Using corpus data for testing Haitian Creole prosody

Alexander M. Teixeira Kalkhoff (2018)

1. Introduction

The paper records the technical and methodological aspects of a corpus-based approach to Haitian Creole (henceforth HC) prosody, such as automatic audio–text alignment of the corpus data and prosodic annotation, exploring the linguistic resources of the Corpus of Northern Haitian Creole (henceforth CNHC).

2. Establishing the workflow for prosodic analysis

Recorded in 2007 and subsequently published by the Indiana University Creole Institute under the auspices of Albert Valdman, the Corpus of Northern Haitian Creole nowadays constitutes a freely available and downloadable online resource for linguistic research on HC. It documents speakers of Capois, i.e. the language variety of HC spoken in the north of Haiti near Cap Haïtien. The mentioned webpage gives access to the audio files of the recordings, to the text files of the transcripts, and to the transcription conventions of ten interviews of altogether approximately ten hours of speech data. Audio files are saved as MP3 and the transcripts as Microsoft® Word files with an overall size of 509 MB. Audio and text files are nor aligned. The transcripts are not translated in any other language.

To establish the workflow for a test run, I chose the longer turn of speaker R from the first interview of the CNHC that can be seen in (1). Selection criteria were, first, the selected test piece should not be too close to the beginning, because a more natural and relaxed social and linguistic interaction between the speakers was expected after a certain time of contact and the fading of the awareness of being recorded, and, second, the test piece should represent a longer turn of one single speaker. The whole conversation was recorded in Thibeau near Cap Haitian and involves three participants, i.e. one interviewer (K), a native school teacher, and two consultants (S, R), two middle-aged farmers, both illiterate monolingual speakers of Capois, the Northern HC variety:[1]

[1] The test piece corresponds to the time interval from 00 h 19 min 26.540 sec to 00 h 20 min 27.165 sec of the audio signal and to the lines 211 to 222 of the transcript. I removed all metalinguistic information, punctuation, and capitalization, which the transcribers originally used for sentence marking of the orthographic transcription, from the transcript. Audible discontinuities were marked by the IPA symbol for intonational grouping ‖.

(1) oke pou mayi ya ‖ m te ka bay yon ti fòmasyon nan mayi ya ‖ mayi ya yon epòk ‖ nan tan lontan lè papanm ‖ tè yo te pi kiltivab ‖ sezon an te pi mache pi byen ‖ ou sè ou konn rive nan tè a ‖ mèt gen rajo ‖ ou fouye tou mayi ou plante mayi ‖ apre lè w fini w sèrkle mayi ya ou fè mayi ‖ men nou rive nan on moman konnya nou vin pa gen sezon menm jan ankò ‖ nou nou sèrkle tè a ‖ lè w fin sèrkle tè a ‖ ou mete latè ‖ tè a vin rive nan lè w rive nan ‖ ou vin rive nan mwa desam ‖ janvye w ap sèrkle tè nou konn plante mayi janvye ‖ nou plante mayi fevriye ‖ nou plante mayi mas ‖ nou plante mayi avril se kat mwa sa n te konn gen n te konn kiltive mayi ya ‖ pou sezon ‖ lè w plante ‖ mayi ya ‖ mayi ya vin i leve ‖ ou sèrkle men nan tan lontan granmoun yo te konn plante i yon lòt jan ‖ tan yo te pi bon ‖ yo konn plante mayi ya senk grenn nan tou ‖ yo fouye tou mayi ya yo mete senk grenn mayi ‖ men nou menm konnya nou vin pa gen sezon menm jan ‖ nou plante mayi ya twa grenn ‖ e nou vin plante a yon distans tou ‖[2]

As a first step of the editing process, the downloaded audio signal was normalized, denoised, and saved as WAV audio file by using Audacity® (version 2.2.2), freely available software for recording and sound editing. It is a big technical deficit of the CNHC audio data for further automatic phonetic analysis that the recordings were saved originally in the lossy compressed MP3 (MPEG-1 Audio Layer 3) audio format (128 kBit per second, 44.1 kHz) instead of WAV (Waveform Audio File Format) (1411 kBit per second, 44.1 kHz). But, both the Praat (Boersma & Weenink 2017) algorithm to calculate automatically the pitch and the MAUS (Munich AUtomatic Segmentation) algorithms (Schiel 1999; Kisler et al. 2012) to align the audio signal to the text and to segment the sound signal turned out to be robust enough to deal with the lossy compressed audio signal. Nevertheless, regarding the pitch track visualized in Praat, there are some artefacts, such as cracks and discontinuities, in our test piece.

The second step was the automatic sound–text alignment and word-and-segment-sized segmentation. This step had rather an experimental character because, of course, for my short sample, I could align

[2] ‘okay for the corn / I can give you a short instruction for the corn / the corn, in a time far in the past, at the age of our fathers / their soil was more cultivable / the season worked better / you know, you could arrive at a ground / you weeded / you dug all the corn, you planted the corn / after you finished to weed, you made corn / but, we arrived at a time where the season does not work longer in the same way / we weed the ground / when we finished to weed the ground / you put into the soil / you come along in the month of December / in January you will weed the ground we can plant corn in January / we plant corn in February / we plant corn in March / we plant corn in April, these four months we could have, we could cultivate the corn / for the season / when you plant / the corn / the corn grows / you weed, but, at the time far in the past, the elderly people they could plant in another way / their time was better / they used to plant the corn with five seeds in a hole / they dig a hole for the corn, they put five corn seeds in it / but we ourselves use, (because) we do not have the seasons in the same way / we plant corn with three seeds / and we plant at a totally different distance’ [My translation].

the audio with the transcript file and segment the words and sound segments manually. I did this step rather in the perspective of taking soundings for a further use of bigger amounts of corpus data. To accomplish this step, I took advantage of the tools and services provided openly and free of charge by the Bavarian Archive for Speech Signals (BAS), which is part of the German CLARIN-D (Common Language Resources and Technology Infrastructure) research infrastructure for humanities and social sciences. The BAS team used the well-trained German acoustic model to perform the HC alignment and segmentation procedure because the German and HC phonemic spelling systems share many features (for HC spelling see Valdman 1981, x-xiii). The output format of the MAUS procedure is a Praat text grid containing the interval tiers ORT, i.e. word-tokenized orthographic transcript, KAN, i.e. canonical pronunciation, and MAU, i.e. SAM-PA (Speech Assessment Methods Phonetic Alphabet) encoded phonemes. As a result, the formerly unconnected audio signal and the text of the transcript are now temporally aligned and decomposed into words and sound segments, ready to be re-used for further phonetic analysis in Praat.

The third step was to create an annotation hierarchy for prosodic analysis in Praat taking into account the prosodic features of HC, i.e. prominences, syllable structure, local tonal movements, duration of segments, and prosodic phrasing. For that purpose, I adapted the Intonational Variation Transcription System (IVTS). IVTS is a language-non-specific annotation system to annotate intonational variation of still unknown phonological systems developed by Brechtje Post and Elisabeth Delais-Roussarie (Post & Delais-Roussarie 2006) on the basis of the language-specific Intonation Variation in English (IViE) annotation system (Grabe & Post 2004).

The IVTS originally encodes orthographic, prosodic, and intonational information on six annotation levels, i.e. (i) a comment, (ii) a phonological (or tonal), (iii) a global phonetic, (iv) a local phonetic, (v) a rhythmic (or prominence), and (vi) an orthographic tier. Starting point for the annotation process is the identification of prominent syllables and a narrow phonetic annotation of local intonational events. Intonation at the discourse level is annotated phonetically at the global phonetic tier. First phonological assumptions can be also annotated at the phonological tier. The IViE/IVTS set of labels is transparent, easy to manage and re-uses well-established ToBI symbols, such as L, H, and % (Grabe 2001).

According to my focus on the prosodic features of HC, I adapted the IVTS and established the following annotation hierarchy for Praat (see figure 1): (i) a segmental tier (SEG) to enable durational measurements, (ii) a syllable tier (SYL) to evaluate syllable patterns and complexity, (iii) a rhythmic tier

(RHY) to annotate prominent syllables and intonational phrasing, (iv) a local phonetic or micro-prosodic tier (MPR) to annotate phonetically the realized pitch levels or glides within the tonal implementation domain[3] of prominent syllables, (v) a break indices tier to annotate prosodic constituency (PCO), (vi) an orthographic tier (ORT) for an orthographic transcription of the speech signal, and (vii) a glossed tier (GLO) to enable to evaluate pitch movements upon function word classes.

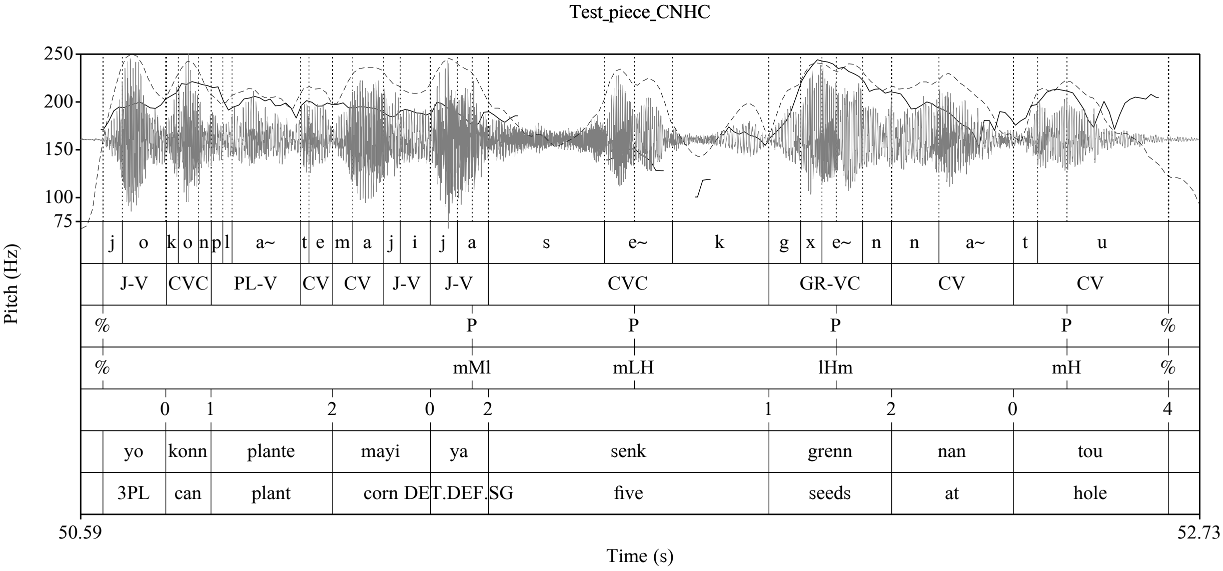

Figure 1: Prosodic annotation of the IP [yo konn plante mayi ya senk grenn nan tou]$_{\mathrm{IP}}$ ('they used to plant the corn with five seeds in a hole').

Sound segments were labeled automatically by the MAUS algorithm with SAM-PA symbols. Labels for syllable constituents are basically C for consonants and V for vowels. On the rhythmic tier, phonetic salience of a syllable relative to adjacent syllables is labeled with P for prominence and rhythmic boundaries are marked by the boundary symbol %. On the micro-prosodic tier, the shape of the local pitch movement relative to the prominent syllables is annotated with six target labels. Capital labels H(igh), M(id), and L(ow) mark the pitch level of the accented syllables; minuscule labels h(igh), m(id), and l(ow) mark the pitch level of unaccentuated syllables preceding or following the prominent syllable (Grabe 2001).[4] Break indices indicate boundaries and junctures between prosodically sensitive constituents. The break index 4 indicates intonational phrase boundaries, 3 indicates intermediate phrase boundaries, 2 indicates clitic group boundaries, BI 1 indicates phonological word boundaries, and 0 indicates junctions between content and function words (Frota 2012). On the orthographic tier,

[3] The tonal implementation domain contains the pre-accentual syllable, the accented syllable, and all following syllables up to the next accented syllable or up to the end of the IP (Grabe 2001).

[4] According to the speaker's pitch range ranging from 130 to 240 Hz, H(igh) tones were assigned to pitch values ranging from 200 to 240 Hz, M(id) tones were assigned to pitch values ranging from 165 to 200 Hz, and L(ow) tones were assigned to pitch values ranging from 130 to 165 Hz.

the words from the transcript appear time-aligned with the audio signal as a result of the MAUS alignment process. Each word is glossed according to the *Leipzig Glossing Rules* (The Leipzig Glossing Rules 2015).

3. Pros and cons of a corpus-based approach to HC prosody

Currently, spoken and written, synchronic and diachronic creole language corpora are a prolific and quite quickly evolving field (see amongst others Hagemeijer et al. 2014 for the building of a Gulf of Guinea Creole corpus and Kriegel 2015 for an overview of French-based creole corpora). Besides the CNHC, online resources for HC of varying quality and functionality are the short APiCS online (2013) HC sample, several recordings of spoken HC from the 1980s provided by the French meta-platform for oral corpora CoCoON, and the Haitian Creole language data corpus (2010) built by the Language Technologies Institute of Carnegie Mellon University's School of Computer Science. But, none of them complies with the full range of desirable corpus functionalities and technical standards, i.e. audio–text alignment, glossing, translation, metadata, annotations, license for free data use, download and publishing (e.g. creative commons), clear indication for citation, high fidelity audio recordings, and TXT or XML text format for transcripts, metadata and annotations (see amongst others Gries & Newman 2013 and the CLARIN-D standards information system).

Corpus-based approaches to language are economic and sustainable, especially if we consider language communities that live far from our own home country. Instead of prior time, money, and infrastructure-consuming fieldwork, we can start with linguistic analysis nearly immediately by re-using already existing data. Of course, besides that undeniable merit, many technical and methodological problems arise while using corpus data. On the one hand, these are related to the general properties of linguistic corpora, which linguists usually create for specific research questions. In the case of the used CNHC, linguists have explored the corpus data for lexical, morphophonological, syntactic, and sociolinguistic concerns (for recent research see, for instance, Valdman et al. 2015). Thus, the design of the interviews and the quality of the recordings did not pay special attention to the usability of the speech data for later phonetic, prosodic, or even intonational analyses. On the other hand, many problems of today's usability have to do with technical standards and limits of the date of creation of the corpus. For instance, in 2007, the audio files of the CNHC have been saved in the disk space-saving but lossy MP3 audio format. Today the memory capacity of electronic recording devices and of computer hard drives as well as the online data transmission are not a major problem anymore. Large WAV files neither limit field recordings nor website functionalities of online-corpora.

In face of all undeniable challenges and inconveniences of field research, creolists in the field should keep in mind well-established technical and methodological standards for language resources and technology (see Maddieson 1999 and Podesva & Zsiga 2013 for phonetic field work and the CLARIN-D standards information system for data formats). Audio files should be recorded and saved as WAV and if any audio compression is needed, it should be executed by lossless compressed digital audio formats such as FLAC (Free Lossless Audio Codec). For the subsequent transcription and annotation process, computer programs such as freely downloadable ELAN or EXMARaLDA should be used instead of creating separate Word or Excel documents to ensure suitable data formats.

A major challenge for creole language prosody and intonation research are the background noises due to the natural environment of the interview settings such as crickets, barking dogs, human voices, and human activities such as playing children or the use of tools. In the case of the CNHC, interviews 5, 6, 7, 8, and 10, i.e. half of the corpus data, cannot be used at all for phonetic analysis because of strong background noises. Therefore, researchers in the field creating new audio or multimedia data should seek noise-reduced recording facilities.

Acknowledgements

I am very grateful to Uwe Reichel from the CLARIN-D infrastructure of the Ludwig Maximilian University of Munich who enabled the use of the BAS tools for HC alignment and segmentation.

6. References

APiCS online (2013) = *Atlas of Pidgin and Creole Language Structures online*, accessed 2018-06-10. URL: http://apics-online.info/

Audacity® (version 2.2.2), accessed 2018-06-10. URL: http://audacity.sourceforge.net/

Boersma, Paul & Weenink, David. 2017. Praat: doing phonetics by computer [Computer program]. Version 6.0.30, retrieved 22 July 2017 from http://www.praat.org/

CLARIN-D (Common Language Resources and Technology Infrastructure), accessed 2018-06-10. URL: http://www.clarin-d.de

CLARIN-D standards information system, accessed 2018-06-01. URL: https://clarin.ids-mannheim.de/standards/index.xq;jsessionid=5vjm9hrn6sjz1nc0ypi9mest3

CoCoON (Collections de Corpus Oraux Numériques), accessed 2018-06-10. URL: http://cocoon.huma-num.fr/exist/crdo

Corpus of Northern Haitian Creole (2007), accessed 2018-06-10. URL: http://www.indiana.edu/~creole/index.shtml

ELAN, accessed 2018-06-10. URL: https://tla.mpi.nl/tools/tla-tools/elan/

EXMARaLDA, accessed 2018-06-10. URL: http://exmaralda.org/de/

Frota, Sónia (2012) "Prosodic structure, constituents and their implementation", in Cohn, Abigail C. & Fougeron, Cécile & Huffman, Marie K. (eds.) *Handbook in Laboratory Phonology*, Oxford: Oxford University Press, p. 255-265.

Grabe, Esther (2001) "The IViE Labelling Guide, version 3", accessed 2018-06-10. URL: http://www.phon.ox.ac.uk/files/apps/IViE/guide.html

Grabe, Esther & Post, Brechtje (2004) "Intonational variation in the British Isles", in Sampson, Geoffrey & McCarthy, Diana (eds.) *Corpus Linguistics: Readings in a widening discipline*, London: Continuum International, p. 474-481.

Gries, Stefan Th. & Newman, John (2013) "Creating and using corpora", in Podesva, Robert J. & Sharma, Devyani (eds.) *Research Methods in Linguistics*. Cambridge: Cambridge University Press, p. 257-287.

Hagemeijer, Tjerk & Généreux, Michel & Hendrickx, Iris & Mendes, Amália & Tiny, Abigail & Zamora, Armando (2014) The Gulf of Guinea creole corpus, *Proceedings LREC 2014*, Reykjavik (Iceland), p. 523-529.

Haitian Creole language data (2010), accessed 2018-06-10. URL: http://www.speech.cs.cmu.edu/haitian/

Kisler, Thomas & Schiel, Florian & Sloetjes, Han (2012) Signal processing via web services: The use case WebMAUS. in *Proceedings Digital Humanities 2012*, Hamburg (Germany), p. 30-34.

Kriegel, Sibylle (2015) "La documentation linguistique des franco-créoles", in Iliescu, Maria & Roegiest, Eugeen (eds.) *Manuel des anthologies, corpus et textes romans* (Manuals of Romance Linguistics 7), Berlin: De Gruyter, p. 647-658.

Maddieson, Ian (2001) "Phonetic fieldwork", in: Newman, Paul & Ratliff, Martha (eds.) *Linguistic Fieldwork*. Cambridge: Cambridge University Press, p. 211-229.

Podesva, Robert J. & Zsiga, Elizabeth (2013) "Sound recordings: acoustic and articulatory data", in Podesva, Robert J. & Sharma, Devyani (eds.) *Research Methods in Linguistics*. Cambridge: Cambridge University Press, p. 169-194.

Post, Brechtje & Delais-Roussarie, Elisabeth (2006) "Transcribing intonational variation at different levels of analysis", in *ISCAM Archives, Speech Prosody 2006*, Dresden (Germany).

Schiel, Florian (1999) Automatic Phonetic Transcription of Non-Prompted Speech, in *Proceedings of the 14th International Congress of Phonetic Sciences*, San Francisco, p. 607-610.

The Leipzig Glossing Rules (2015), accessed 2018-06-10. URL: https://www.eva.mpg.de/lingua/resources/glossing-rules.php

Valdman, Albert (ed.) (1981) *Haitian Creole-English-French dictionary*, Bloomington: Indiana University, Creole Institute.

Valdman, Albert & Villeneuve, Anne-José & Siegel, Jason F. (2015) "On the influence of the standard norm of Haitian Creole on the Cap Haïtien dialect: Evidence from sociolinguistic variation in the third person singular pronoun". *Journal of Pidgin and Creole Languages* 30(3), p. 1-43.