ARTÍCULOS

Revista Española de **Antropología Americana**
ISSN: 0556-6533

EDICIONES COMPLUTENSE

https://doi.org/10.5209/reaa.64967

A Computational Approach to Scribal Practice

Ignacio Cases Martín[1] and Alfonso Lacadena García-Gallo[2]†

Recibido: 4 de febrero de 2019 / Aceptado: 4 de marzo de 2019

Abstract. The study of the construction of social meaning in ancient Maya communities of Mesoamerica poses a variety of methodological problems in historical sociolinguistics due to the reliance on written records by means of a writing system that exhibits variation itself. While variation in writing systems has been previously studied in terms of diachronic shifts and dialectal variation, systematic approaches still remain elusive. This paper explores new avenues for the computational extraction of sociolinguistic features, resulting in the automatic extraction of useful sociolinguistic information from written corpora using Machine Learning algorithms. We show that these features can help illuminating the contribution of pragmatic choices in the selection of graphemes to stylistic practices that are key in the construction of Mayan scribal communities of practice.

Keywords: scribal practice; computational sociolinguistics; Maya writing; natural language processing.

[es] Aproximaciones computacionales al estudio de las prácticas escriturarias

Resumen. El estudio de la construcción del significado social en las antiguas comunidades mayas de Mesoamérica plantea una variedad de problemas metodológicos en sociolingüística histórica, en gran medida debido a la dependencia de un sistema de escritura que, en sí mismo, presenta variaciones. Si bien estas variaciones en el sistema escriturario se han estudiado previamente en términos diacrónicos y dialectales, aún carecemos de enfoques cuantitativos sistemáticos. Este artículo explora nuevas vías de aproximación computacional a ciertos rasgos de origen sociolingüístico, lo que permite la extracción automática de información sociolingüística utilizando algoritmos de aprendizaje automático. Estas características pueden ayudar a iluminar la contribución de las elecciones pragmáticas en la selección de grafemas a las prácticas estilísticas que son clave en la construcción de comunidades mayas de práctica escrituraria.

Palabras clave: práctica escrituraria; sociolingüística computacional; escritura maya; procesado del lenguaje natural.

How to cite: Cases, Ignacio, and Alfonso Lacadena. 2019. «A Computational Approach to Scribal Practice». *Revista Española de Antropología Americana* 49 (número especial): 209-224.

[1] Linguistics Department and Stanford NLP Group (Artificial Intelligence Lab), Stanford University. cases@stanford.edu

[2] Departamento de Historia de América y Medieval y Ciencias Historiográficas. Universidad Complutense de Madrid.

1. Introduction

The analysis of the construction of social meaning in ancient Maya communities of Mesoamerica rests on the understanding of the social meaning of material signs, discourses, and actions that have to be adequately situated in their context of the physical world (Geertz 1973; Scollon and Scollon 2003; Houston 2004). Such reliance on written records pose a variety of methodological problems in the historical sociolinguistics of Mesoamerican languages, only incremented by the difficulty of accessing to traditional sociolinguistic variables. Only in a few cases it is possible to categorize an approximate age, sex, and status of a writer. In addition, the writing system used to encode language exhibits a complex relation with variation in the host language due to an array of sociolinguistic factors. Although there has been interest in the interaction between writing and language among sociolinguistics since the early work by Weinreich (1953), this relation has been studied mainly in terms of diachronic change and dialectal variation (Weinreich 1953; Labov 1972a, *inter alia*). In this paper, we explore the contribution of pragmatic choices in the selection of graphemes to stylistic practices that are key in the construction of scribal communities of practice (Justeson 1978). In particular, we follow up on recent studies addressing change and variation in both the encoded languages and their writing systems in Mesoamerica (e.g. Lacadena and Wichmann 2002) by means of a systematic analysis of a proxy corpus of Classic Maya inscriptions.

2. Sociolinguistic Variation in Classic Maya Writing

Language variation and change was reflected in Classic Maya writing system, and result in an important source of linguistic information for the reconstruction of ancestral varieties of several modern Mayan languages (Lacadena and Wichmann 2002). However, systematic sociolinguistic analysis of variation and change as observed from variation and change in the writing system is still to be developed, essentially for two reasons. First and foremost, the unavailability of comprehensive, large corpora of Classic Maya, associated to a general lack of appropriate tools for analysis for even small corpora. Second, the partially deciphered writing system maintains a complex relation with the languages represented in the texts, as seen above. In our opinion, the complex relation between Classic Maya writing system and the hosted languages can be better addressed by using the same methodological approaches traditionally used in sociolinguistics (Labov 1972b; Shibamoto Smith and Schmidt 1996, *inter alia*). As such, variation and change in both the languages and the writing system can be productively analyzed in sociolinguistic terms by introducing relevant linguistic and social features Labov (1972b).

The emerging picture during the Classic Period is that of a diverse linguistic area in which several language vernaculars are in contact and percolate to different degrees into the high-prestige language of the inscriptions (Houston et al. 2000; Lacadena and Wichmann 2002). For the most part of the corpora the determination of social features is difficult. Access to traditional social variables such as age, sex, and status of a writer is partial in most cases. However, it is also possible to consider most of the additional types of congruence introduced by Weinreich as archetypical of languages in contact (Weinreich 1953). Therefore, variables such as geographic

areas, ethnicity, cultural or ethnic groups, religion, occupation, and rural vs. urban population, for which access is less of a problem, emerge as important candidates to understand the dynamics between linguistic varieties.

The bilingual individuals constituting the locus of study in language contact (Weinreich 1953) are to be found in our setting as bilingual scribes that form part of a scribal community of practice—workshops that enforces normativity through the establishment of grammatical rules and sets of prescribed scribal practices. A central question to address is, then, what are the possible features, linguistic and graphemic, that enable us the identification of communities of scribal practice. Classic Maya corpus shows a wealth of information of sociopolitical actors and their interactions (Martin and Grube 2000), resulting in dense data that enables social network analysis. Once these communities are determined, it could be possible to contrast the sociolinguistic differences between variable communities of practice (Milroy and Margrain 1980), and examine the relation between these communities and the sociopolitical landscape.

The mechanisms and structural causes of transfer at the graphemic level are to be traced back to linguistic variation and change. The following sections discuss, following the layout established by Weinreich (1953), some of the preliminary findings in the literature and also found during the early stages of corpus construction regarding linguistic features that index variation and change. Then, new features at the graphemic level are introduced and analyzed in a synchronic case study.

2.1. Diachronic variation

Language change in the diachronic axis (Labov 1972a) is reflected in the phonological, morphological, and semantic levels. By the end of the Classic Period, around 900 A.D., Classic Ch'olan experienced a series of phonological changes, in a series of changes that could have happened in short period of time (cf. Trudgill 2002). Long vowels shortened and glottalized vowels disappeared (Lacadena and Wichmann 2004). The distinction between velar and glottal fricatives, once predicted by Kaufman and Norman (1984 [1978]) in their initial reconstruction of Proto-Cholan and found in early Classic Cholan by Grube (2004), also vanished by the end of the Classic Period. Houston et al. (2000) have documented a shift of *-h - ... - aj* from intransitive positional marker to passive, with the old passive marker *-V_1y* becoming a marker of the mediopassive. After that process, a positional *-wan* marker is introduced as an innovation, probably after a process of percolation from the vernacular Ch'olan in the Tabasco region into the high-prestige variety. Lexical change is more difficult to track without the availability of computational methods, but some examples can be recognized: early logograms, probably borrowed from close-by written traditions such as the Epi-Olmec, changed their lexical values while presumably keeping the same semantic reference. This seems to be the case of the word for sun or sun god, initially rendered using the term **JAMA** borrowed from Mixe-Sokean languages, and used only in early stages of Maya writing. Later, the same logogram carries the Mayan word *K'IN*, *sun* or *sun god*. Semantic changes are also difficult to trace. A possible example involves the adjective *k'uh*, which is originally taken to mean sacred but experiences a shift in the semantic space towards venerable by the time of the Spanish contact.

2.2. Synchronic variation

Synchronic variation in Mayan languages during different stages in the Classic Period points towards the existence of a plurality of spoken vernaculars that sometimes percolated into the written high-prestige variety. Although it follows the paradigmatic situation of diglossia introduced by Ferguson 1959), there are several questions to be answered related to the manifestation of the interaction between the different vernaculars in the contact situation, and their interaction with the high prestige variety. Phonemic variation shows a differential phono-logical system between southern (Ch'olan) and northern (Yucatecan) varieties. Among the numerous examples we can mention the spelling of the number 4 in a series of texts from Ek B'alam, in the north of the Yucatan peninsula, using the Yucatecan phonology as **ka-na**, *kan*, instead of the attested Ch'olan version *chan* registered in southern sites. The abstractive suffix *-il* appears to derive abstract nouns from concrete nouns, such as *'ajawil*, 'kingdom', from *'ajaw*, 'king', in most of the Southern Maya Lowlands. However, the same suffix is replaced by the abstractive *-lel* in Western Maya Lowlands, home to modern day Ch'ol and Chontal that retain reflexes of that suffix (Lacadena and Wichmann 2002). In the Northern Maya Lowlands of Yucatan, the abstractive attested is *-lil*, which is the ancestor of the abstractives found in Yucatecan languages. Lexical variation seems to be less common, but there are some unequivocal cases such as the word for month, *winik* in Western Lowland Mayan, and *winal* in Eastern varieties.

3. A Proxy Corpus for the Analysis of Mayan Scribal Practice

The analysis of scribal practices requires the construction of appropriate corpora of Classic Mayan texts (Cases et al. 2014). This process involves the designing of the corpus, the collection of data, the encoding in machine-readable format, and a proper assemblage and storage of relevant metadata, including linguistic and epigraphic annotation (McEnery and Hardie 2012).

The question of what exactly is an appropriate corpus is important, and it has different answers depending on the overall objectives of the research. To the bare minimum, a well designed corpus for research should avoid the confirmation bias, i.e. a design that favors the interpreter's initial hypothesis or beliefs (McEnery and Hardie 2012: 14). The confirmation bias can be avoided using the principle of total accountability, which states that the researcher must not select a favorable subset of the data.

A long term objective aims for the construction of a monitor corpus of Maya writing, including all the glyphic texts from the Classic and Post-Classic periods, as well as Colonial and Post-Colonial alphabetic texts. The medium term objective, and the aim in this paper, has been the creation of a balanced corpus of Classic Maya texts with specific attention to diachronic balancedness, genre and provenience. The short term objective considered the development of an opportunistic corpus with a manageable size, but enough large as to obtain relevant data for a given area. The area selected was the Western Maya Lowlands, with texts from the sites of Palenque (PAL) and Comalcalco, Tabasco (CML), covering a small variety of genres in a span

of 200 years in the Late Classic Period. The results in the following sections have been obtained using this proxy corpus.

In order to illustrate the procedure used in the construction of the corpus, we will use a text example from Comalcalco. In a remarkable discovery in 1998, archaeologist Ricardo Armijo-Torres found a sealed urn in the platform between Temples II and II-A that are situated in the Main Plaza. The archaeological contextualization shows that the urn contained the burial of a male individual, accompanied by 74 beads of jade, 52 shark's teeth, a series of ornamental shells, prismatic obsidians, eccentric flints, specular hematite counters, seven stingray needles, sixteen stingray spines with hieroglyphs carved, and 82 small pendants, 36 of which were carved with glyphs (Armijo 1999; Armijo, Gallegos y Zender 2000; Armijo, Zender y Gallegos 2000). Other organic remains of the interior could have been due to the presence of leather, opening the possibility of the material to have functioned as a bag for the rest of artifacts (Armijo, personal communication to Cases, 2006).

One of the texts from this urn is the stingray spine number 3, technically referred to as CML Urna 26 Spine 3. The epigraphic rendition of the text from the Spine 3 is shown in Figure 1. An epigraphic contextualization includes the transcription, transliteration, and morphological segmentation as follows:

13 'AJAW 18-HUL-'OL-la CHUM-TUN-ni 'u-17-WINIKHAB' wa– 'i-ja K'IN-TUN-ni wa-'i-ja WI'-na-li tu-13-TUN-ni

13 'Ajaw 18 Hul-'O'hl chumtuun
'u 'uuklaju'n winikhaab'
wa'iij k'intuun wa'iij wi'naal
tu['] 'uuxlaju'n tuun

13 'Ajaw 18 Hul-'O'hl chum-tuun-ø
'u 'uuklaju'n-winik-haab'-ø
wa'-iij-ø k'intuun wa'-iij-ø wi'naal
tu'-'uxlajun-tuun

NUM(13)-NOUN('Ajaw) NUM(18)-NOUN(Hul-'O'hl) VERB(seat)-NOUN(stone)-ABS(3s)
ERG(3s) NUM(17)-NOUN(year)-ABS(3s)
PART NOUN(drought)-NOUN(period of time)-ABS(3s) PART NOUN(famine)-ABS(3s)
PREP(on)-NUM(13)-NOUN(year)

On 13 'Ajaw, 18 Hul 'O'hl,
(it is) the seating of the *tuun*, the 17th winikhaab',
will there be (a period of) drought, will there be famine,
on the 13th *tuun*.

These components constitute a wireframing for the first set of contextualizations at the archaeological and epigraphical level, necessary to perform historical sociolinguistic analysis, and serve to illustrate the procedure used in the corpus construction.

```
/*!
  @site: CML
  @mon: Urn 26
  @object: Spine
  @objectOther: 3
  @facture:
  @collation: ic
  @colVersion: 0.1
  @since: 20/6/2012
  @notes:
  @references: Armijo-Gallegos and Zender (2002)
  @imageFile:
  @drawingAuthor: Marc U. Zender
  @textDisposition: column
  @textDimensions: 1x13
 */
A1: 13-?.#AJAW
A2: 18-HUL-'OL-la
A3: CHUM TUN-ni
A4: 'u-17-WINIK-HAB'
A5: wa-['i]ja [K'IN]TUN-ni
A6: wa-['i]ja WI'-na-li
A7: tu-13-TUN-ni
pA8: hi-HIX-li
pA9: 'a-'AJAW-wa
pA10: 'u-?-ji
pA11: ya-'o-la
pA12: ti-su-tz'i-li
pA13: ti-CHUM[mu]
```

Figure 1. Machine readable epigraphic transcription for the Spine 3 from Comalcalco Urn 26. Machine readable transcriptions like this one used in this work contain two types of content: the epigraphic transcription itself (in this case using a format similar to Lacadena and Wichmann 2004 with the conventions in Lacadena and Cases 2010), and meta-information in the header, including the provenance of the text, details of the material artifact, manufacture date in Long Count, details about the collation of the transcription, among others. The transcription itself also contains meta-information regarding the position of the graphemes in the text and specifics to the grapheme.

4. Identifying Scribal Practices

In Classic Maya writing, graphemic types are usually combined in glyphic blocks following a series of internal rules that restricted the permissibility of such combinations, resulting in observable patterns in the written texts. Analysis of these patterns, that could be termed graphotactics, has been the object of study since the beginning of the decipherment of the writing system. Most of the early work focused primarily in the arrangement of signs inside glyphic blocks (e.g., Thompson 1950, 1962; Kelley 1976; Justeson 1978; Grube 1990; Lacadena 1995) and graphemic chains, resulting in spelling rules (Kelley 1976; Justeson 1978; Bricker 1986; Grube 1990; Houston et al. 1998; Kaufman with Justeson 2003; Lacadena and Wichmann 2004). The constraints inside a graphemic chain or block will be referred to as short range graphotactics. A medium range graphotactics would consider graphemic restrictions inside a given text. This section briefly considers the evaluation of medium range

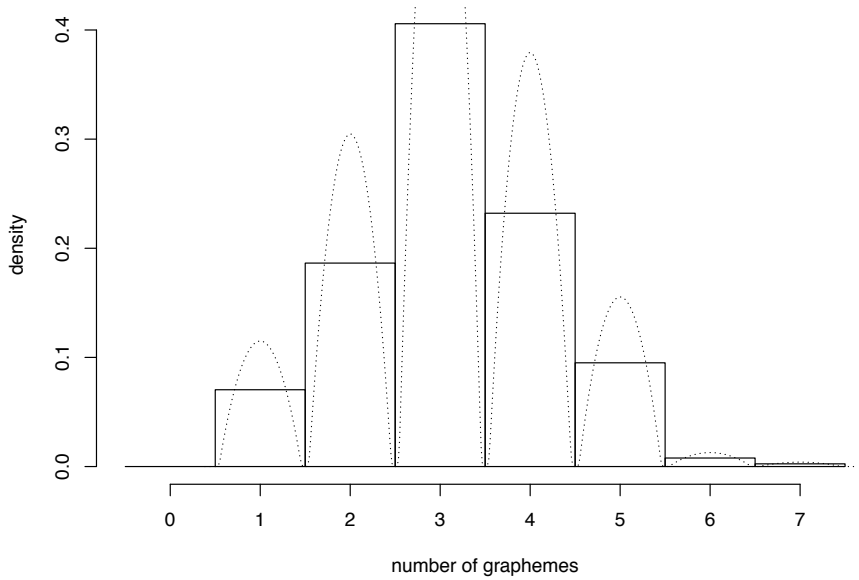


Figure 2. Distribution of the graphemic size of chains across the corpus, i.e., the normalized number of examples versus the size of each example in terms of number of graphemes. The distribution shows that in this corpora the glyphic blocks are most frequently composed out of three graphemes, followed by blocks composed of four graphemes, etcetera. However, as useful as this can be as an initial description of the text, this type of distribution merges all types of graphemes and is mostly a representation of short-range selection of the graphemic types —a selectional restriction on the types like this can only serve as a simple metric for downstream tasks. The dotted line red is the result of applying a simple model (adjustment with Epanechikov's kernel).

statistical graphotactics, i.e., in the analysis of possible constraints in graphemic selection as resulted from statistical data that can eventually serve as graphemic features for a sociolinguistic analysis.

The selectional restriction of graphemic types in short range graphotactics results from the scribal practice of applying a series of spelling rules with the object of representing an utterance of the underlying language, and a series of compositional rules constraining their graphic arrangement. These constraints have been studied by Justeson (1978). The rules were probably learned as part of the scribal training in the high-prestige variety inside the community of practice (Ferguson 1959).

A question to be raised is the size of the range, measured in number of graphemes for example, to which these rules would apply. Spelling rules range of action seems to be limited to the extent of graphemic chains, whether they create constraints in the nucleus or suffix domains. Figure 2 shows the distribution of the number of graphemes per graphemic chain, together with the adjustment using Epanechnikov's kernel. The distribution is centered in three graphemes per graphemic chain, that could belong to any type of sign. This type of representation merges all kind of signs for all the graphemic chains. This information can be analyzed by taking into account the sign type for known graphemes with reading value, i.e. either logographic or syllabic.

Figure 3 represents the distribution of syllabic versus logographic signs for all the texts in the corpus, where the radius of the circles are proportional to the relative frequency. Therefore, it can be seen that combinations with one logogram and two phonograms are the most common, followed by graphemic chains with three syllables and no logograms, observations compatible with the distribution mentioned before. More interesting observations appear when the information is plotted taking into account parameters like provenience, authorship, or genre. Figures 4, 5 and 6 portrait the frequencies of combinations in the syllabic-logographic plane for the texts produced by the scribe workshops of Kan B'ahlam, K'an Joy Chitam, Ahku'l Mo' Naahb' (K'uk' B'ahlam in Palenque, and Aj Pakal Tahn from Comalcalco not shown). In the case of Kan B'ahlam, the texts are represented adjacent each other in function of the artifact, that ultimately is closely related to both size and genre. It is notable that the plots for the panels from the Temple of the Inscription show similar patterns, with higher frequencies of graphemic chains with highly frequent three phonogram graphemic chains, and one logogram and two phonograms graphemic chains. The main texts from the Group of the Cross are also remarkably similar,

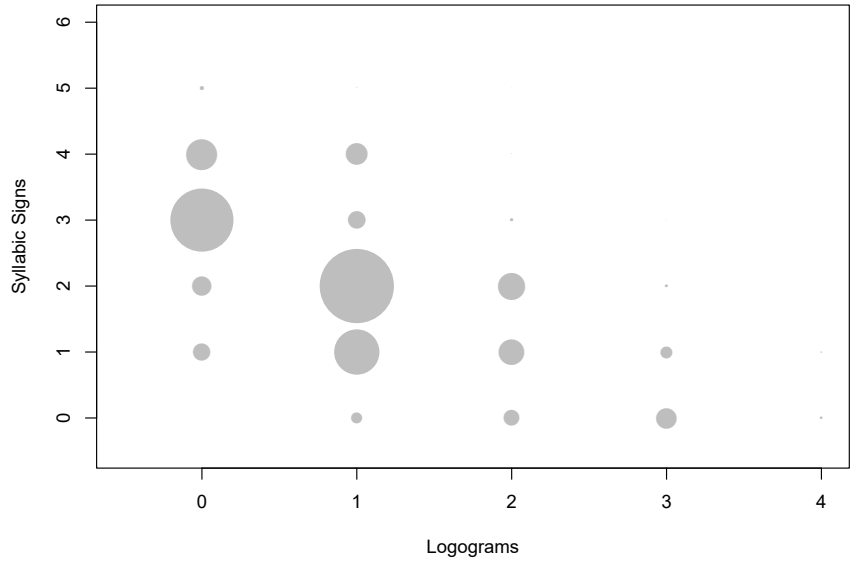
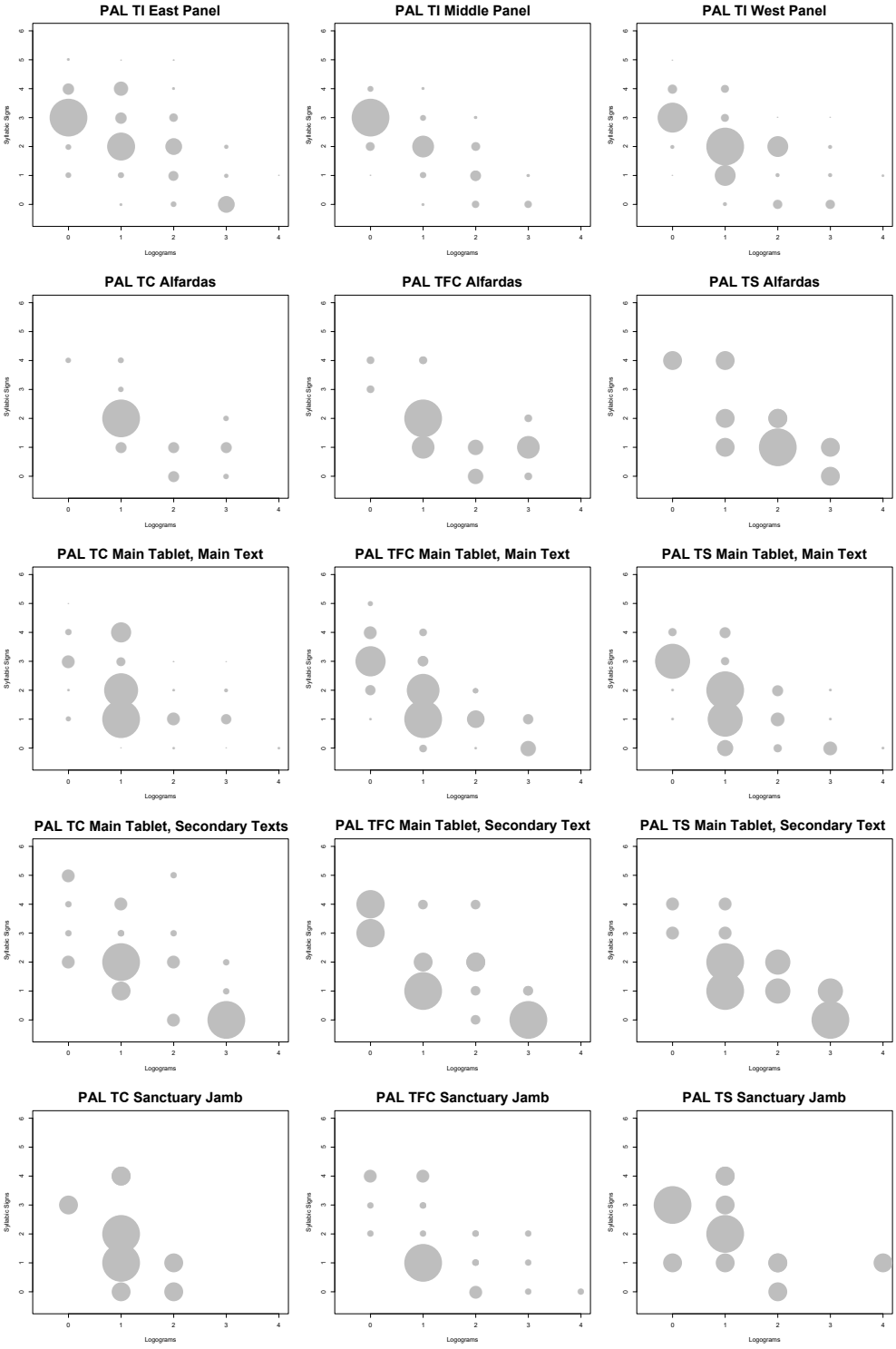


Figure 3. Distribution of syllabic versus logographic signs for all the texts in the corpus, where the radius of the circles are proportional to the relative frequency. The bubbles with larger radius correspond to graphemic chains composed out of one logogram and two phonograms (1 in the logogram axis, 2 in the phonogram axis). Graphemic chains with three syllables and no logograms are next in terms of frequencies (0 logograms and 3 syllabic signs).

Figure 4 (next page). Distribution of graphemic chains in the syllabic-logographic plane (syllabic versus logographic graphemes) for the texts produced by the scribe workshops of K'inich Kan B'ahlam from Palenque. The texts are arranged adjacent each other in function of the textual artifact, which is closely related to both text size and the literary genre. Note how graphemic chains consisting of one logogram and one and two phonograms have higher frequencies.


PAL TI East Panel
PAL TI Middle Panel
PAL TI West Panel
PAL TC Alfardas
PAL TFC Alfardas
PAL TS Alfardas
PAL TC Main Tablet, Main Text
PAL TFC Main Tablet, Main Text
PAL TS Main Tablet, Main Text
PAL TC Main Tablet, Secondary Texts
PAL TFC Main Tablet, Secondary Text
PAL TS Main Tablet, Secondary Text
PAL TC Sanctuary Jamb
PAL TFC Sanctuary Jamb
PAL TS Sanctuary Jamb
Syllabic Signs
Logograms

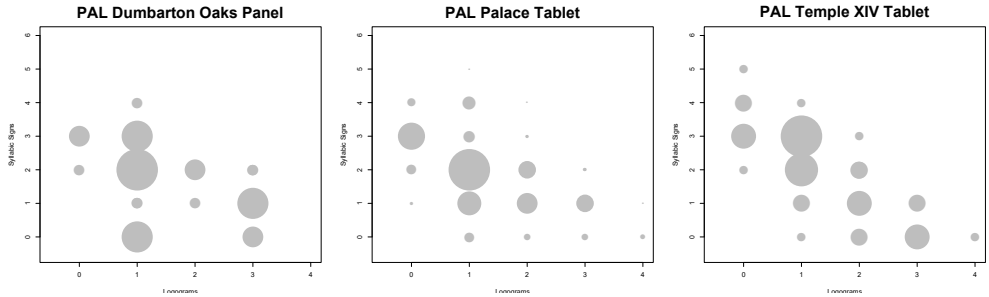


Figure 5. Distribution of graphemic chains in the syllabic-logographic plane (syllabic versus logographic graphemes) for the texts of K'inich K'an Joy Chitam from Palenque, arranged by artifact kind. The text distributions from this workshop are similar to the distributions from K'inich Kan B'ahlam, suggesting some sort of continuity in the scribal practices.

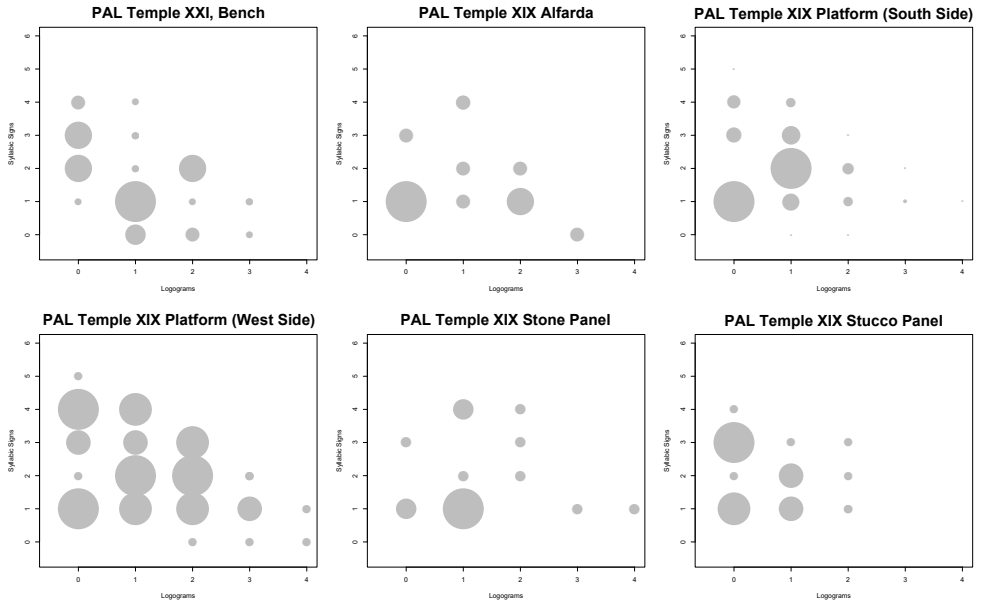


Figure 6. Distribution of graphemic chains in the syllabic-logographic plane for the texts of K'inich Ahku'l Mo' Naahb' from Palenque, arranged by artifact kind. The text distributions from this workshop, specially the remarkable Temple XIXth platform, have higher weights in the lower parts of the diagrams. This suggests a change in the scribal practice that favored a more logographic writing in clear departure from the previous tradition. This change in practice could be correlated to more profound changes brought by K'inich Ahku'l Mo' Naahb' (Stuart 2005).

using graphemic chains of one logogram and one and two phonograms very frequently. Graphemic chains with three syllables increase their frequency in the main text of the Temple of the Sun when compared with the main texts from the Temple of Foliated Cross and Temple of the Cross. It is important to keep in mind, though, that

these representations depend on very short-range properties of the text, and these graphs are not easily comparable. This is one of the main reasons to look for medium range quantities in the next section.

In middle range graphotactics, the derived measurements result from quantities that are averaged across the full length of the text. Therefore, in contrast to short range, they provide better estimates for style or authorship detection. Nonetheless, it is the combination of the information obtained from both ranges what ultimately helps to identify scribal practices.

5. Graphemic Signatures

It is now interesting to introduce a quantity that will be termed *morphographicity*, referring to the ratio of logograms and phonograms over tokens calculated across the full length of a given text. As it will be shown, morphographicity can help to estimate a measure of the amount of phoneticism that, deliberately or not, the scribe chose in his practice of writing.

In a given point t in the text, where $t$indicates the number of tokens, if γ_l is the number of logograms, γ_s the number of phonograms, and γ_r is the number of other graphemes, these quantities satisfy the trivial sum

$$t = \gamma_l(t) + \gamma_s(t) + \gamma_r(t)$$

In relative terms, the relative amount of logograms and phonograms respect to the total signs up to that point can be defined as

$$M_i(t) = \frac{\gamma_i(t)}{t}$$

These quantities would sum 1 if all the signs are known, but this is not always the case due to eroded or unknown graphemes. The text T with N_T tokens is therefore partitioned as

$$N = \gamma_l^{N_T} + \gamma_s^{N_T} + \gamma_r^{N_T}$$

and therefore, the text satisfies

$$\sum_{i \in l,s,r} X_i^{N_T} = 1$$

The final values are obtained with $t = N_T$, and therefore are $M_s^{N_T}$ and $M_l^{N_T}$.

Figure 7 includes the morphographicity diagram for Aj Pakal Tahn's CML Urn 26 Spines 3. These texts start with a number of logograms higher than the number of phonetic signs, a result expected considering that the initial clauses are dates and these are usually represented with logograms for numbers and day names. At a given point in Aj Pakal Tahn's texts, the number of phonograms starts exceeding the number of logograms, indicated by the crossing of both lines. The morphographicity values then estabilize, reaching values $M_s^{N_{T_k}}$ and $M_l^{N_{T_k}}$ at the end of each of the k texts. The morphographicity diagrams for all Aj Pakal Tahn's texts (not shown) provide similar behavior: phonograms end up dominating the morphographicity. This occurs

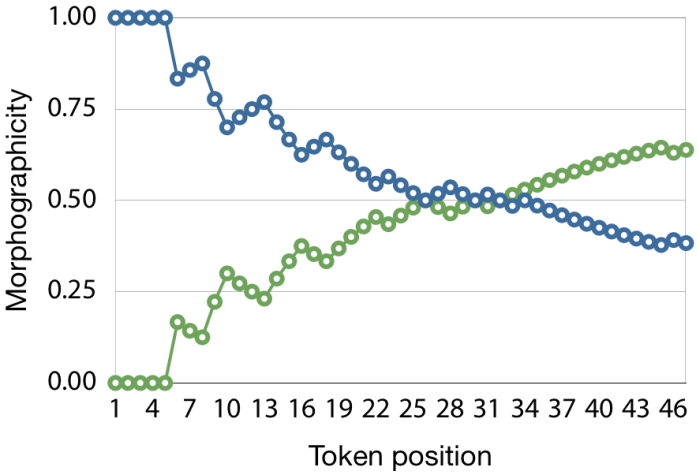


Figure 7. Distribution of graphemic types (logograms and phonograms) versus token position for Aj Pakal Tahn's Spine 3 (Comalcalco). Green lines indicate the relative amount of phonograms M_s (t) at a given position t inside the text, while blue lines represent the relative amount of logograms M_l (t) for that position. This representation makes explicit the overall ratio of types: the text start with more logograms than syllabic signs. The crossing of lines indicates that the number of phonograms starts exceeding the number of logograms.

in all Aj Pakal Tahn's texts but one: 22 texts out of the 28 start with $M_l > M_s$, and all but Spine 8 show a crossing point, therefore ending with $M_s > M_l$.

In a marked contrast with the observed behavior of Aj Pakal Tahn's texts, the subcorpus of Kan B'ahlam shows different patterns. As it appears in Figure 8, Kan B'ahlam's texts also start with a $M_l > M_s$, but the morphographicity values stabilize without crossings, and therefore end with $M_l > M_s$. This happens with only one exception, the Middle Panel from the Temple of the Inscriptions.

In order to analyze morphographicity, it is convenient to represent each text in the logographic-syllabic plane, where the texts have coordinates M_s, M_l, as it appears in Figure 8. In this morphographicity plane, those texts for which all the graphemes are known lay in the rect depicted in black. Depending on the number of graphemes eroded—a physical parameter— or graphemes with unknown reading value—an observer interference—, texts will deviate from this line of perfect accessibility into a not perfectly accessible region. In other words, from a graphemic perspective, researchers have more knowledge about those texts close to the line than in the region beneath. Fully phonetic texts will have coordinates (1,0), and fully logographic texts (0,1). Thus, it is interesting to note that K8885, a conch shell of unknown provenance, lays closely to the fully phonetic point in this plane (1, 0). An example of fully logographic text comes from an early pendant from Kaminaljuyu, where the scribe chose to write a text with no morphological marking. Most texts range in the middle region, with a notable gap between the fully logographic point and the populated area. The reason for this can be found behind the complexities of constructing a narrative with a fully-logographic text, where inflectional and derivational markers will not be possible to render in most cases.

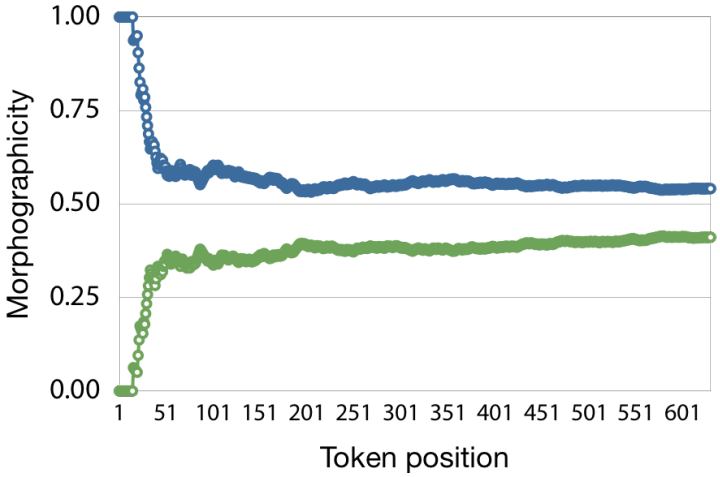


Figure 8. Distribution of graphemic types (logograms and phonograms) versus token position for Kan B'ahlam's Temple of the Cross (Palenque). The other texts from the Group of the Cross have very similar distributions, in striking contrast with the distribution of graphemic types from Aj Pakal Tahn's texts. The final values (values at the end of the text) of both curves can be plotted in a logographic-syllabic plane as presented in Figure 9.

For the corpus at hand, it can be appreciated interestingly enough two main clusters in this plane. A color plot with symbols for each author helps to identify that one cluster is formed by Aj Pakal Tahn's texts and one more from Comalcalco, around the point (60,30), while Palenque's texts cluster around (40, 50) (Figure 8).

In order to better discern the clusters, it is possible to project the texts along the orthogonal line between each point and the perfect accessibility line. This is equivalent to perform a linear extrapolation of the coordinates M_l and M_s that a given text would have in case of being perfectly accessible. After the projection, the texts in the perfect accessibility line form a linear distribution[3].

The distribution's density has two main modes, corresponding to the two main clusters from Palenque and Comalcalco scribes, and a smaller third one corresponding to K8885 in the extreme right. Obviating this last case, an unsupervised Machine Learning model using a mixture distribution of multimodal Gaussians provides the parameters for the former two distributions. These parameters fully characterize the distribution of the morphographicity of the texts.

The parameters show that Aj Pakal Tahn distribution model is centered in a syllabicity of around 69%, and therefore a logographicity of 31%, with a deviation of only 7.5%. Palenque scribes are centered in a syllabicity of around 47%, logographicity of 53%, with also a deviation of 7.5%. Insofar these distributions model morphographicity, the values of their parameters are determined by a combination of factors including linguistic features—prominently genre and topic—, graphemic features, including functional constraints, graphemic style and authorship, or in other words,

[3] Essentially this is achieved by an affine transformation composed of a clockwise rotation of 45 degrees with center the origin and a scale to the range [0,1].

scribal practice. Consequently, these distributions are considered here as the *scribal graphemic signature*, understanding the fact that they represent a mixture of factors that results in series of observable patterns in the graphemic level. A representation of the adjustment appears in Figure 9.

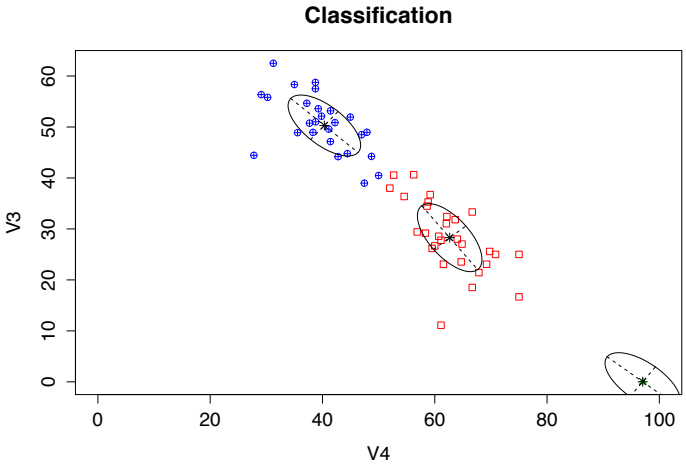


Figure 9. Distribution of texts in the morphographicity plane scaled up to 100 (horizontal axis is syllabicity, vertical axis is logographicity). Each point represents the final values of the graphemic type distributions for each text (cf. Figures 7 and 8), with coordinates M_s,M_l. A text for which all the glyphs can be read (or at least the type of the graphemes is known) would be represented by a point lying in the line drawn on the top right section of this plane. Texts with a number of missing graphemes will lie any place in the area between the upper bound represented by that line and the origin. A text with a relatively large number of logograms will be closer to upper left part, while a text with relatively large number of phonograms will be closer to the lower right part. There is a gradient in the use of graphemic types that can be associated to different scribal practices. Ellipses show the unsupervised clustering of texts that generate the notion of graphemic signature.

6. Conclusions and Future Work

The case study in Comalcalco and Palenque shows the possible application of graphemic features to sociolinguistic analysis. It is important to note that in this analysis only two distributions have been considered: it is possible, and desirable, to extend the analysis for intra-site discrimination of scribal traditions, inter-site analysis, and a diachronic evolution of the graphemic signatures. In the general framework, the next step involves the analysis of the interaction between scribal practices detected at the graphemic level with linguistic variation and change as exposed previously, according to Weinreich's types. Once the dynamics of this interaction is established, we will be able to link variation between communities of scribal and linguistic practice to sociopolitical networks.

Acknowledgements: We would like to express our sincere gratitude to John Justeson, Danny Law, John Rickford, Penny Eckert, and Dan Jurafsky for the comments received in the different phases of this project. We would like to thank very specially María Josefa Iglesias Ponce de León, Andrés Ciudad Ruiz, and our anonymous reviewer for the valuable comments and insights. This manuscript is closely based on a draft the authors where working when the tragic passing away of Alfonso Lacadena happened. Despite all the shortcomings sensibly pointed out by our reviewer, the junior author took the decision to keep its current form, closer to that original version that was lively discussed during numerous, truly special hours with the senior author.

7. References

Armijo Torres, Ricardo. 1999. «Nuevo hallazgo en Comalcalco». *Arqueología Mexicana* 37: 71–72.

Armijo Torres, Ricardo, Miriam J. Gallegos y Marc U. Zender. 2000. «Urnas funerarias, textos históricos y ofrendas en Comalcalco», in *Los Investigadores de la Cultura Maya* 8, Tomo II, pp. 312–323. Campeche: Universidad Autónoma de Campeche.

Armijo Torres, Ricardo, Marc U. Zender y Miriam J. Gallegos. 2000. «La urna funeraria de Aj Pakal Than, un sacerdote del siglo VIII en Comalcalco, Tabasco, México». *Temas Antropológicos* 22 (2): 242–253.

Bricker, Victoria R. 1986. *A Grammar of Mayan Hieroglyphs*. Middle American Research Institute 56. New Orleans: Tulane University.

Cases, Ignacio, Alfonso Lacadena and Christopher D. Manning. 2014. *Stochastic Modeling of Graphemic Chaining in Classic Maya Writing*. Technical Report CS224N. Stanford: Stanford University.

Ferguson, Charles A. 1959. «Diglossia». *Word* 15 (2): 325–340.

Geertz, Clifford. 1973. *The Interpretation of Cultures: Selected Essays*. New York: Basic Books.

Grube, Nikolai. 1990. «The Primary Standard Sequence in Chocholá Style Ceramics», in *The Maya Vase Book, Volume 2*, Barbara Kerr and Justin Kerr, eds., pp. 320–330. New York: Kerr Associates.

———. 2004. «The Orthographic Distinction between Velar and Glottal Spirants in Maya Hieroglyphic Writing», in *The Linguistics of Maya Writing*, Søren Wichmann, ed., pp. 61–81. Salt Lake City: University of Utah Press.

Houston, Stephen D. 2004. *Writing in Early Mesoamerica*. Cambridge: Cambridge University Press.

Houston, Stephen D., David S. Stuart and John S. Robertson. 1998. «Disharmony in Maya Hieroglyphic Writing: Linguistic Change and Continuity in Classic Society», in *Anatomía de una civilización. Aproximaciones interdisciplinarias a la Cultura Maya*, Andrés Ciudad, Yolanda Fernández, José Miguel García, Mª Josefa Iglesias, Alfonso Lacadena and Luis T. Sanz, eds., pp. 275–296. Madrid: Sociedad Española de Estudios Mayas.

Houston, Stephen D., John S. D. Robertson and David S. Stuart. 2000. «The Language of Classic Maya Inscriptions». *Current Anthropology* 41 (3): 321–356.

Justeson, John S. 1978. *Mayan Scribal Practice in the Classic Period: A Test-Case of an Explanatory Approach to the Study of Writing Systems*. PhD dissertation. Stanford: Stanford University.

Kaufman, Terrence S. with John S. Justeson. 2003. *A Preliminary Mayan Etymological Dictionary*. Foundation for the Advancement of Mesoamerican Studies (FAMSI). http://www.famsi.org/reports/01051/pmed.pdf.

Kaufman, Terrence S. and William M. Norman. 1984 [1978]. «An Outline of Proto-Cholan Phonology, Morphology, and Vocabulary», in *Phoneticism in Mayan Hieroglyphic Writing*, John Justeson and Lyle Campbell, eds., pp. 77–166. Institute for Mesoamerican Studies 9. Albany: State University of New York.

Kelley, David H. 1976. *Deciphering the Maya Script*. Austin: University of Texas Press.

Labov, William. 1972a. *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

———. 1972b. «Some Principles of Linguistic Methodology». *Language in Society* 1(01): 97–120.

Lacadena, Alfonso. 1995. *Evolución formal de las grafías escriturarias mayas: implicaciones históricas y culturales*. PhD dissertation. Madrid: Universidad Complutense de Madrid.

Lacadena, Alfonso and Ignacio Cases. 2010. *Introducción a la Escritura Jeroglífica Maya – Glosario Maya Jeroglífico*. Madrid: 15th European Maya Conference.

Lacadena, Alfonso and Søren Wichmann. 2002. «The Distribution of Lowland Maya Languages in the Classic Period, in *La organización social entre los mayas. Memorias de la Tercera Mesa de Palenque*, Vol. 2, Vera Tiesler, Rafael Cobos and Merle G. Robertson, eds., pp. 275–319. México: Instituto Nacional de Antropología e Historia y Universidad Autónoma de Yucatán.

———. 2004. «On the Representation of the Glottal Stop in Maya Writing», in *The Linguistics of Maya Writing*, Søren Wichmann, ed., pp. 100–164. Salt Lake City: The University of Utah Press.

Martin, Simon and Nikolai Grube. 2000. *Chronicle of the Maya Kings and Queens: Deciphering the Dynasties of the Ancient Maya*. London: Thames and Hudson.

McEnery, Tony and Andrew Hardie. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

Milroy, Lesley and Sue Margrain. 1980. «Vernacular Language Loyalty and Social Network». *Language in Society* 9 (1): 43–70.

Riese, Berthold. 2004. *Abkürzungen für Maya-Ruinenorte mit Inschriften*. Wayeb Notes 8. https://www.wayeb.org/notes/wayeb_notes0008.pdf.

Scollon, Ron and Suzie Wong Scollon. 2003. *Discourses in Place: Language in the Material World*. London: Routledge.

Shibamoto Smith, Janet S. and David L. Schmidt. 1996. «Variability in Written Japanese: Towards a Sociolinguistics of Script Choice». *Visible Language* 30 (1): 46–71.

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

Thompson, John Eric S. 1950. *Maya Hieroglyphic Writing: An Introduction*. Norman: University of Oklahoma Press.

———. 1962. *A Catalog of Maya Hieroglyphs*. Norman: University of Oklahoma Press.

Trudgill, Peter. 2002. «Linguistic and Social Typology», in *The Handbook of Language Variation and Change*, J. K. Chambers, Peter Trudgill, and Natalie Schilling-Estes, eds., pp. 707–728. Malden: Blackwell Publishing Ltd.

Weinreich, Uriel. 1953. *Languages in Contact: Findings and Problems*. The Hague: Walter de Gruyter.