Analysing rhythm in ritual discourse in Yucatec Maya using automatic speech alignment

Valentina Vapnarsky[1], *Claude Barras*[2], *Cédric Becquey*[1], *David Doukhan*[2^], *Martine Adda-Decker*[2] *& Lori Lamel*[2]

[1]Centre EREA du LESC, CNRS & Université Paris Ouest, France
[2]LIMSI-CNRS, Univ. Paris-Sud, 91403, Orsay, France
[1]valentina.vapnarsky@cnrs.fr, [2]firstname.lastname@limsi.fr

Abstract

Over the years, research in ethno-linguistics contributed to gather corpora in a wide range of languages, cultures and topics. In the present work, we are investigating ritual speech in Yucatec Maya. The ritual discourse tends to have a cyclic structure with repetitive patterns and various types of parallelisms between speech sections. Previous studies have revealed an intricate connexion between a speech's structure and vocal productions, in particular through temporal aspects including rhythm, pauses and durations of different speech sections. To further investigate our findings by relying more strongly on the acoustic recordings, automatic speech recognition tools may become of great help, in particular to test various linguistic and ethnolinguistic hypotheses. Unfortunately, Yucatec Maya, with less than one million native speakers, is an under-resourced language with respect to digital resources. As a total, 24 minutes of ritual speech from three performances were manually transcribed by expert linguists in Yucatec and a basic pronunciation dictionary for Yucatec was created accordingly. The transcribed acoustic recordings were then automatically time-aligned on a phonetic and lexical basis. Automatic segmentations were used to measure tempo changes, durations of breath units as well as to examine their link with the structure of the ritual text.
Index Terms: ethnolinguistic, Yucatec Maya, ritual discourse, automatic alignment, phonetic segmentation, tempo.

1. Introduction

Over the years, research in social sciences like anthropology, linguistics and ethnomusicology contributed to gather corpora in a wide range of languages, cultures and topics. In this context of growing amount of sound archive databases, the DIADEMS[1] research project aims at integrating automatic audio analysis tools in an indexing platform for ethnomusicological and ethnolinguistic archives and initiated a collaboration between several research laboratories in computer science and social science [5]. As part of this project, a special research is conducted on Mayan ritual speech to test ethnolinguistic hypotheses[2]. Ritual discourse has been a main topic of research, with its pragmatic and enunciative specificities, but previous studies emerged from very small-scale manual analyses, whereas it is clear that more precise, systematic and semi-automatized analyses would be indispensable to properly test the anthropological and linguistic hypotheses.

In the present work, we are investigating ritual speech in Yucatec Maya. The ritual discourse tends to have cyclic organisations with repetitive patterns and various types of parallelisms between speech sections. Previous studies revealed an intricate connexion between a speech's structure and vocal productions, in particular through temporal aspects including rhythm, pauses and duration of different speech sections. To further investigate our findings by relying more strongly on the acoustic recordings, automatic speech recognition tools may become of great help, in particular to test various hypotheses on the relation between the vocal production, the text and the performance. Unfortunately, Yucatec Maya, with less than one million native speakers, is an under-resourced language with respect to digital resources. As a total, 24 minutes of ritual speech from three performances were manually transcribed by expert linguists in Yucatec and a basic pronunciation dictionary was created accordingly. The transcribed recordings were then automatically time-aligned on a phonetic and lexical basis. Automatic segmentations were used to measure tempo changes, durations of breath units as well as to examine their link with the structure of the ritual text.

In the following, we first present the Yucatec Maya language and the ritual speech and vocal production. We then describe the corpus of three transcribed speeches. Section 5 presents the automatic alignment methodology before an analysis of the results in Section 6 and the conclusions and perspectives in the last section.

2. Yucatec Maya language

Yucatec Maya is one of some 30 languages of the Mayan family. It is spoken by about 800,000 speakers in the northern lowlands of the Maya area, throughout the Yucatan Peninsula. It shows less dialectal variation than Highland Mayan languages, but three main regional dialects have been distinguished [19]. Yucatec Maya is a mildly polysynthetic, initial predicate, head-marking language. It has a complex voice and aspectual system, with split ergativity [2, 1, 24]. Most of its lexical roots are CVC, where the vowel can have four distinct values (see below) [3]. Most clauses end by a final-clause vowel clitic, from a five-term deictic paradigm [12].

Yucatec Maya has a typologically original phonological system. Firstly, two series of consonants are contrasted by a glottalization feature: ejective consonants ([k'], [p'], [t'],[ts'],

This work was partly funded by the French National Agency for Research (ANR) under grant ANR-12-CORD-0022-05 (project DIADEMS).

^D. Doukhan is now with IRCAM.

[1]http://www.irit.fr/recherches/SAMOVA/DIADEMS/en/welcome/

[2]Other participants are M. Chosson and A. Monod Becquelin on Tseltal Maya.

[t͡ʃ']) vs. voiceless stops and affricates ([k], [p], [t], [ts], [t͡ʃ]). Secondly, its five vowels ([a], [e], [i], [o], [u]) may be realized in four distinctive ways involving length (short vs. long), and only for long vowels, tone (high *ún* vs. low *ùn*) and glottalization for long high-tone vowels (modal *ún* vs. creaky *un'un*) [10, 6]. Vowel distinctions have lexical and grammatical functions [15]. Besides the length opposition, vowel lengthening is a commonly used, yet unstudied, prosodic means of expression, with specificities depending on speech genres.

Table 1: *Consonants and vowels in Yucatec Maya. In italic, phonemes found in loan words and/or rare.*

	Labial	Alveolar		Palatal	Velar	Glott.
Stop	p	t *d*	ts	ch [t͡ʃ]	k *g*	‘
Ejective	p’	t’	ts’	ch’ [t͡ʃ']	k’	
Implosive	b					
Fricative	*f*	s		x [ʃ]		h
Nasal	m	n				
Trill		*r*				
Approx.		l		y	w	

	Front	Back
Close	i, ìi, íi, i’i	u, ùu, úu, u’u
Mid	e, èe, ée, e’e	o, òo, óo, o’o
Open	a, àa, áa, a’a	

3. Yucatec Maya ritual speech and vocal production

Yucatec Maya has various elaborated ritual speech genres, performed in agricultural, cynegetic and therapeutic ceremonies dedicated to, invoking guardian-spirits of the land and the forest, as well as in religious festivals with omnipresent prayers and ritual dialogues [11, 13, 20, 22, 23]. Yucatec Maya ritual speech is distinguished from mundane speech by a complex cyclic organisation and various types of parallelisms working through the text on different levels and speech sections [11, 16, 21]. These textual repetitive patterns contribute to distinguish one genre from another, but they are never completely fixed; variation (under certain limits defined by the genre) is a crucial aspect of Mayan ritual performativity [18].

Mayan ritual speech is also characterized by its specific vocal production, in particular temporal aspects including rhythm, pauses, tempo, vowel lengthening patterns and pitch modulation. This has been long noticed by ethnographers working on different Mayan languages [9, 7, 11, 17], however no systematic study has yet been done on this respect. Preliminary studies, based on small-scale manual analyses led to the hypotheses of an intricate connexion of ritual vocal production (1) with the ritual actions performed or aimed at as the ritual unfolds and (2) with the textual structure. As regards (1), tempo and pause variations in particular seem to be linked with special phases of the ritual, as well as with the energy required to move spiritual entities and to adapt to the state and the capacities of the other participants in the ritual (e.g. patients in therapeutic rituals, dialogue partner in ritual dialogues). As for (2), length variations of the speech segments appear to add an extra layer of structural complexity to the discourse, with the introduction of discrepancies between the textual unit and the vocal unit boundaries. These partial overlaps and the non-correspondences between textual and vocal units also vary during the ritual and brings us back to question (1).

4. Corpus description

The three speeches analysed come from a village located about 30km. south of Felipe Carrillo Puerto (Eastern region of the Yucatan Peninsula). They were audio-recorded in 1995 and 1996, by V. Vapnarsky during two rituals dedicated to guardian-spirits of the forest. The same speaker, a well-known ritual specialist in the sub-region, was involved in each case. He knew about and agreed to be recorded. The recording method was intended to be the least intrusive possible, it was done in the context of long term fieldwork, making the performance as natural as possible. Although the content of the discourse is very similar, the speaker shows some stylistic differences or variations between the two performances, the analysis of which is part of our objective. In particular, in the second ritual he was influenced by another ritual specialist he had heard a couple of weeks before, and this is reflected by the more emphatic tone and melodic contours of his speech.

Speech 1 and 2 were performed in a first-fruit ceremony for the maize harvest (ho’olbesah nal). In Speech 1, the ritual specialist goes with his voice to different places of the terrestrial and cosmological world to gather the spiritual entities and bring them to the altar; in Speech 2, he brings them back to their dwelling place, after the offerings. Speech 3 corresponds to the first phase (spirit gathering) of a ritual dedicated to the same entities, which was performed for prophylactic purposes. The three speeches have similar lengths (1/ 13:54 ; 2/ 11:25 ; 3/ 13:15) and their textual structure is very homogeneous. The general organisation of the text is the following: contextual anchoring of the ritual (space, time, sponsor, purpose); long invocation of the spiritual entities with textual cycles (see extract in Table 2); *credo* (part of the offering); invocation of saints and other spiritual entities; closing. The speaker makes the sign of the cross at specific moments, kissing his hand (kissing-sound audible in the recording).

The recordings were manually segmented into vocal units (or breath groups) and transcribed using the ELAN tool[3] [4]. In addition, textual units boundaries (cycles and verses) were marked with special characters in the flow of the transcription. However the corpus is very noisy due to the acoustic background and not all vocal units were transcribed. Also, due to its specific pronunciation and esoteric content, one section appeared to be very difficult to transcribe precisely and was left for future research. Out of the 342 segmented vocal units in the three speeches, 261 were fully transcribed for a cumulated duration of 24 min of speech.

These ritual performances involve special vocal production features, with fast tempo, alternations of series of short and extra long breath units (as long as possible until running out of breath), extra lengthening of some vowels especially at the beginning and the end of some textual cycles, high intensity and pitch at the beginning of speech units.

5. Automatic alignment

Various versions of the dictionary were tested and the resulting alignement were checked by the Yucatec Maya experts before achieving a satisfying version. With respect to phonemic inventories described in Section 2, some simplifications have been carried out. We did not use different models for tones and length in vowels as in these, modal voice tends to produce

[3] `http://tla.mpi.nl/tools/tla-tools/elan/` provided by Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands

Table 2: *Extract of Speech 3 (04:37 to 05:19) # breath units boundary, // cycle textual boundary, ; verse boundary.*

Yucatec Maya	English translation
# // Hats'aknaak topoknaak;	*Hats'aknaak topoknaak*
kubin int'àan ;	my words go
tunoh uk'a' bin;	to the right hand they say
u'ah kanan kàakbilo';	of the guardians of the earth
ti' bin ti' bin u'ah kanan montanya'ilo';	to they say the guardians of the high forest
ubàalamk'áaxilóo';	to the jaguars of the forest
ti' bin u'ah tepalilóo';	to they say the rulers
# Tsukha'ase;	at Tsukha'ase'
yuumeeen;	my looord.
// Hats'aknaaak topoknaak;	*Hats'aknaak topoknaak*
kubin int'aan;	my words go
tunoh uk'a' biin;	to the right hand they say
u'ah kanan kàakbilóo';	of the guardians of the earth
ti' bin u'ah kanan ; montanya'ilóo';	to they say the guardians of the high forest
# beeh San Bartolo'e';	towards San Bartolo
yumeeeen;	my looord.
// Hats'aknaaaak topoknaak;	*Hats'aknaak topoknaak*
kubin int'àan;	my words go
tunoh uk'a' bin ;	to the right hand they say
u'ah kanan kàakbilo' xaan;	of the guardians of the earth also
# le San Merehìildo'e'	of San Merejiildo'e'
Yumeen ;	my loord.

similar acoustic parameters for modeling, but we used distinct acoustic models for modal and glottalized vowel productions for each of the five vowel qualities. Some phones, introduced by foreign Spanish names, are very rare, especially the 'f' which was only observed in the word 'fiyadòore' and was replaced by a close Mayan phoneme. The dictionary contains thus 10 vowels plus 22 consonants. Given that the convention for orthographic transcription for Yucatec Maya are very close to the pronunciation, the pronunciation dictionary was built through a set of simple transliteration rules.

A set of context-independent phone models were estimated on the available transcribed audio data. The models were trained from a flat-start, meaning that no prior segmentation information was used [25]. The acoustic models are tied-state, left-to-right 3-state HMMs with Gaussian mixture observation densities (about 10-15 components) [8]. The set of phones was determined by the pronunciation dictionary which provided one or more phonemic form for each lexical item. The acoustic feature vector contains 42 components formed by the concatenation of 39 PLP-like [14] features with a 3-dimensional pitch feature vector (pitch, Δ and $\Delta\Delta$ pitch). More precisely, the PLP features are derived from a Mel frequency spectrum, with cepstral mean removal and variance normalization carried out on a segment-cluster basis.

The resulting acoustic models were used to align the speech data with the provided transcriptions, according to the available word level pronunciations and with optional pauses allowed between the words. The alignment failed for 18 out of the 261 transcribed vocal units, representing about 2 minutes of speech; this usually occurs due to a discrepancy between the acoustic

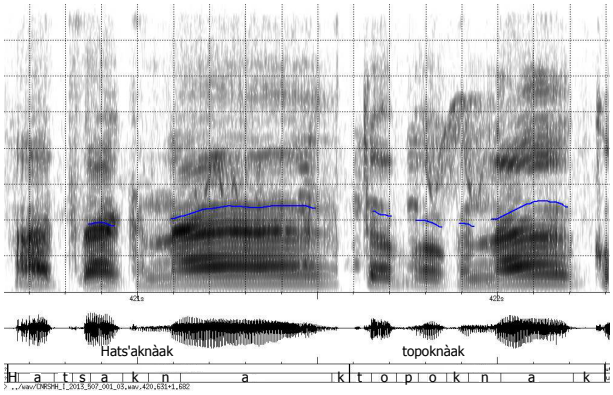

Figure 1: *Spectrogram illustrating an excerpt of ritual discourse with automatically aligned words and phones.*

and phonetic levels, eg. in our corpus when the voice is covered by the crowing of a cock. For all other segments, the alignment provides accurate time stamps to each word and also at the phonetic level. Figure 1 shows a spectrogram of an excerpt of the corpus. The precise temporal location is also propagated to the textual unit boundaries inside the transcription, ie. the cycles and verses. Finally, optional pauses between words allow to identify short silences in the recording and to refine the boundaries of the breath groups which were manually segmented.

6. Results and discussion

The upper graph in Figure 2 synthesizes the temporal organization of the successive breath groups for the third speech recording. For each breath group, a vertical bar displays the duration of the vocal unit. The duration of each inter-segment pause is also visible below the breath group it precedes as a black line under the x-axis. When the transcription was incomplete (units between 45 and 60) or when the alignment failed (eg. for unit 68), the bar is shown grey and dashed. The temporal structure of the textual units is visible through green circles located at the beginning of the cycles and colored fragments for each of the verses. We can observe very long breath groups up to 18 seconds which are specific of the rituels. We also see that most of the cycles start at the beginning of a breath group.

For the same vocal units, the lower graph in Figure 2 displays the syllabic rhythm in the same vocal units. The number of syllables in a breath group was counted as the number of vocalic nuclei in the phonetic transcription, and it was divided by the cumulated speech duration as obtained with the phonetic alignment. The local rhythm is shown relative to the average rhythm which is 5.6 syll/sec. in the 3rd recording with much lower values at instants (eg. 1 syll/sec for the 12th unit which consist of a single short verse 'yumeen' finishing the cycle with an extremely lengthened final syllable).

The statistics and the graphs obtained provide results which are relevant for the characterization of the vocal production of ritual speech, compared to other genres (conversation, narratives, etc.). They also provide evidence to test the hypotheses (1) and (2) mentioned above. In particular, they allow a quantitative analysis and a user-friendly visualization of the relations between textual and vocal units, and its progression through the discourse (e.g. in the 3rd recording, we observe a correspondence between cycle and vocal units at the beginning, followed by more complex patterns as the ritual unfolds). In further stud-

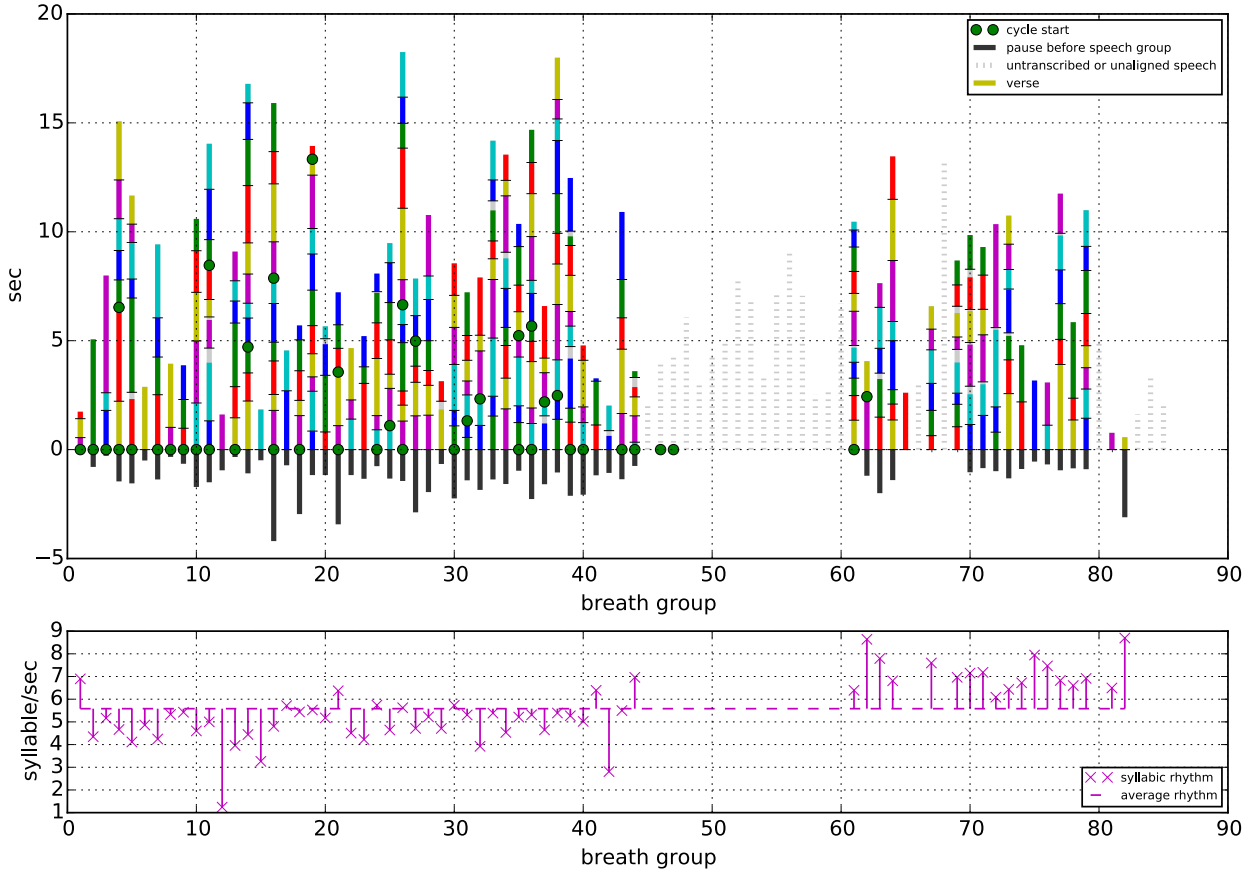

Figure 2: *Temporal structure of the 3rd recording, above the duration of the pauses and of the speech groups and the internal division into cycles and verses, below the evolution of the syllabic rhythm.*

ies, each type of verse could be coded with specific colours, which would allow to follow verse recurrence patterns. The tempo analysis also offers crucial data to the understanding of the rhythmic variations during the performance (e.g. in the 3rd recording, a tempo acceleration is clearly visible in the final section, which reveals a distinct phase of the ritual). Overall, these tools appear to be very useful for the comparison between ritual performances as well as between ritual genres in Yucatec Maya.

7. Conclusions

We have presented a collaborative work between researchers in social sciences and computer science to design audio analyses tools within an information indexing and retrieval platform for ethnomusicological and ethnolinguistic archives. More specifically, we worked on an under-resourced language, Yucatec Maya, in order to support and speed up phonological and prosodic quantitative studies of ritual discourse and to compare these to other everyday-like speech genres. A pronunciation dictionary was created and 24 minutes of manually transcribed ritual speeches were sufficient to train Yucatec Maya acoustic models. A speech tempo metric based on number of syllables per breath (or vocal) units was defined. Automatic speech alignment produced accurate time-stamps for the ritual discourse speech data on cycle, verse, word and phone levels. These allowed us to give a quantitative description of tempo changes within and across verses and cycles in ritual discourse.

The phonemic and lexical alignments are a good starting point for further analyses of vowel lengthening and pitch modulation. The proposed work is easily extendable to the whole corpus of Maya Yucatec ritual discourses but also, for comparative purposes, to other discourse genres and to other languages of the Mayan family. Potentially, the methodology could be applied beyond this family, to other under-ressourced languages, fostering cross-cultural/linguistic analysis, a crucial tenet of human social sciences. This will contribute to renew the ethnolinguistic analysis of oral discourse genres in important ways.

8. References

[1] J. Bohnemeyer, *The grammar of time reference in Yucatec Maya*. Muenchen: LINCOM Europa, 2002.

[2] V. Bricker, "The source of the ergative split in Yukatek Maya". *Journal of Mayan Linguistics* 2(2), pp. 83–12, 1981.

[3] V. Bricker, E. Po'ot Yah and O. Dzul de Po'ot. *A Dictionary of The Maya Language as Spoken in Hocabá, Yucatán*. University of Utah Press, 1998.

[4] H. Brugman and A. Russel, "Annotating Multimedia/ Multimodal resources with ELAN". In: *Proceedings of LREC 2004, Fourth International Conference on Language Resources and Evaluation*. 2004.

[5] T. Fillon, J. Simonnot, M.-F. Mifune, S. Khoury, G. Pellerin, E. Amy de La Bretèque, D. Doukhan, D. Fourer, J.-L. Rouas, J. Pinquier, and C. Barras, "Telemeta: An open-source web framework

for ethnomusicological audio archives management and automatic analysis," in The 1st International Digital Libraries for Musicology workshop (DLfM 2014), London, UK, 2014.

[6] M. Frazier, "The phonetics of Yucatec Maya and the typology of laryngeal complexity". *Language Typology and Universals* 66, pp. 7–21, 2013.

[7] F. Furbee, "To Ask One Holy Thing : Petition as a Tojolobal Maya Speech Genre". In *Tojolabal Maya : ethnographic and linguistic approaches*. M. J. Brody & J. S. Thomas (eds). pp. 39–53. Geoscience and Man 26. Geoscience Publications. Baton Rouge : Louisiana State University, 1988.

[8] J.-L. Gauvain, L. Lamel, and G. Adda, "The LIMSI Broadcast News transcription system," *Speech Communication*, vol. 37, no. 1–2, pp. 89–108, mai 2002.

[9] G. Gossen, "To speak with a Heated Heart: Chamula Canons of Style and Good Performance". In *Explorations in the ethnography of speaking*, R. Bauman & J. Sherzer (Eds.). Cambridge: Cambridge University Press, 1974.

[10] C. Gussenhoven and T. Renske, "A moraic and a syllabic H-tone in Yucatec Maya". In *Fonología instrumental: Patrones fónicos y variación*, Esther Herrera Z. & Pedro Martín Butrageño (Eds.). pp. 49–71. Mexico City: El Colegio de México, 2008.

[11] W. Hanks, "Sanctification, structure and experience in a Yucatec Maya ritual event". *Journal of American Folklore*, pp. 131–166, 1984.

[12] W. Hanks, "Explorations in the Deictic Field". In *Current Anthropology* 46, pp. 191–220, 2005.

[13] W. Hanks, "Joint commitment and common ground in a ritual event". In *Roots of Human Sociality. Culture, cognition and Interaction*, Enfield, N.J. & S. Levinson (Eds). pp. 299–328. Oxford: BERG, 2006.

[14] H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech," *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[15] X. Lois and V. Vapnarsky. "Polyvalence of root classes in Yukatekan Mayan Languages". Muenchen: LINCOM, 2003.

[16] A. Monod Becquelin and C. Becquey, "De las unidades paralelísticas en las tradiciones orales mayas". In *Estudios de cultura maya* 32, pp. 111–153, 2008.

[17] A. Monod Becquelin and A. Breton, *La guerre rouge ou une politique maya du sacré. Un carnaval tzeltal au Chiapas*, Mexique. Paris: CNRS Editions, 2002.

[18] A. Monod Becquelin, V. Vapnarsky, C. Becquey, and A. Breton. "Paralelismo, variantes y variaciones : decir, contar y rezar la diversidad maya". In *Figuras mayas de la diversidad*, Aurore Monod Becquelin, Alain Breton et Mario Humberto Ruz (Eds). pp. 101–152. Universidad Nacional Autónoma de México, 2010.

[19] B. Pfeiler and A. Hofling, Apuntes sobre la variación dialectal en el maya yucateco. *Península*. I(1):27-45, 2006.

[20] V. Vapnarsky, "De dialogues en prières, la procession des mots". In *Les Rituels du dialogue*, A. Becquelin & P. Erikson (Eds). pp. 431–479. Nanterre: Société d'Ethnologie, 2000.

[21] V. Vapnarsky, "Paralelismo, ciclicidad y creatividad en el arte verbal maya yucateco". *Estudios de Cultura Maya* XXXII, pp. 155–199. México, 2008.

[22] V. Vapnarsky, "Briser les vents et échanger les cœurs : art et performance dans les discours rituels mayas contemporains". Mayas. In *Mayas. Révélations d'un temps sans fin*, D. Michelet (ed.). pp. 123–129. Paris: Musée du quai Branly, Réunion des musées nationaux, 2014.

[23] V. Vapnarsky and O. Le Guen. "The guardians of space and history: Understanding ecological and historical relations of the contemporary Yucatec Maya to their landscape". In *Ecology, Power, and Religion in Maya Landscapes*, C. Isendahl & B. Liljefors Persson (Eds). Verlag Anton Saurwein. pp. 191–206, 2011.

[24] V. Vapnarsky, A. Monod Becquelin, and C. Becquey. "Passive and ergativity in three Mayan languages". In *Ergativity, Valency and Voice*, G. Authier & K. Haude (Eds). pp. 51–110. Mouton de Gruyter, 2012.

[25] S. J. Young and P. C. Woodland, "State clustering in hidden Markov model-based continuous speech recognition," *Computer Speech & Language*, 8(4), pp. 369–383, Oct. 1994.