Contributions in New World Archaeology 7: 37-46

CORPUS EPIGRAPHY: LINGUISTIC IMPLICATIONS AND DIDACTIC APPLICATIONS

Harri Kettunen

Academy of Finland and University of Helsinki, Finland

Abstract

Corpus epigraphy deals with statistical analyses of hieroglyphic signs in the corpus of ancient scripts. The initial research and the main focus of this study concerns Maya texts. The primary objective of corpus epigraphy discussed in this article is to detect whether there are temporal and/or regional differences in the choice of words, phrases, and clauses in the Maya hieroglyphic corpus. By detecting these distribution patterns, the aim is to reach a more comprehensive understanding of regional and temporal variance and diversity – and the phenomena associated with them – and to reinforce the hypothesis that the ancient Maya culture is not a homogenous entity.

The implications and applications of corpus epigraphy can be utilized for scholarly as well as didactic purposes. The preliminary results of this research project indicate that some of the rules that have been applied to the writing system of the ancient Maya by modern scholars do not seem to operate flawlessly. In brief, instead of imposing our somewhat rigid linguistic theories upon an ancient writing system, we should explore carefully the practices of the scribes who wrote these texts in the first place.

The didactic dimension of this research is that it facilitates the learning process of the ancient Maya writing system among the students who embark upon deciphering the script. For example, frequency lists of the most common hieroglyphs in a certain area or specific time period can be utilized during workshops on Maya hieroglyphic writing. Also, students can detect patterns and rules that govern phonetic complementation as well as learn how to reconstruct the contents of eroded inscriptions.

Resumen

La epigrafía de corpus se ocupa del análisis estadístico de los signos jeroglíficos en el corpus de escrituras antiguas. La investigación inicial y el mayor punto de interés del presente estudio tiene que ver con los textos mayas. El objetivo principal de la epigrafía de corpus comentado en este articulo es detectar las posibles diferencias temporales y/o regionales en la selección de las palabras, frases y oraciones procedentes del corpus jeroglífico maya. Al detectar estas pautas de distribución se intentará llegar a una comprensión más amplia de la variedad y la diversidad regional y temporal, así como de los fenómenos asociados con ellas, y reforzar la hipótesis de que la antigua cultura maya no tenía un carácter homogéneo.

Las implicaciones y aplicaciones de la epigrafía de corpus pueden aprovecharse con fines académicos y didácticos. Los resultados preliminares del presente proyecto indican que algunas de las reglas que han sido aplicadas al sistema de escritura de los antiguos mayas por los investigadores modernos no siempre funcionan de manera perfecta. Por ello, en vez de imponer nuestras teorías lingüísticas rígidas a un sistema de escritura antigua, en primer lugar deberíamos investigar cuidadosamente las prácticas de los escribanos que crearon los textos en cuestión.

La dimensión didáctica de la investigación consiste en facilitar el proceso de aprendizaje del antiguo sistema de escritura maya a los estudiantes que se introducen en el desciframiento de la escritura. Por ejemplo, las listas de frecuencia de los jeroglíficos más comunes en un área determinado o periodo de tiempo concreto pueden ser utilizados durante los talleres dedicados a la escritura jeroglífica maya. Además, los estudiantes pueden detectar los modelos y reglas que rigen la complementación fonética, así como aprender a reconstruir el contenido de las inscripciones erosionadas.

"We have to remember that what we observe is not nature herself, but nature exposed to our method of questioning"

Werner Heisenberg
Physics and Philosophy: The Revolution in Modern Science (1958)

INTRODUCTION

Corpus epigraphy, to coin a term, stems from corpus linguistics, with the exception that besides language itself, the focus is on the writing system. The study is a marriage between corpus linguistics and epigraphic studies, with methodological support from iconographic studies, inspired by the work of Tatiana Proskouriakoff (1950). This methodology was developed in Kettunen 2006, dealing with regional and diachronic distribution of iconographic motifs, followed by a pilot project (2007–2009) and an ongoing project (2013–2018) on Maya epigraphy, both funded by the Academy of Finland. In 2013 cooperation was started with Ignacio Cases to further develop the methods involved in corpus epigraphy in the future, especially as relates to employing artificial intelligence in the process of analyzing the data.

METHODOLOGY

To commence a corpus epigraphic study, one has to amass a considerable data set. At the time of writing this article (2013), the database contains approximately 15,000 hieroglyphs. However, the aim of the project is to come up with a database of ca. 50,000 hieroglyphs from all over the Maya area, all time periods and different types of media and context. As there are ca. 20 units of analysis per each sign (see below), this will lead to ca. one million items of analysis.

The methodology involves the following steps:

1. Assembling a statistically significant corpus of texts
2. Creating a list of database entries
3. Cataloging and analyzing each sign individually in the database using a variety of entries (modes of analysis)
4. Utilizing the database for research purposes and for pedagogical objectives

In contrast to traditional epigraphic analysis, corpus epigraphy takes into account all conceivable variables that can have an influence, importance, and implications on the reading of individual glyphs or whole texts. An example of traditional epigraphic analysis looks as follows (see Figure 1: A2):

Transliteration:	**u-tz'a-pa-wa TUN-ni**
Transcription:	*utz'apaw tun~tuun*
Morphological segmentation:	u-tz'ap-aw-Ø tun~tuun
Morphological analysis:	3SE-insert/plant-THM-3SA stone.N
Grammatical description:	CVC-transitive verb, active voice
Translation 1:	'he/she inserted/planted it, the stone'
Translation 2:	'he/she inserted/planted the stone'
Translation 3:	"he erected a monument"

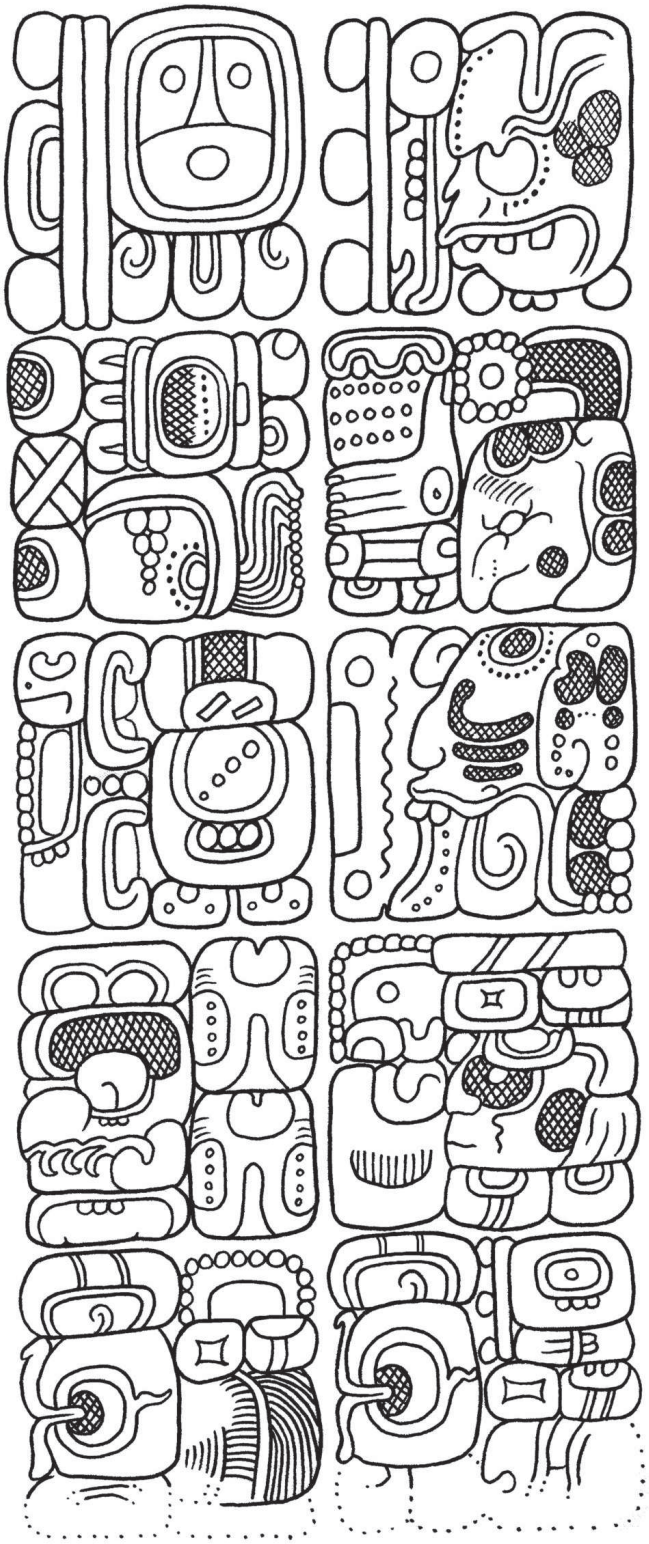

Figure 1. Stela 4 (A1-B5), Ixtutz, Guatemala (drawing by the author).

In contrast, a corpus epigraphic analysis has ca. 50 units of analysis for each text, with ca. 20 of them for each segment in the text:

Site:	Ixtutz
Code:	IXZ
Monument /artefact / context 1:	Stela 004
Monument /artefact / context 2:	Stela
Material:	Honey-colored limestone
Dimensions:	H:1.83/MW:1.16/MTh:0.40/Rel:0.01
Carved areas:	Front
Original location:	Ixtutz
Archaeological context:	W side of Strs. 9-11
Discovered by:	Ian Graham and Eric von Euw, 1972
Current location:	MNAE, Guatemala
Condition:	Pristine to poor
Publication (photo & drawing 1):	CMHI, Vol. 2:171
Publication (drawing 2):	Kettunen and Helmke 2003
Drawing 1 & 2 by:	Ian Graham and Harri Kettunen
Contemporary event (narrative date):	09.17.10.00.00
Last date (LC):	09.17.10.00.00
Last date (CR):	12 Ajaw 8 Pax
Last date (GD):	780-12-02
Approximate dedication date (k'atun interval):	09.18.00.00.00
Sequence number of hieroglyphic units:	8
Sequence number of hieroglyphic blocks:	3
Position:	A2
Transliteration:	u / u^{N}
T-number:	T0013
Description (iconic):	n.a.
Description (semantic):	n.a.
Inferred?:	n.a.
Condition:	good
C/V-construction 1:	V
C/V-construction 2:	V-
Underlying sound:	u [cf. /ni/: -n]
Morpheme:	u-
Glyph compound:	u
Part of Speech:	pronominal
Collocation:	*utz'apaw tun~tuun*
Morphological segmentation:	*u-tz'ap-aw-Ø tun~tuun*
Morphological analysis:	3SE-insert/plant-THM-3SA stone.N
Grammatical description:	CVC-transitive verb, active voice
Translation 1:	'he/she inserted/planted it, the stone'
Translation 2:	'he/she inserted/planted the stone'
Translation 3:	"he erected a monument"
Glyphs per block:	6
Sentence/clause boundary:	n.a. (vs. beginning/end of text/clause)
Sign typology 1:	independent (vs. conflated / infixed)
Sign typology 2:	phonogram / syllabogram (vs. logogram)
Sign typology 3	part of glyph compound (vs. stand-alone)
Sign typology 4:	independent (vs. phonetic complement)

All the records above are entered to the database using a computer program that is flexible enough so that the data can be exported to other programs in the future. The (partial) raw data looks as follows:

Position:	Transliteration:	Position:	Shape:	Arrangement:	Position of "affixes":	T-number:	Graphemic CV-composition:	Morpho-graphemic CV-composition:	Morpho-graph. CV-composition (surface level):	Grapheme (surface level):	Morpheme:	Glyph compound:
A2	**u**	complete	flat	vertical	left	T0013	V	V-	V-	u	u-	u
A2	**tz'a**	first	round	sideways		T0366v	CV	-CV-	-CV-	tz'a	tz'ap	tz'apaw
A2	**pa**	penultimate	round	sideways		T0586	CV	-CV-	-CV-	pa		
A2	**wa**	last	flat	vertical	bottom	T0130	CV	-CV	-C	w	-aw	
A2	**TUN**	first	round	upright		T0528	CVC	CVC-	CVC-	tun	tuun	tuun
A2	**ni**	last	flat	vertical	right	T0116	CV	-CV	-C	n		
B2a	**u**	complete	flat	horizontal	top	T0001	V	V-	V-	u	u-	u
B2a	**CHOK**	first	round	upright		T0710v	CVC	-CVC-	-CVC-	chok	chok	chokow
B2a	**ko**	penultimate	flat	horizontal	bottom	T0110	CV	-CV-	-CV-	ko		
B2a	**wa**	last	flat	horizontal	bottom	T0130	CV	-CV	-C	w	-ow	
B2b	**ch'a**	first	flat	horizontal	top	T0093	CV	CV-	CV-	ch'a	ch'aaj	ch'aaj
B2b	**ji**	last	round	upright		T0758	CV	-CV	-C	j		

Figure 2. Detail of the raw data of corpus epigraphy.

As regards the different levels of translating Maya texts (or any other texts for that matter), it is important to realize that there are different ways of translating words and sentences from one language to another, from glosses and direct translations to free prose translations. A good example is the compilation "Forty-Three Ways of Looking at Narihira" by Anthony H. Chambers (n.d.) of different translations of a single poem by the Japanese 9th century poet Ariwara no Narihira[1]. Looking only at the first line of the poem, one can detect a great variation in the translations:

Moon? There is none.
No moon!
Can it be that there is no moon?
The moon [and spring] come round as in old days.
And the light of the moon was not so serene.
Is it then true that there is no longer a moon?
Is the moon changed?
What now is real? This moon, [this spring], are altered.
Can it be that the moon has changed?

[1] I would like to thank David Carrasco for pointing out this reference to me.

Is not the moon the same?
The moon – is it gone?
The moon is not the moon of that year!
Is that moon the same?
Here, here is the moon.
Moon? Is that not you?
etc.

In the graphemic level, when dealing with database entries and relying on Latin alphabet, it is important to make a distinction between a transliteration of an utterance vs. its iconic and semantic description. Transliteration without graphemic description (or without a T-number or any other code) is problematic when working with databases. A good example is the word *chan*, which can mean 'snake', 'four' or 'sky' in Classic Maya (Fox and Justeson 1984; Houston 1984). Without a graphemic description, the database is limited and defective.

In the case of Ixtutz Stela 4: A4a (see Figure 1), the transliteration of the ophidian sign is **CHAN**, although a more logical and transparent transliteration would be, for example, **CHAN**$_1$ (to distinguish it from **CHAN**$_2$ 'sky'). Another way to make the distinction is to add icon(ograph)ic descriptions to the database, such as "CHAN-snake" vs. "CHAN-sky". Furthermore, it is necessary to include a semantic description to the database, as logograms are not pictograms. Consequently, in the case of the snake head on Ixtutz Stela 4: A4a, the semantic description would be "CHAN-guardian". It is also noteworthy that dealing with late monuments such as this one, the vowel length or vowel complexity is not present any longer and, consequently, the spelling of 'snake' and 'guardian' is identical, both using the same phonetic complement **na**.

As regards phonetic complements, another outcome of corpus epigraphy is the fact that it reveals details of the rules governing phonetic complementation. Based on statistics, the use of phonetic complements was very restricted. For example, pre-logographic phonetic complementation is extremely rare and with some sounds (especially plosives/stops) almost non-existent (except when marking dialect variance). Consequently, it is also important for students in Maya epigraphy to realize that examples of the workings of the script in the scholarly literature, such as **ba-BALAM** are imaginary examples that do not appear in the corpus at all, which is to say that they were never part of the scribal practices of the ancient Maya.

Yet another advantage of corpus epigraphy is that it exposes regional and temporal patterns in the writing system. It also informs students what to expect and what not to expect before or after a given sign in particular areas and/or specific time periods, including exposing patterns on partially eroded monuments.

CORPUS EPIGRAPHY: APPLICATIONS

Besides the regional and temporal patterns in Maya texts, corpus epigraphy can shed light on the harmony/disharmony discussion in epigraphy. Based on the patterns in the corpus, it appears that some of the features of the quality of vowels of certain words are conditioned by the tradition of using specific types of signs as final CV-phonograms. Consequently, the harmony rules put forth by Houston, Stuart, and Robertson (1998), Houston, Robertson, and Stuart (2000) as well as Lacadena and Wichmann (2004) are not to be taken as a dogma when studying the system. For example, based on the rules by Lacadena and Wichmann (2004), **JUN-na** should produce *ju'n* (CV1C / CV1-CV2 > CV'(V)C [V1 = e, o, u; V2 = a]). However, there is no indication in Mayan languages that this is indeed the case. Another example is the morphology of active transitive CVC-verbs: ERG-CVC-

V1w-ABS, where the thematic suffix –Vw represents a vowel resonating the vowel of the verbal root, as in *u-chok-ow* ('he/she threw it') and *u-tz'ap-aw* ('he/she inserted/planted it'). In all the cases, the *graphemic* suffix is the **wa**-phonogram that undoubtedly simply cues the –Vw rather than –V'w thematic suffix for active transitive constructions.

Also, it is worth noticing that ancient scribes were – and modern epigraphers are – faced with a challenge due to the absence of certain phonograms in the syllabary. It appears as if the Maya scribes only employed a limited set of final phonograms without specifically indicating complexity in the root vowel (or any preceding vowel). Statistically, these final phonograms tend to take primarily /a/, /i/, or /e/ vowels (–Va, –Vi , and –Ve), and particularly the first two, with /o/ and /u/ (–Vo and –Vu), being infrequent. Consequently, it seems that disharmonic spelling by itself does not necessarily denote vowel complexity, and nor does synharmonic spelling always indicate short vowels. Also, it is important to indicate in each case *which phonogram* we are dealing with. Working with Latin alphabet when analyzing Maya texts is always problematic. As such it is necessary to indicate which Maya sign we are talking about in each case, instead of merely referencing to, for example, a phonogram **na**. As a whole, the Maya writing system is not a sterile and mechanical apparatus (no more than any other writing system in the world) and it should not be forced to fit a fixed pattern of linguistic theory.

Besides linguistics, the statistical analyses can also be used for didactic purposes. For example, frequency lists of the most common hieroglyphs (see Kettunen 2010) can be utilized when teaching Maya epigraphy to students. It is, for instance, important to note that a handful of the most common hieroglyphs also make up the largest volume of all existing examples of hieroglyphs in the corpus. The same goes with many other writing systems, including Japanese where ca. 20% of the kanji constitute ca. 80% of everything that is written in Japanese (Japanese Agency for Cultural Affairs 1981). This rule, also known as the *Pareto principle*, can be applied to teaching and learning the Maya script: just as in Japanese, it makes sense to learn the most common kanji or logograms first.

As regards the frequency of signs in the corpus, it comes not as a surprise that the most common logogram in the Maya hieroglyphic writing is **AJAW** 'lord' followed by **K'UH**/**K'UHUL** 'god/godly' and a plethora of calendrical signs. When it comes to phonograms, the most frequent sign by far is **u** – again unsurprisingly – as most of the Maya texts were written in third person singular. Based on the initial analysis of 15,000 signs, the next most frequent phonograms in the corpus of monumental inscriptions are **ya**, **wa**, **ni**, **la**, **na**, **a**, **ti**, **ji**, **ja**, **ma**, **li**, **ki**, **ba**, **ta**, and **ka**, with some variance as to the area and time period of the monuments. The order of the frequency list will obviously alter as new monuments are added to the database. Consequently, this list is to be understood only as a preliminary inventory of the signs, comprising merely 30% of the expected final inventory. The explanation for the frequency of each phonogram has to be analyzed carefully, as in some cases the rationale lies behind the frequency of sounds in the Classic Maya language, in some cases in the frequent use as part of grammatical affixes, and in many cases also as phonetic complements. Furthermore, the motivation for using certain phonograms in the final position of written words may actually stem from scribal practices (see below).

From a didactic point of view the list of most common phonograms (and logograms) is important, as it is the most common signs that one should start to learn and identify as a beginning student in epigraphy. Here, however, it is important to note that the concept of the most common sign in the writing system is not the same as the concept of the most common sound in the writing system. Consequently, frequent phonograms in the writing system, **ya**, **wa**, **ni**, **la**, **na**, **ji**, **ja**, and **ma** are regularly written by using only one graphic variant when marking grammatical suffixes, whereas the phonogram **u**, a pronominal affix marking subjects of transitive verbs as well as working as a possessive pronoun, exhibits a profusion of graphic variance. Therefore, as the saying in epigraphic studies goes, "when in doubt, it's an **u**". The same goes with logograms, where the most common word can be written in very uncommon ways.

However, analyzing the statistics becomes more interesting when we look at the distribution of phrases, words, and morphemes – rather than merely logograms and phonograms – across the Maya area in time and space. As an example, one can study the distribution of verbs or verbal phrases in the corpus and identify the most common verbs in different areas, time periods, media, and objects/artefacts.

Based on a sample data comprised of 410 monuments from 56 sites, yielding 92 different verbs (=types) out of 2013 instances of verbs (=tokens), by far the most common verbal root in the corpus is, unsurprisingly, *uht-* 'to happen'. However if all the verbs in calendrical contexts are excluded (with *uht-*, *tz'ak-*, *k'al-*, and *hul-* leading the ranking list), we get a clearer picture of the actions that the ancient Maya inscribed in their monuments, namely *k'al-* 'to present/bind', *sih-*/*siy-* 'to be born', *ux-* 'to carve', *uht-* 'to happen', *chum-* 'to sit down (into rulership)', *chok-* 'to scatter', *chuk-* 'to seize/capture', *il-* 'to see', *tz'ap-* 'to plant/insert (monuments)', *joy-* 'to bind/reveal', *och-* 'to enter', *?* 'starwar', 'to wage war', *ahk'-* 'to dance', *hul-* 'to arrive', and *cham-/kam-* 'to die'. With more data from additional monuments, the statistics will evidently vary. As they unquestionably will if texts on portable artefacts, such as ceramic vessels and codices, are included in the data set. Consequently, it is important to study the Maya texts based on *genre* as well, instead of looking at the writing systems as a monolith.

Nonetheless, examining the whole corpus yields noteworthy statistical observations. Interestingly, regarding the distribution of verbs, the percentage of the 20 most common verbs in the sample corpus discussed above is 81.7%, i.e. the 20 most common verbs make up ca. 80% of all occurrences of the verbs in the data set, following quite ideally the Pareto principle mentioned above.

Also, as pointed out by Juan Ignacio Cases (personal communication 2013), one should make a distinction between *readable graphemes* and *deciphered types* in the script. While the percentage of *readable graphemes* in the corpus is ca. 95%, the percentage of *deciphered types* is as low as 65%. In practice this means that when an epigrapher is asked to give an estimate as to the percentage of deciphered hieroglyphs in the system, the answer depends on what is meant by "deciphered" and "percentage of". We do not even need to go as far as to question whether we really understand what the ancient Maya meant by each word and phrase, but merely to examine the "of" of "percentage *of*". If we examine the volume of Maya texts, we can say that ca. 95% of the texts have been deciphered to a satisfactorily degree. However, if we examine each sign as a *type*, disregarding the occurrences of each sign in the corpus, the percentage of deciphered signs as *types* rather than *tokens* drops down to 65% – or even lower. As pointed out by Cases, the undeciphered (or only partly deciphered) signs consist predominantly of *hapax grammena* or *dis grammena* signs (i.e. signs appearing only once or twice in the corpus, also referred to in the literature as *hapax legomena* and *dis legomena*, respectively, when referring to words rather than signs), as well as semantically transparent but phonetically opaque signs.

Interestingly, similar pattern applies to other writing systems and languages, including English where 10 of the most common words (or word forms, *lemmas*) constitute 25% of the entire corpus of English language, while 100 most common words provide 50%, 1000 comprise 75%, 7000 cover 90%, etc. (The Oxford English Corpus). It is also interesting to note that while over 50% of English words are of Latinate (Latin and or French) origin, nearly all of the 100 and the majority of the 1000 *most common* words in the English language are of Germanic origin.

CONCLUSION

To conclude, the importance of corpus epigraphy is twofold: (1) it generates a profuse amount of linguistic and epigraphic data and analyses, including the recognition of scribal tradition and its linguistic implications, as well as (2) provides students with tools to embark upon deciphering the

script. The foundation of corpus epigraphy is an extensive database that functions as a flexible raw data that can be modified as the understanding of the writing system progresses. Finally, the advantage of corpus epigraphy is that it is not restricted to Maya hieroglyphic writing: methods used in corpus epigraphy can be applied to any writing system in the world – or, when modified, to any other system devised by human beings.

REFERENCES

CHAMBERS, ANTHONY H.

n.d. *Forty-Three Ways of Looking at Narihira*. URL: <http://www.public.asu.edu/~achamber/narihira.html> [Accessed: January 1st, 2013]

FOX, JAMES A. AND JOHN S. JUSTESON

1984 Polyvalence in Mayan Hieroglyphic Writing. *Phoneticism in Maya Hieroglyphic Writing*, edited by John S. Justeson and Lyle Campbell, pp. 17-76. Institute for Mesoamerican Studies Publication 9. Albany: State University of New York at Albany.

HEISENBERG, WERNER

1958 *Physics and Philosophy: The Revolution in Modern Science*. New York: Harper & Brothers Publishers.

HOUSTON, STEPHEN D.

1984 An example of homophony in Maya script. *American Antiquity* 49 (4): 790-805

HOUSTON, STEPHEN, DAVID STUART AND JOHN ROBERTSON

1998 Disharmony in Maya Hieroglyphic Writing: Linguistic Change and Continuity in Classic Society. *Anatomía de una civilización: Aproximaciones interdisciplinarias a la cultura maya*, edited by Andrés Ciudad Ruiz, Yolanda Fernández, José Miguel García Campillo, Maria Joseja Iglesias Ponce de León, Alfonso Lacadena García-Gallo and Luis T. Saenz Castro, pp. 275-296. Publicación No. 4. Madrid: Sociedad Española de Estudios Mayas.

HOUSTON, STEPHEN, JOHN ROBERTSON AND DAVID STUART

2000 The Language of Classic Maya Inscriptions. *Current Anthropology* 41 (3): 321-356.

JAPANESE AGENCY FOR CULTURAL AFFAIRS

1981 常用漢字表 [*Jōyō Kanji Hyō*]. <http://www.bunka.go.jp/kokugo_nihongo/pdf/jouyoukanjihyou_h22.pdf> [Accessed: February, 2013]

KETTUNEN, HARRI

2006 *Nasal Motifs in Maya Iconography: A Methodological Approach to the Study of Ancient Maya Art*. Helsinki: Annales Academiæ Scientiarum Fennicæ.

2010 *Frequency Chart of Logograms in Maya Hieroglyphic Writing*. Materials for the Advanced Workshop on Maya Hieroglyphic Writing. Helsinki: Department of World Cultures, University of Helsinki.

LACADENA ALFONSO AND SØREN WICHMANN

2004 On the Representation of the Glottal Stop in Maya Writing. *The Linguistics of Maya Writing*, edited by Søren Wichmann, pp. 103-162. Salt Lake City: University of Utah Press.

THE OXFORD ENGLISH CORPUS

n.d. Oxford: Oxford University Press.

PROSKOURIAKOFF, TATIANA

1950 *A Study of Classic Maya Sculpture*. Publication No. 593. Washington, D.C.: Carnegie Institution of Washington.