriers, for example) and secondary data (e.g. drawings or interpretive annotations) [Stührenberg 2012]. This strategy can be implemented using stand-off annotations for data. Similarly, such an annotation should permit multiple descriptions, i.e. alternative statements concerning the data [Stührenberg 2012]. For example, sign sequences in Classic Mayan can be variously analyzed depending in a particular vernacular context [Gronemeyer 2014]. Only in this manner can various ways of thinking be appropriately recorded and relayed to the scientific community for discussion.

A key objective of the XML-based markup of hieroglyphic, cuneiform, or linear texts should thus be to represent the original spelling and arrangement of the signs in their respective contexts. A linear transcription alone cannot represent the original text or primary source in its entirety, as many details remain undocumented. For example, signs in Maya hieroglyphic texts are not arranged according to their literal reading order, but instead in spatially distinct, square or rectangular units (so-called "blocks"), each of which in most cases corresponds to a word or morpheme sequence. A detailed markup of the

original text is therefore of great importance, particularly for partially deciphered and undeciphered graphemes and writing systems in general. In such cases, an alphanumerical or numerical nomenclature is often used to refer to the signs in order to carry out a corpus linguistic analysis of the texts. The arrangement of each block, as well as the text and its position on a text carrier, should be documented as well. To fully understand a writing system – i.e. the language and information expressed in it - a detailed representation of the primary source and its context must be a key objective in the digital markup of documents. When studying complex writing systems (respectively non-alphabetic writing systems), digital documentation of the original spelling using annotation standards like TEI is a basic prerequisite for conducting a detailed graphemic and graphetic analysis of the relevant script and for providing a basis for a linguistic and corpus linguistic investigation. This need represents a significant desideratum in epigraphic research, and it also constitutes a core pillar of computational linguistics [McEnery & Wilson 2001].

In an interdisciplinary effort, epigraphers and experts of digital markup-languages alike need to discuss methods for investigating syllabic and logo-syllabic writing systems using the XML-based standards like TEI or EpiDoc. When doing so, they should discuss the following points of emphasis, among others:

2.1.2.2. Graphetics

Graphetics concerns the formal structure of linguistic units and the structure of texts [Crystal 1997]. Classic Maya hieroglyphic inscriptions, for instance, may display a very high degree of variation in writing styles for arranging texts, a phenomenon which is deserving of investigation. In this manner, various systems of notation are studied in their individual, social, and typographic aspects, for instance [Bußmann 2002]. Similarly, in paleography, script decipherment is examined from graphetic perspective [Bußmann 2002]. The arrangement of texts, for instance, different grapheme or block sizes of text fields, or different styles of hieroglyphic writing (equivalent to alphabetic font style and size) are analyzed.

Research addresses the degree to which attested variation in the composition of hieroglyphic texts may be ascribed to specific scribal schools or to regional forms of expression, rather than to differences in the meaning of the signs' contents.

2.1.2.3. Graphemics

When analyzing the contents of these texts, it is important to investigate meaningful variation underlying the manner in which the writing is arranged and to thereby identify distinct units [Bußmann 2002]. Graphemics is devoted to this area of research, and consequently to exploring the meaningfully distinct features of signs.

2.1.2.3.1. Allography of Graphemes

A very important example of a subject of graphemic research is provided by allography, a phenomenon which describes a 1:n relationship between a grapheme and its various graph representations [Crystal 1997].

The process of establishing a suitable annotating-tool thus must be oriented toward a series of questions: What is the significance of allography for the respective grapheme inventories? Are allographs annotated in the transliteration or transcription of the respective writing systems?

For example, in Maya writing, there are over twenty different allographs for the vowel sign <u>, which is used e.g. as the 3rd person ergative pronoun u- "he, she, it". Additional research questions may comprise 1) the reading order of the signs in context; 2) the existence of word separators or other graphic aids used to differentiate between meaningful units or words; 3) multiple possibilities for establishing a reading order of signs within a block that are equally meaningful (with alternatives documented or marked by epigraphers); 4) the reading order of meaningful units within the texts; 5) variations or violations of these orders (and if so, with a markup as well); 6) cases where images appear with the text and their possible relation; 7) integration of texts into the images or vice versa.

2.1.2.3.2. Graphotactic Strategies of a Writing System

For analysis, it is important to convey the graphemics of a writing system, which are lost in an alphabetic transliteration. Graphemics concerns sign function, or graphotactic strategies for constituting a meaningful unit or word. Representing the reading order of signs in the text and in each block is similarly imperative. To graphotactically link individual signs in a hieroglyphic block in Maya writing, various representational rules may be used (see Fig. 1.2.), e.g. affixation, infixation, conflation, or superimposition [Zender 1999]. Additional typical, functional traits of logo-syllabic writing systems include underspellings, diacritics, phonemic complements and indicators, semantic classifiers and determinatives, and sign class convergences. However, these

characteristics may have very diverse graphotactical manifestations in differ-
ent writing systems, which nonetheless need to be standardized and commu-
nicated in a digital structure in order that researchers may thus take note of
various interpretations under discussion.
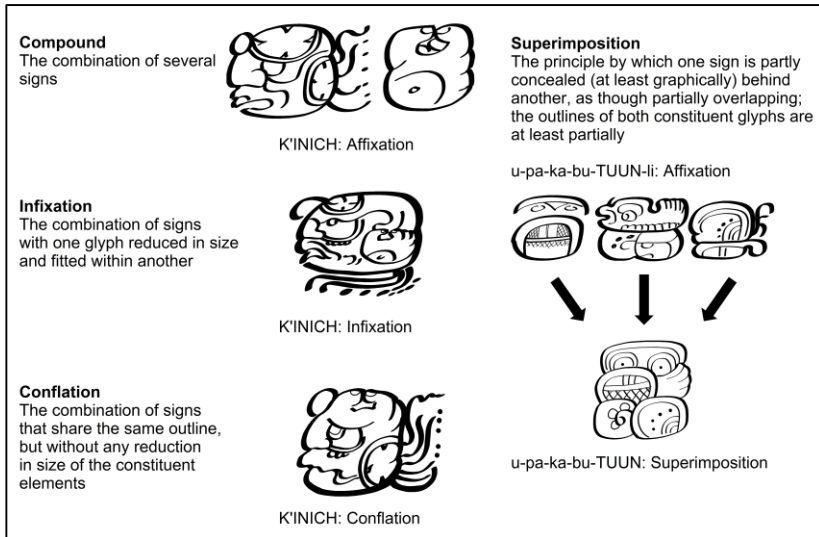
Fig. 1.2. Some possible sign combinations in Classical Mayan hieroglyphs [after Zender 1999],
drawings by C. Prager 2015

2.1.2.4. Signary: Sign Classification and Sign Catalogs

The graphemic lexicon provides a central reference for marking up texts
composed in non-alphabetic writing systems. In the case of the Maya script,
signs whose phonemic reading is unknown or unconfirmed are denoted using
reference systems from various classification catalogues (e.g. Thompson
1962). Use of question marks for indicating unclear readings, for example,
should be avoided, because such signs impair machine readability and some-
times represent control characters or part of an escape sequence. In this con-
text, it should be determined what significance sign classifications and no-
menclatures have in their respective branches of epigraphic research. Also
their role in transliterating and transcribing texts in the respective writing sys-
tem, as well as in sign classification is to be examined. Sign inventories con-
stitute a central authority in an epigraphic VRE, and they need to be modified
or supplemented in accordance with the ever-changing state of research as
soon as newly discovered material is entered and signs are identified that had
not previously been isolated or identified as discrete graphemes.

2.1.2.5. The State of Decipherment and the Readability of Texts

In the case of unreadable or only partially deciphered writing systems, the question in the forefront is the state of decipherment of the relevant writing system and the handling of undeciphered signs or text passages. As archaeological artifacts, texts are also subject to processes of decay that inhibit visible legibility. It has several implications: 1) the treatment of undeciphered signs or text passages; 2) the indication of physical gaps (as texts are also subject to processes of decay that inhibit visible legibility) and their marking in the transliteration and transcription of the text; 3) the treatment of hypothetical, yet plausible, readings of singular signs or passages; 4) the handling of alternative readings of the same passage; and 5) a review to what extent the EpiDoc Metadata standard, into which the Leiden Conventions have also been incorporated, is able to capture these features of readability and alternative interpretations of inscriptions.

These and other characteristics of non-alphabetic scripts, which to a certain extent reach beyond the epigraphic problematic of alphabetic writing systems, illuminate a desideratum here and the need to formulate ideas that contribute to the design and modelling of appropriate XML-based metadata schemas.

3.1. Open Science Strategy

A Digital Humanities project working in the spirit of Open Science fundamentally intends the work environment that it develops to be reused. As such, a VRE for non-alphabetic writing system such as that of the Classic Maya should be developed as generically as possible, in order that structural interoperability may ensure that they can be reused by similar epigraphic projects.

In this respect, any metadata schema that is created should be widely intelligible and ensure the data's syntactic and semantic interoperability by using common, given, XML-based annotation and metadata standards, like TEI and EpiDoc, and controlled vocabularies, such as SKOS, that are already established in the Digital Humanities [Diederichs 2015].

Another objective of database projects in the Digital Humanities should be to ensure open access to their data and metadata, and thus to maximize usage of their research by maximizing user access to it. As all current and future research and innovation stands on the shoulder of giants, an efficient system for broadly disseminating and allowing uninhibited access to the pro-

ject's research (raw data and metadata), as well as for guaranteeing the contents' productive reuse, must be ensured through the use of free licenses [DFG 2014].

As such, all the contents of a database and data infrastructure that are being planned will be made available under so-called "Open" licenses (Open Access, Open Source, Open Methodology, Open Data, etc.), with the goal of pursuing a comprehensive Open Science strategy [Diederichs 2015].

All methods and the process of developing tools must thus be carefully documented to maximize reuse of all schemas and software. In particular, all research information and results should be made available not only to the scientific community, but also to the general public, to contribute to the digital safeguard and dissemination of humanity's cultural heritage.

4.1. Summary

The creation of digital infrastructures and metadata schemas that facilitate recording and further analysis of the appropriate annotation for non-alphabetic writing systems, including their respective cultural contexts, is a desideratum in the Digital Humanities that needs to be addressed.

The recommendations articulated here for an object database, as well as for a structure for recording epigraphic and linguistic data, can make a significant contribution to understanding how traditional epigraphy can be transferred to digital form. In working towards this goal, methods and tools from the field of Digital Humanities will be suitably adapted to the needs of epigraphy, and new methods and structures from the digital world will find their way into digital epigraphy.

Bibliography

BUẞMANN, H., *Lexikon der Sprachwissenschaft*, Stuttgart 2002.

CRYSTAL, D., *The Cambridge Encyclopedia of Language*, Cambridge 1997.

DAY, P. L., *Dies diem docet: The Decipherment of Ugaritic,* Studi Epigrafici e Linguistici sul Vicino Oriente antico, 19 (2002), 37-57.

DEUTSCHE FORSCHUNGSGEMEINSCHAFT (DFG), *Praxisregeln "Digitalisierung"*, Bonn 2013.

DIEDERICHS, K., *The "Open Science" Strategy of the Project "Text Database and Dictionary of Classic Mayan"*, Textdatenbank und Wörterbuch des Klassischen Maya Working Paper 1, Bonn 2015 (in print).

GRONEMEYER, S., *E pluribus unum: Embracing Vernacular Influences in a Classic Mayan Scribal Tradition*, in Ch. Helmke, F. Sachse (edd.), *A Celebration of the Life and Work of Pierre Robert Colas*, Markt Schwaben 2014.

– *Class Struggle: Towards a Better Understanding of Maya Writing Using Comparative Graphematics*, in H. Kettunen, Ch. Helmke (edd.), *Proceedings of the 17th European Maya Conference, 2012*, Markt Schwaben 2015.

MCENERY, T. & WILSON, A., *Corpus Linguistics: An Introduction, Second edition*, Edinburgh 2001.

LACADENA, A. & WICHMANN, S., *The Distribution of Lowland Maya Languages in the Classic Period*, in V. Tiesler Blos, M. Greene Robertson, R. Cobos (edd.), *La organización social entre los mayas*, Mexico City 2015.

LACAU, P., *Métathèses apparentes en égyptien*, Recueil de travaux relatifs à la philologie et à l'archéologie égyptiennes, 25, 3-4 (1903), 139-161.

NEUROTH, H. & LOHMEIER, F., *TextGrid - Virtual Research Environment for the Humanities*, The International Journal of Digital Curation, 6 (2011), 222-231.

PRAGER, C., *Das Textdatenbank- und Wörterbuchprojekt des Klassischen Maya. Möglichkeiten und Herausforderungen digitaler Epigraphik,* in H. Neuroth, A. Rapp, S. Söring (edd.), *TextGrid: Von der Community - für die Community*, Göttingen 2015.

SACHSE, F. & DÜRR, M., *Morphological Glossing of Mayan Languages under XML: Preliminary Results*, Textdatenbank und Wörterbuch des Klassischen Maya Working Paper 4, Bonn 2015 (in print).

STÜHRENBERG, M., *Auszeichnungssprachen für linguistische Korpora : theoretische Grundlagen, De-facto-Standards, Normen*, Bielefeld 2012.

THOMPSON, J. E. S., *A Catalog of Maya Hieroglyphs*, Norman 1962.

ZENDER, M., *Diacritical Marks and Underspelling in the Classic Maya Script: Implications for Decipherment*, (MA Thesis), University of Calgary, 1999.