Crossroads Corpus creation: Design and case study

Abbie Hantgan-Sonko

DE GRUYTER OPEN

Yearbook of the Poznań Linguistic Meeting 3 (2017), pp. 167–198
DOI: 10.1515/yplm-2017-0009

Crossroads Corpus creation: Design and case study

Abbie Hantgan-Sonko
Independent researcher
ahantgan@gmail.com

Abstract

This paper illustrates a methodological approach to the design of an annotated corpus using a case study of phonetic convergences and divergences by multilingual speakers in southwestern Senegal's Casamance region. The newly compiled corpus contains approximately 183,000 annotations of multilingual, spoken data, gathered by eight researchers over a ten year span using methods ranging from structured lexical elicitation in controlled contexts to naturally occurring, multilingual conversations. The area from which the data were collected consists of three villages and their primary languages, and yet many more contribute to the linguistic landscape. Detailed metadata inform analyses of variation, the context in which a speech act took place and between whom, the speakers' linguistic repertoires, trajectories, and social networks, as well as the larger language context. A potential path for convergence or divergence that emerged during data collection and in building and searching the corpus is the crossroads in the phonetic production of word-initial velar plosives. Word-initial [k] emerges in one language where only [g] is present in the other; the third utilizes both. The corpus design makes it feasible, not only to identify areas of accommodation, but to grasp the context, enabling a sociolinguistically informed analysis of the speakers' linguistic behavior.

Keywords: multilingualism; corpus design; socio-phonetics; African languages; social networks.

1. Introduction[1]

Despite evidence that corpora have existed for four centuries (Kennedy 1998) and linguistic research has been gathered from speakers in the world's remotest

[1] This work was funded by SOAS Linguistics Professor Friederike Lüpke's Leverhulme Trust Research Leadership Award Grant, "Crossroads – Investigating the unexplored side of multilingualism". I am indebted to Professor Lüpke and the contributions of the Crossroads Project team members, Alain Christian Bassène, Alexander Cobbinah, Cheikh Sadibou Sambou, Samantha Goodchild, Chelsea Krajcik, Tricia Manga, Anne-Laure Vielle, Rachel Watson, and Miriam Weidl, to the transcribers and participants at the Crossroads, and to the participants of 46th Poznań

settings for nearly the same time, abundant raw data have remained inaccessible to most living outside those languages' speaking communities until this century. Since digitized corpora have been available since the 1960s (Gibbon et al. 1997), and many researchers are now required to deposit their data in a publicly available online archive, we cannot attribute the impasse solely to a lack of technology. The main cause has been an absence of time-aligned, digitized annotations, minimally including a phonetic transcription and translation into a more widely understood language, a necessary but challenging component of a corpus of a relatively under described language.

The first step in the process of making a language accessible outside of its speech community is establishing a writing system. Only slightly more than half (3,866 out of 7,099) of the world's living languages has a writing system (Simons and Fennig 2017), and only one-fourth (26 percent) of the world's adult population is literate (UNESCO 2015), so even established writing systems are unlikely to be widely used.

In common practice, spoken languages are first written by a missionary or one or a team of linguists documenting related languages as part of a larger project. The primary aim of these documentation projects is often to produce a grammatical description and lexicon or an analytical dissertation on a particular feature of a given language. The data are gathered using elicitation methods ranging from the word to the paradigm level (for criticisms on traditional language documentation methodologies, see Lüpke and Storch 2013; Newman 2013). The most natural conversation recorded is usually a requested narrative or performative demonstration. Few linguists have adequate resources (time, funding, talent) to become proficient in speaking the language of inquiry or to collect naturally occurring conversational data with an adequate transcription and translation into a more widely understood language. With limited budgets, most documentary linguists transcribe their own data, sometimes with the assistance of their language consultant(s).

Therefore, most raw documented data are deposited in an archive without annotation and, consequently,without search capabilities. The SOAS Endangered Languages Resource Project and DOBES archives provide notable exceptions; some of the files housed at the Paradisec archive are also transcribed and

Linguistic Meeting. Additionally, I would like especially to thank SOAS Crossroads Postdoctoral Fellow Rachel Watson and the two anonymous reviewers of an earlier version of this paper for their detailed and helpful feedback and suggestions. The views and opinions expressed in this article are those of the author and do not necessarily reflect those of the SOAS Crossroads project or its members.

viewable via EOPAS, University of Melbourne principal investigators Nick Thieberger and Rachel Nordlinger's web-based interlinearization viewer.

Although digitized annotated corpora were once created solely by computational linguists, now, more user-friendly linguistic documentation software has put the task within reach of most researchers. This capability is a great step forward. The Somali Language Corpus, compiled and annotated mostly by one speaker of a relatively well-known African language with a long tradition of writing, represents an amazing achievement. However, this amount of work is untenable for most researchers, especially given the considerations listed above.

The benefits of a corpus of spoken data, as opposed to an archival depository or lexical database (differences are outlined in O'Keeffe and McCarthy 2008), are enormous, especially for researchers unfamiliar with the target language(s). One crucial difference is a corpus of speech data can be searched, and the data and metadata provide crucial information about the context of the speech event and the speaker. While resources, such as Segerer and Flavier's (2011–2016) searchable lexical database of African languages, are invaluable to linguistic researchers, relying solely on words without context can lead to erroneous interpretations.

The design of a spoken data corpus of an under-described language differs sharply from corpora with huge inputs of written data from a widely studied language such as in English, the British Academic Written English (BAWE) provides one example, or those with both written and spoken texts, such as the Corpus of Contemporary American English. As Lüpke (2005) discusses in detail, corpora of less-widely spoken languages without a written tradition may be smaller, but they are equally, perhaps more, interesting and useful. A specific advantage of smaller corpora is the ability to conceptualize sampling quite differently than Biber (1993) recommends for corpora. Illustrated by the case study featured in here, representativeness notwithstanding, a small corpus can inspected in its entirety at a level impractical if not impossible in large corpora.

Further, a collection of natural language conversations by multilingual speakers, such as the Crossroads Corpus presented here, entails its own set of considerations. Mikhailov and Cooper (2016) give examples of large, online, publicly available multilingual corpora; the parallel bilingual Evrokorpus has about 240 million words. However Schmidt and Wörner (2012) point out that parallel multilingual corpora differ from data collections containing conversations gathered in multiparty speaking environments, requiring different structuring considerations. Additionally, the contributions to Ortega et al.'s (2016) edited volume mine structured corpora composed of data gathered in naturalistic

settings, or "multilingualism in the wild" (Achard and Lee 2016), to propose new interpretations of language "blending" as opposed to code-switching.

The presentation of the Crossroads Corpus, compiled and managed by the author from March 2016–March 2017, contributes to the growing discussion of methodological approaches to the design of an annotated corpus of spoken multilingual data. The author's background is not in computational or corpus linguistics; the design and illustrative study presented here reflect a discovery process. An introduction to the Crossroads Project and relevant demographic details of the target study population are presented in Section 2. The paper then emphasizes the techniques and tools used to assemble the corpus (Section 3) and considerations regarding its representativeness in Section 4, followed by an illustration of how the author interacts with it to investigate hypotheses concerning accommodation practices witnessed during fieldwork in Section 5. The author, a field linguist with an interest in theoretical phonology, hopes researchers with a similar background will gain insights into the resources available for corpus compilation and the ways they can be used to produce searchable, analyzable data for meaningful results. The results of the case study, presented in Section 6, touch on the need for data collection balanced with corpus compilation methods.

2. Background

A diverse linguistic area in South-western Senegal is the focus of Principle Investigator SOAS Professor Friederike Lüpke's ongoing research project, "Crossroads – investigating the unexplored side of multilingualism". The locus of the linguistic diversity is located at a physical crossroads: the meeting point of three villages. Geographically, the crossroads is positioned where a paved highway which runs north-south in between the adjacent villages Djibonker and Brin meets an unpaved path that leads to Essil, a village located amongst a larger group of ten Jóola villages known as The Kingdom, approximately one kilometer perpendicularly to the west.

Each of the villages situated at the crossroads is, at least nominally, associated with its own patrimonial language (based on the concept of "patrimonial deixis" (Lüpke 2016a), as opposed to matrilineal language). The three villages and their patrimonial languages are Djibonker: Baïnounk Gubëeher, Brin: Jóola Kujireray, and Essil: Jóola Banjal. Linguistically, two of the languages, Kujireray of Brin and Banjal of Essil, are classified as Jóola languages, while Gubëeher

of Djibonker is classified within a more distantly related Baïnounk grouping (Pozdniakov and Segerer, in press).

However, the linguistic situation at the crossroads cannot be simply summed up by the addition of its parts. Casamance, the wider geographic area in which the crossroads is situated, has its own long-standing linguistic diversity (Dreyfus and Juillard 2005) which contributes to the language landscape and thus to individuals' multifaceted language use (Cobbinah 2010; Lüpke 2016a). Discussed in detail below in Sectoni 4.2, corpus data illustrate there are at least 18 identified languages spoken at the crossroads and its surrounding area.

Corpus data are being gathered not only at the actual crossroads, but in the region as a whole. The Crossroads Project investigation focuses on the interactions among the three languages spoken directly at the crossroads and those with which the speakers are in contact, with a goal of determining which areas the languages influence each other least and most in structure, lexicon and speech-accompanying gesture, as well as the ideologies that underlie speakers' language use patterns.

A given participant who may have three to five languages at his or her disposal is constantly having to choose from which language to communicate. The degree of what Lüpke (2016c) refers to as "small-scale multilingualism" depends on various factors, some of which are touched on in this paper. Additionally, whether by a common lineage or long-standing contact, the crossroads' inhabitants share about one-third of their lexical inventory (based on calculations from Crossroads Postdoctoral Fellow Rachel Watson's compilation of a 1300 item comparative word-list).

A key component of the study of multilingualism at the crossroads is finding a way to delineate different spoken languages. Lüpke (2016b) and Goodchild and Weidl (2016b) demonstrate the complexities of this goal with examples of speakers' proclaimed intertwined identities. Thus, Watson (in press) has also been working towards establishing a relevant methodology based Rosch's (1973) prototype theory. One variable that has been identified as constituting a prototypical pronunciation that distinguishes the two Jóola languages is word-initial velar plosives. Further, Hantgan (2017) argues that the use of word-initial velar plosives indexes (cf. Silverstein 2003) identity.

An exposition of the Crossroads Corpus is given in the following Section 3 and 4, followed by a case study (Section 5) involving the use of velar variable and its consequences for corpus compilation methodologies and multilingual contexts.

3. Corpus overview

This section gives an overview of the corpus. The overall contents of the corpus are presented in Section 3.1 and its organization in Section 3.2. The linguistic data management tools used to compile and to search it are presented in Section 3.3, and the specific tier structure that forms the corpus backbone is described in Section 3.4.

3.1. Contents

The Crossroads Corpus, compiled and managed by the author with the assistance of three research assistants from March 2015–March 2016, includes just over 100 hours of spoken data (101:39:17 of audio, 47:00:03 of which have accompanying video) from 18 different languages. It is comprised of 516 annotated recording sessions, containing approximately 182,963 words[2] gathered during fieldwork conducted by the Crossroads project's eight researchers (three postdoctoral fellows, three London-based and two Senegal-based PhD students) since December 2014. It also contains data gathered since 2008 by two of the project's postdoctoral researchers, Cobbinah (2013) and Watson (2015), during fieldwork for their PhD theses on Baïnounk Gubëeher and Jóola Kujireray, respectively.

Currently, the Crossroads team is using a working corpus that comprises data, not only from the three villages, but also collected by Lüpke's (since 2008) investigations into the Baïnounk language Gujaher in a different area of the Casamance, which serve as control data for comparison with the Baïnounk data gathered at the crossroads in Djibonker as the speakers of both regions are in contact with different languages.

The larger corpus also includes data collected by Serge Sagna, a linguist originally from Essil, for his thesis Sagna (2008), archived with the Endangered Language Documentation Project at SOAS. The entire corpus contains 306 hours of audio recordings (.wav format), 119 hours of video (.mp4 format), and 134 hours (52 video and 82 audio) that have been transcribed, representing a total of 621 annotated files.

Most of the recordings in the corpus were annotated by a team of five residents of the crossroads area: Laurent Manga and Lina Sagna are from Djibonker; Aime Cesaire Biagui is from Brin; Davide Sagna is from Essil; and Jérémie

[2] The entire corpus has yet to be parsed at the word level. This calculation comes from the 33,882 transcribed annotations with an average of 5.4 words per annotation.

Fahed Sagna spent part of his childhood in Essil but currently lives in Djibonker. In addition to providing transcriptions, time-aligned by speech utterance, the transcribers tag each utterance with an identification of language and participant. The case study presented in the second portion of this paper looks into the advantages and disadvantages of the transcribers' ideologies concerning multilingual language use.

3.2. Organization

Following the workflow composed by Crossroads postdoctoral fellow Rachel Watson, annotated recording session files are transferred to SOAS in London for the corpus manager and assistants to integrate into the corpus structure.

Discussed in more detail in Section 3.5, in order to facilitate the use of the corpus, the annotated recording sessions are partitioned into one of four communicative events (cf. Himmelmann 1998). The following subsections describe the content of each of the genres and accompanying annotated data included in the Crossroads working corpus (excluding those of the wider corpus containing Lüpke's and Sagna's data).

3.2.1. Observed communicative events

Observed communicative events in the corpus include the Crossroads Social Network Study and Sociolinguistic Study of Multilingualism. For reasons described in Paragraph 1, these data comprise the most naturally occurring in the corpus, a component of language documentation rarely captured, and thus truly make the corpus unique in this respect.

(1) Social network study

Following a similar study conducted by Beyer (2015), ongoing since November 2016, two participants from each of the three villages are recorded while wearing a portable digital recorder for an entire day as they go about their normal activities and interactions. The researcher then debriefs the participant, asking him/her with whom s/he interacted during the day and if any sections of the recording should be omitted for ethical or confidentiality reasons. These recordings are spliced into ten-minute cuts that are then given to the transcriber best suited for the language(s) included in the recording.

(2) Sociolinguistic study of multilingualism

Cheikh Sadibou Sambou, Samantha Goodchild, and Miriam Weidl, currently conducting research for their PhD theses, are also focusing on recordings gathered in naturalistic settings: within the household, outside of local shops, during work in the rice fields, and building houses. These recordings are often accompanied by video recordings, which add knowledge on turn-taking and other subtleties of multilingual conversations not captured by audio content alone.

(3) Dissertation fieldwork

Alexander Cobbinah, former Crossroads postdoctoral fellow who, during his fieldwork for his doctoral dissertation on Baïnounk Gubëeher, learned to speak the language with enough fluency to be able to document naturally occurring conversations in and around the village of Djibonker which have been included in this communicative genre.

3.2.2. Staged communicative events

Staged communicative events are subdivided into narratives and experiments. Unlike elicited data, which is most often inherently bilingual (the researcher asking for translated lexical or phrasal items) staged communicative events may, and often do (see Section 6), represent monolingual language usage.

Narratives, such as story-telling, are a common type of speech event found in the area. Experiments are those in which a participant has been asked to perform a specific task, but normally no other constraints have been placed on the speaker's language use.

(4) Narratives

The Crossroads Corpus includes Alexander Cobbinah's contributions to a study of Baïnounk language and culture, the DOBES Project on Baïnounk language and culture.

Current Crossroads Postdoctoral fellow Rachel Watson's is also proficient in the language through her PhD dissertation research on Jóola Kujireray and thus narratives she obtained are included in this genre.

(5) Experiments

Rachel Watson performed tests, primarily based in Brin, for the Spatial Language and Cognition in Mesoamerica Project (PI University of Buffalo Professor Jürgen Bohnemeyer) to assess linguistic specifications of spatial relations among objects and meronymy. A selection of these recordings are included in the Crossroads Corpus.

Chelsea Krajcik and Tricia Manga's ongoing experiments on gesture and deixis for their PhD theses are included in the Crossroads Corpus.

Alexander Cobbinah used the Pear Story (Chafe 1980) to determine the frequency of argument ellipsis among the three Crossroads languages. These audio and visual recordings are incorporated into the Crossroads Corpus.

Abbie Hantgan asked participants to describe short video clips with targeted lexical items to another participant in the language of their choosing. These sessions are included in the Crossroads Corpus.

3.2.3. Interview

Although many interviews were collected in French, the former colonial language of Senegal, a few were collected in the area's languages and are included because they provide examples of a genre not found among the other types of data collected, and also because these recordings enhance the participants' metadata.

(6) Social network study interviews

As noted above, six participants (two from each village) are primarily involved in the Crossroads Social Network Study. Interviews were conducted with the six participants as well as two levels of their social networks. Those interviews which have been transcribed are included in the Crossroads Corpus.

(7) Sociolinguistic interviews

Sociolinguistic questionnaires were gathered by project PhD students for their sociolinguistic studies of the area, and by Jérémie Fahed Sagna of the transcription team, who is currently working toward his Master's at the University of Ziguinchor.

3.2.4. Elicitation

Elicitation is an essential element of any language documentation project. While conversation represents one of the most natural types of language use, elicitation is also necessary for a linguist to ascertain the composition of the language's grammar, including the phonology, morphology, syntax, and lexical semantics. For a researcher who wants to interact with the corpus but is unfamiliar with the area's languages, this genre is also the best place to begin. Additionally, as illustrated by the case study, elicitation was a necessary step in determining the phonemic inventories of the area languages and thus establish an objective point (cf. Watson's, in press) Crossroads prototype) from which they may be distinguished.

(8) Lexical elicitation

Another component of the Crossroads project is obtaining a comparative wordlist for the area. Rachel Watson has diligently gathered 1,300 items from each of the three Crossroads languages. The author targeted phonetically similar lexical items with the same meaning over the three languages for a study of pronunciation that thought to constitute a "foreign accent" (Hantgan 2016).

Bangor University's Sarah Cooper collaborated with SOAS researchers to collect data to understand intonation among the crossroads languages (Goodchild et al. 2013). The data files' tier structure (see Section 3.3) was modified to be included in the Crossroads Corpus.

(9) Grammatical elicitation

Watson and Cobbinah's fieldwork data, gathered to complete reference grammars for Jóola Kujireray and Baïnounk Gubëeher, are included in the corpus.

3.3. Tools

ELAN, annotation software created by the Max Planck Institute for Psycholinguistics, provides the means by which the Crossroads working corpus is transcribed, translated, annotated, accessed, and searched. The Senegal-based team

of transcribers uses an ELAN template designed in conjunction with Endangered Languages and Documentation Project digital archivist Sophie Salffner. The ELAN template enables them to annotate the transcription, translation, and speaker and language identification for the sound and/or video files produced during fieldwork. Figure 1 shows an example of an ELAN-incorporated corpus file; more detail about the tier structure is presented in Section 3.4.

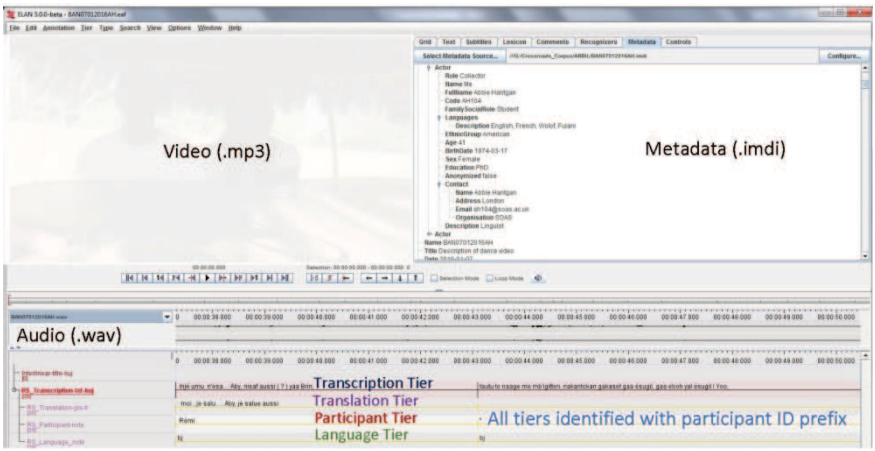

Figure 1. Example ELAN file.

Once the transcriber has processed a file and sent it to a team member at SOAS in London, the corpus manager and assistants link each ELAN file to its respective media (audio and/or visual) and metadata files on the SOAS server. Due to restrictions on what may be linked to ELAN and uploaded to the remote server (discussed below), all video files are converted to .mp4 and audio files to .wav. Metadata files are created in Arbil, a program developed by the Max Planck Institute for Psycholinguistics to create, search, and organize metadata files. Using .wav format for audio files allows them to be easily opened for detailed examination or exported to Boersma and Weenink's (2017) phonetics software package, PRAAT (version 6.0.29).

Some of the corpus files have been exported for interlinearization and incorporation into the 5,723 lexical- entry, multilingual dictionary (with an excerpt included in Appendix B). The lexical database was compiled by the author using Watson, Cobbinah, and Hantgan's databases for Kujireray, Gubëeher, and Banjal, respectively with the SIL lexical database program, Fieldworks, otherwise known as FLEx. Following Gaved and Salffner's (2014) teaching set, ELAN files, once re-imported and time-aligned, can be searched in ELAN at the level of the lexeme, morpheme, or part of speech, as shown in Figure 2.

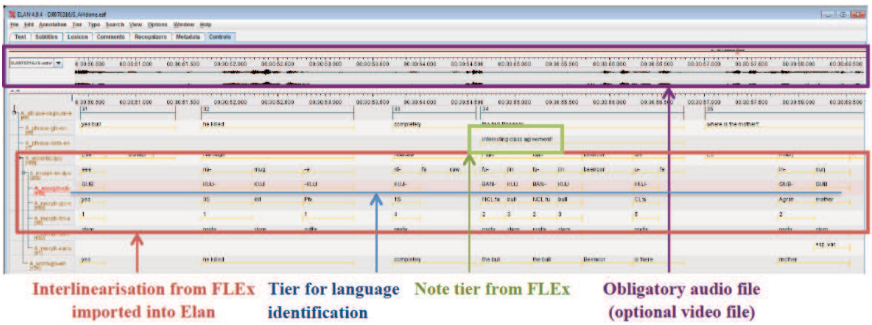

Figure 2. ELAN corpus example.

The working Crossroads Corpus is currently accessible only to researchers on the Crossroads Project. It is housed on a local SOAS server and shared with the entire London-based Crossroads team. Using exported .imdi files from Arbil, the author and assistants have uploaded a remote copy of the corpus to the SOAS Endangered Language Archive (ELAR) LAT server. Paralleling the local corpus copy's nonhierarchical folder structure (see Section 3.4), each researcher has his or her own node with a folder, including a session's recording (audio and/or video), transcription, and metadata. Figure 3 presents an example.

For the time being, the corpus cannot be accessed remotely. An online corpus will be released for public use at the end of the project.

Other tools developed for the Crossroads Corpus are University of Glasgow professor Dale Barr's package ELAN for R and Max Planck for Psycholinguistics software developer Peter Wither's genealogy software KinOath.

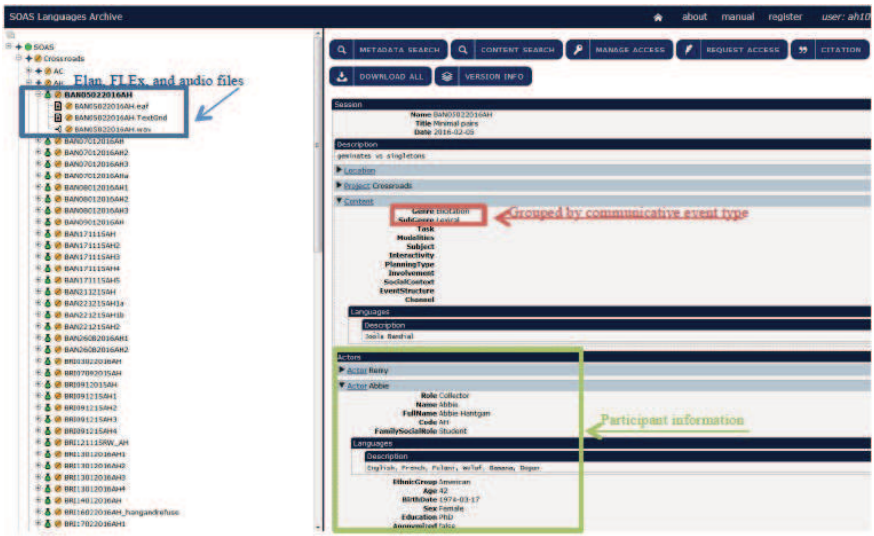

Figure 3. Online corpus at ELAR.

3.4. Structure

As described in the previous section, the corpus is composed of four communicative genres, each with its own sub-genres. The files are structured to easily permit selection of a given genre (its folder) as a domain for searching in ELAN (discussed in detail in Section 3.5). With the assistance of Professor Barr and his R package ELAN, the author searched the project's metadata and created a separate folder and subfolder for each communicative genre/sub-genre. Each subgenre folder contains the individual ELAN files, linked to the audio, visual, and metadata files, which are housed in separate folders for ease of search and duration measurements.

Each ELAN file within the corpus has been modified to adhere to a strict, hierarchical structure. As discussed below in Section 4.1, each participant represented in the corpus has been assigned a unique ID. Shown in Figure 1, the participant ID, a combination of the participant's initials and in cases of duplicate initials, an accompanying number, is shown as a prefix to that participant's tier name in the ELAN file. To ensure accurate identification, the participant ID is also represented in the Arbil metadata file linked to the ELAN file; the two must match.

Each participant in an ELAN file has at least three tiers: a transcription, translation, and language tier. In terms of the finalized corpus incorporation, the fourth tier that contains the speaker information becomes superfluous as it is copied into the ELAN tier structure by the corpus manager and assistants. Additional tiers are added if a file has been exported and reimported from FLEx (as shown above in Section 3.3). Ideally, although see Crossroads team members Goodchild and Weidl (2016a) presentation for practicalities of transcriptions done by speakers not trained in linguistic documentation methods, the Senegal-based transcriber team segments an ELAN file by utterance, and then transcribes what they hear into the Transcription tier. Then, in the Translation and Language tiers, the transcribers provide a French translation and identify the language they think the speaker is using (see Section 5) with project-standardized two-letter abbreviations.

3.5. Searching

The Crossroads Corpus was designed to facilitate searches. At present, searches can be performed in ELAN at the phoneme, word, or utterance (phrase) level, and an investigator can search a given participant's utterances across multiple files or examine any number of instances(s) of a particular language. Detailed instructions on how to search the Crossroads Corpus are provided in the Crossroads Corpus Manual.

The most efficient way to search the corpus is to use the Multiple Layer Search tool in ELAN. Through the creation of domains based on the existing communicative genres (see Section 3.4), searches of a specific variable can be performed for a given participant (or tagged language) across all of the files in which they appear. Figure 4 illustrates a search of all of one participant's word-initial uses of the voiceless velar plosive and lists the identified languages for the given utterance.

Clicking on the results of the search will produce a file and its annotation. The utterance can then be viewed in the context in which it was spoken, and the linked metadata can be consulted to verify whether the utterance is expected based on the participant's background.

ELAN files with multiple participants can be difficult to analyze. Without video evidence, it is often unclear to whom a participant is speaking. In ambiguous cases, exporting an ELAN tier structure to an Excel or SPSS file (instructions provided in the Crossroads Corpus Manual), eases line-by-line comparison.

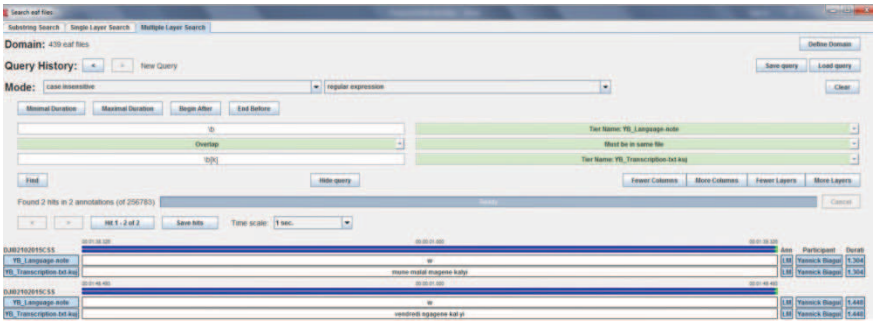

Figure 4. Multilayered search in ELAN.

The current study used a combination of word, variable, and participant searches to obtain the results discussed in Section 5.

4. Representativeness

As mentioned in the introduction, Section 1, the manageable size of the Crossroads Corpus in comparison to more widely used corpora can be considered an advantage. Rather than exclude data or make assumptions that might restrict potential findings, the entire working Crossroads Corpus can be inspected on a detailed level unparalleled in large corpora, providing at least an equally encompassing view of the target community and associated languages without relying on a compilers' sampling restrictions. The following subsections outline the demographic makeup of the participants in the study.

4.1. Participants

In total, 211 participants (83 female, 128 male) are represented in the corpus. Since much of the data were gathered by individual researchers consulting with trusted individuals for lexical, experimental, and narrative purposes, the duration of transcribed utterances for some participants is longer than for others. Certain types of studies, such as frequency counts, could be affected by this disparity. Gries and Berez (to appear) advocate for resolving disproportionality by creating samples. With regards to the present study, in cases where relative pro-

portionality is a crucial factor, samples have been created (see Section 4.2). Asymmetric demographics can also be resolved by examining each utterance in its own context with the speaker's accompanying metadata.

4.2. Languages

A total of 17 spoken languages and one sign language are represented in the corpus. Figure 5 illustrates the relative utterance durations for each language found in the corpus based on the transcribers annotations.

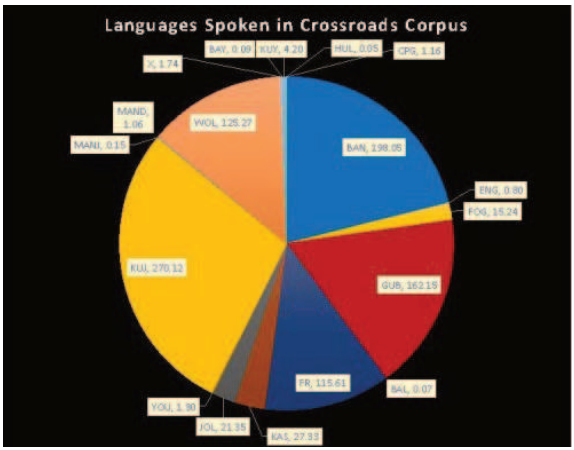

Figure 5. Language representation in the Crossroads Corpus.

The figure shows that Jóola Kujireray is the most represented language in the corpus; however, the fact that more audio was recorded in Djibonker (44 hours) than in Brin (32 hours) illustrates a village cannot be directly associated with its patrimonial language. Banjal annotations also exceed those of Baïnounk Gubëeher, but the recording time in Essil was far less than in the other two villages (27 hours). A reason proposed to explain the perceived discrepancy is that Djibonker community members accommodate their choice of language to that of the surrounding villages, as discussed in detail in (Cobbinah et al. 2017; Hantgan 2017).

A component of the Crossroads Social Network Study gathered data on participants' reported speech repertoires. Focusing on the three languages associated with the immediate crossroads area, we see a somewhat surprisingly low number of participants (29 out of 113) who claim proficiency in all three. Although speakers' reported proficiency may not be a reliable diagnostic of speech usage (see Goodchild (2016) for issues surrounding both self-reported and researcher-prompted language repertoires), it is still worth acknowledging in terms of the speech accommodation patterns discussed in the case study.

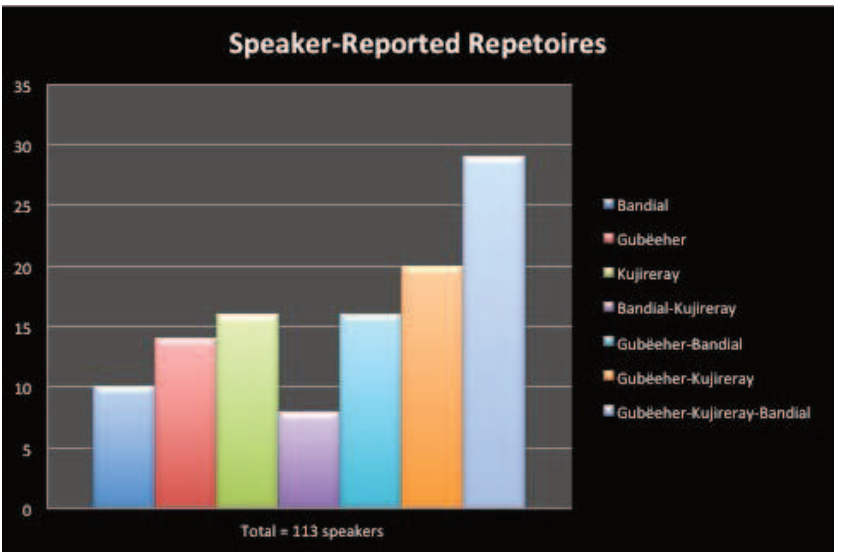

Figure 6. Speaker reported repertoires from social network study.

To compare these reported repertoires with actual language usage, the author selected approximately ten minutes randomly from files classified within the observed communicative event genre with at least five participants, from each of the three villages. As Figure 7 shows, the relative proportionality differs, not only from the reported repertoires, but also depending on the village in which the participants are speaking.

Note that even though both Kujireray and Banjal are geographically and genetically proximate Jóola languages, few sampled speakers at the crossroads

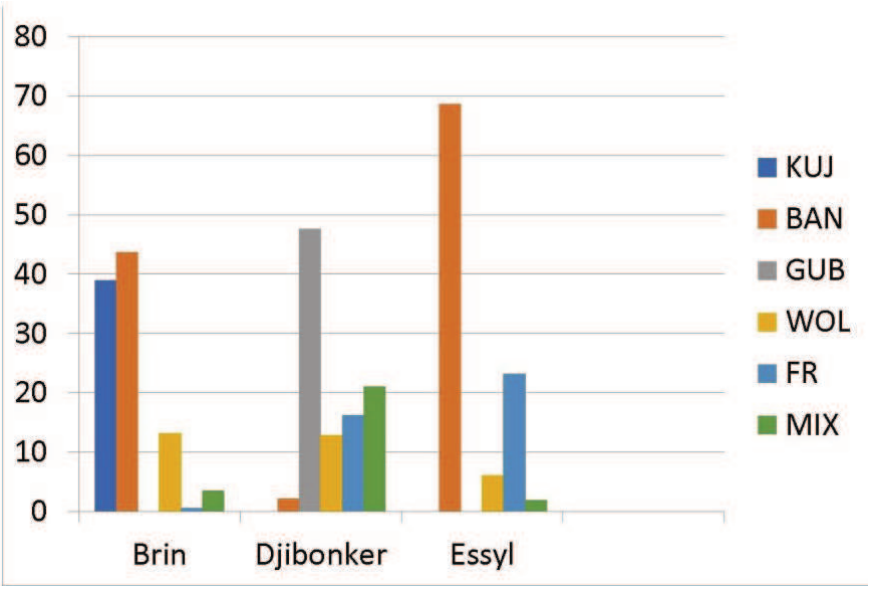

Figure 7. Language usage in observed communicative events sample.

claim to speak both. In the data obtained from Brin, we see a nearly equivalent amount of time devoted to both Jóola languages, although in Essil, the same generalization does not hold. The Figures 6 and 7 illustrate the need to look at the context of, not only the crossroads as a whole, but individual villages. Sagna (2016) confirms that speakers from Essil tend to mix languages less than the communities of Djibonker and Brin; Lüpke (2016a) argues, based on historical sources, that speakers from Djibonker accommodate others in line with their first-comer status.

4.3. Genres

An important observation enabled by the organization of the Crossroads Corpus is the relative representation of the communicative genres presented in Section 3.2. Table 2 compares the total amount of time represented in the media files (hrs:min:sec format) to the number of transcribed ELAN files.

Table 2. Crossroads Corpus contexts.

File Type	Observed	Staged	Interview	Elicitation
Transcription	110	331	27	48
Audio	16:55:58	31:23:04	04:46:01	05:21:35
Video	5:14:59	31:58:17	01:55:07	02:26:41

The largest proportion of data belongs to the genre of staged communicative events. Following suppositions put forth by Green and Abutalebi (2013), we do see that the use of multiple languages is in direct proportion to the naturalness of the speech event's genre. As described in Section 1, the corpus is organized by communicative genre ranging from the most unnatural (elicitation) to the most natural (observed) type of speech event; as Biber (1993) asserts, a corpus should include all types of speech events found in the community. By design, many of these genres include only monolingual speech events, which is one source of unnaturalness. However, not all sessions recorded in unnatural settings, especially narratives and experiments, are inherently monolingual. Therefore, the current study, presented in the following sections, attempted to identify instances of multilingual speech in the corpus, irrespective of communicative genre.

5. Case study

While the Crossroads Corpus does not constitute a phonological corpus like that presented by Durand et al. (2014), with ELAN's ability to extract portions of an audio .wav file in PRAAT, the author was able to use it to conduct a thorough investigation into the phonological patterns of the three primary crossroads languages. This section provides an in-depth discussion of a case study of accommodation patterns at the crossroads with a focus on a specific variable: stem-initial velar plosives.

As a means by which we may continue to present the methodological approaches used in the design of the Crossroads corpus, Section 5.1 evaluates hypotheses about language use by exploring the conversational data it houses. In Section 5.2, the author describes how she used the ELAN corpus to determine the phonemic inventory of the three languages. In line with Voormann and Gut's

(2008) recommendations for cyclical corpus creation and analysis, results of data inspection fed the creation of new hypotheses and vice versa.

5.1. Hypotheses

As indicated in Section 2, Watson (in press) has been working towards establishing a methodology for determining what constitutes a prototypical feature and therefore makes it identifiable to others as a specific Crossroads language. An observational study of the three main languages spoken at the crossroads has shown the potential for phonetic divergence or convergence in accommodating speech patterns among velar consonants. A dichotomy is present stem-initially between the Jóola languages; in Banjal, the consonant [g] is found to the exclusion of its voiceless counterpart [k]. In Kujireray, the opposite is found; [k] to the exclusion of [g]. In the third language, Baïnounk Gubëheer, both the voiced and voiceless velar plosive may appear at the beginning of the word.

Therefore, the question emerged: How would multilingual crossroads speakers pronounce a velar variant, especially if it is not phonemic to their language. The literature on foreign accent development among speakers who learned more than one language from an early age and relative voice-onset time (VOT) in bi/multilingual speakers (Ayala 2011; Beyer 2015; Chang 2013; Flege and Eefting 1987; Fowler et al. 2008; Sancier and Fowler 1997) indicates that multilingual speakers should be capable of pronouncing the target variants with minimal phonetic divergence from the target output.

Because of the velar plosives non-phonemic status stem-initially in the Jóola languages, /k g/ were predicted to emerge categorically as [k] in Jóola Kujireray and [g] in Jóola Banjal, irrespective of an individual's geolinguistic background (the place they first learned a language). Velar plosives in Baïnounk Gubëeher were predicted to surface according to their underlying specification /k/ as [k] and /g/ as [g].

Since, as shown in Section 4.1, inhabitants of Djibonker accommodate to others rather than the reverse, an opportunity to align their pronunciation with that of the prototypical form for either of the two Jóola languages presents itself in the context of stem-initial velars. The results of a preliminary study, reported in (Hantgan 2017), indicate that, at least in the context of greetings, not only those from Djibonker, but many crossroads speakers may make a conscious choice to align with a specific linguistic identity through their pronunciation of stem-initial velar plosives.

The Crossroads Corpus was consulted to determine how multilingual speakers pronounce word-initial velar plosives in their day-to-day speech and to confirm the observation of a phonemic split between voiced and voiceless word-initial velar plosives in the Crossroads Jóola languages.

5.2. Metholodolgy

First, words that are phonetically and semantically equivalent across the three languages were elicited by the author to determine if speakers' relative VOT differed according to the language in which they were most proficient (Hantgan, in prep.).

Word-initial velar plosives are found among noun stems that are prefixed with a nominal classifier. Examples found in the comparative word-list (see Section 3.2), that have the same (or a similar) form and meaning across the three languages are shown in Table 3.

Table 3. Crossroads lexical correspondences.

GUB	KUJ	BAN	ENG
ka-law	ka-law	ga-law	'to ask'
ka-lak	ka-lah	ga-llax	'field'
gʊ-pəl	ka-pəl	ga-pəl	'skin'
gʊ-fɔs	ka-fɔs	ga-fɔs	'grass, weed'
gʊ-məy	ka-məy	ga-məy	'eyelash/brow'
gʊ-bɪfʊm	kə-bifum	gə-bifum	'fan'
gʊ-ñaak	ka-ñahah	ga-ñañ̃ax	'palm wine funnel'

It was hoped that these specific words and other phonetically similar items could be examined in the corpus in order to determine if the pronunciation altered in various contexts. Unfortunately, the target words were found with low or no occurrence in the corpus, save when elicited for the comparative study of pronunciation. Consequently, an ELAN search using regular expressions (following Mosel 2015) was performed across the corpus data for other instances of stem-initial velar plosives [k g] among the three crossroads languages: Jóola Kujireray and Banjal, and Baïnounk Gubëeher.

In order to avoid the issue of circularity, (at present, the only way to search the corpus by language is through the annotations of the transcribers), the goal

of the corpus-wide search of stem-initial velars was revised to confirm the hypothesis that the velar dichotomy was indeed emblematic of a language's identity. In other words, any given speaker could potentially be speaking any given language at any given time at the crossroads. Further, the transcribers ability to interpret language is not limited to any one language. The question becomes, How does a transcriber identify a crossroads language as such? The theory is that, based on our understanding of the phonemic status of stem-initial velars, a salient attribute in the identification of the Jóola languages at least would be the presence of an initial [k] or [g].

Therefore, tokens of stem-initial velar plosives were separated by language, and only those identified by the transcriber as one of the three Crossroads languages were kept in the sample. Clitics, nonintegrated loan words from French or Wolof, and ideophones were removed from the results so that the remainder of words were nouns in a class with a velar-initial prefix or verbs prefixed with the 3rd person plural pronominal. A total of 75 tokens were used for the study: 25 tokens for each of the three primary crossroads languages across 33 randomly selected participants, 15 of whom were male and 18 female.

Because the tokens were chosen randomly, without consideration of communicative genre, the study weighed heavily in the direction of more naturally occurring speech events. The following subsection addresses this imbalance.

Table 4. Distribution of tokens across genres.

Observed	Staged	Interview	Elicitation
20	36	8	10

All utterances used in the study were examined in PRAAT. The FLEx multilingual language database was used to interlinearize each utterance, and any lexical items with which the author was unfamiliar were discussed with the researcher most familiar with the identified target language to confirm its meaning.

6. Results

The results, displayed in Table 5 showed near categorical identification of the expected (prototypical) variant for the intended language.

Table 5. Velar variant distributions across languages.

TOK	KUJ	BAN	GUB
k	24	1	7
g	1	24	18

In all but three instances, a stem-initial voiceless velar plosive [k] was associated with an instance of Jóola Kujireray and that of the voiced velar plosive [g] with Jóola Banjal. Additionally, the transcriber's intuitions matched those of the linguistic researcher: the 24 words with initial velar voiceless plosives aligned both semantically (a corresponding lexical item listed for that language in the multilingual language database) and phonetically (the transcribed [k] or [g] was confirmed as such).

6.1. Jóola phonemic velar split

The one instance of a stem-initial [g] identified as being Jóola Kujireray by the transcriber was uttered by a resident of Djibonker who was, at the time, participating in an observed conversation taking place in Brin. While it is possible that the participant used a non-prototypical pronunciation of the word, it is more likely that she was speaking in Jóola Banjal but that the transcriber identified the language as being Jóola Kujireray based on the participant's setting.

The instances of a stem-initial voiced velar plosive being associated with a Jóola Banjal utterance is also worth exploring, as it also took place in observed conversational context in Brin, this time in front of a shop. In this instance, the author and other researchers disagreed with the transcriber's interpretation: the plosive was voiced rather than voiceless.

These interpretations provide two valuable insights. As said in the introduction, an asset of the Crossroads corpus is its ability to be inspected at the utterance level with the addition of the rich metadata collected over a span of almost a decade in the area by unequivocally experienced researchers. Further, the fact that the transcriber heard a [g] even though the utterance was spoken by a member of the Jóola Kujireray speaking community and labeled it as being Jóola Banjal confirms the hypothesis that the voicing specification of an initial velar plosive plays a key role in the identification of language at the level of the listener.

6.2. Baïnounk Gubëeher accommodation

As noted above in Section 5.2, stem-initial velars in Baïnounk Gubëeher may be either voiced or voiceless phonetically. There is an overall higher frequency of stem-initial [g] over [k], in part due to the frequency of the noun class prefix [gu-] in Baïnounk Gubëeher.

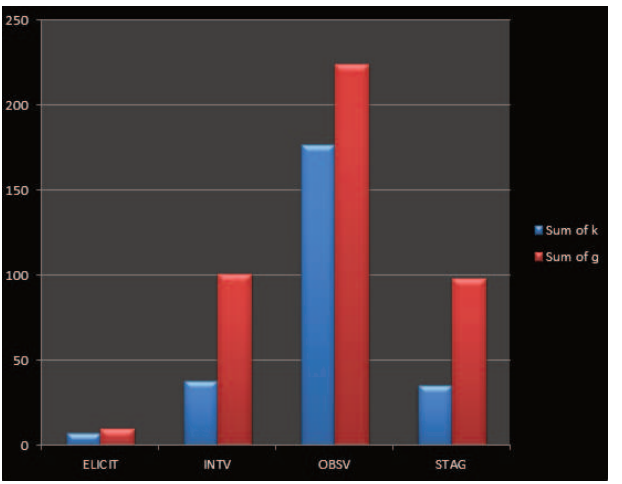

Figure 8. Preference of [g] over [k] in Baïnounk Gubëeher among all speakers/genres sampled.

The higher frequency of the voiced variant of the velar plosive stem-initially in Baïnounk Gubëeher is interesting because, although curiously not often found in the corpus, the author has observed a divergence from the expected pronunciation among crossroads speakers in one word: [gə- ~ kə-ssumaj] 'peace', used as an introduction to the greeting sequence. She argues (Hantgan 2017) that speakers at the crossroads chose whether or not to align their speech patterns with their interlocutor for reasons of identity projection. In the corpus and in observation, speakers of Baïnounk Gubëeher use the voiceless variant in greetings more often than the voiced plosive, which is unexpected, given the morpho-phonological tendency to use stem-initial voiced velars. It supports the author's hypothesis that Baïnounk speakers of Gubëeher align their speech with that of

Kujireray speackers from Brin, while Banjal speakers from Essil diverge from what they perceive as a (negative) Baïnounk identity.

6.3. Observed communicative contexts

Further, a key issue which was identified in the process of the study's development and the results is that of communicative context. Although the case study was biased for more naturally occurring communicative genres, and thus a finding as such based on these data would be considered problematic, other examinations of the corpus have revealed that, in accordance with expectations outlined by Green and Abutalebi (2013), speakers tendency to pronounce unexpected variants (speakers of Jóola Kujireray using a stem-initial voiced velar plosive) increased in proportion to the degree of naturalness of the speech event.

As described in Section 2, the corpus is organized by communicative genre ranging from the most unnatural (elicitation) to the most natural (observed) type of speech event; as Biber (1993) asserts, a corpus should include all types of speech events found in the community. By design, many of these genres include only monolingual speech events, which is one source of unnaturalness. However, not all sessions recorded in unnatural settings, especially narratives and experiments, are inherently monolingual. Therefore, the current study attempted to identify instances of multilingual speech in the corpus, irrespective of communicative genre.

To account for the unequal representation of participants' participation in the corpus mentioned in Section 4.1, only the utterances of the participants who contributed to both staged and observed speech contexts were compared for degrees of multilingual speech. Four participants were found to be active in both speech genres; Table 6 shows the durations of their contributions in minutes.

Table 6. Duration (in minutes) of speakers who participated in both staged and observed events.

PAR	OBSV	STAG
LM	40	12
GS	8	5
HPS	4	2
JHS	12	1

Again, the duration for each participant was not equal, therefore, the relative percentage of their participant was measured along with the duration of the languages spoken in the speech event (as identified by the transcriber). The author, with the help of the research assistants, exported the transcription, duration, and identified languages of the staged and observed communicative events from ELAN into text format and then re-imported the data into SPSS (*IBM SPSS Statistics for Windows* 2016). Next, we tagged these annotations for instances of inter- and intra-utterance language changes based on the language and speaker identified for each utterance. We then counted the number of times a speaker changed languages and compared these instances across the two speech genres.

Taking into account the unequal amount of time spent speaking in each genre, the percentages in Figure 9 show that the duration of single language usage all but eclipses that of multi-language usage in the staged communicative events.

Figure 9. Proportionality of language mixing in communicative event samples.

The results show that, although one language remained pervasive in each type of speech event, speakers changed languages less in the staged than in the observed communicative genre.

7. Conclusion

The Crossroads corpus is among the first corpora of multilingual spoken languages data collected from relatively unknown languages, over a long period of time. A large component of the corpus is drawn from naturally occurring, conversational data, somewhat ironically, an all but unexplored area among lesser-

known languages. Given the capability of current technology to capture and analyze speech data and then to store and organize it into a meaningful and searchable way, it is hoped that more corpora like this one will begin to emerge. This paper hopes to provide a preliminary methodology for those seeking to construct such corpora, and to provide an example of the uses of such a body of data. In particular, a corpus of this type can provide key insights into the study of phonological contrasts and contexts, within the realm of socio-phonetics and beyond.

Unique strengths of the Crossroads Corpus are its capacity to be searched and studied at a minute level and the richness of the participants' metadata, provided by integrated linguistic researchers who have not only spent extensive time in the community, gaining residents' trust, but also have learned to speak the area's languages. Additionally, each of the multidisciplinary team of researchers has contributed his or her own individual piece of the puzzle to bring our vision of an apparent opacity better into focus.

One difficulty discovered through the case study of the Crossroads Corpus is its potential towards circular introspection with regards to the identification and partitioning of languages. As Watson (2017) so eloquently elaborates in her attempts to define what it means to speak "prototypical" Jóola Kujireray, we must not attempt to define what a language is by our standards, but yet by how the speakers' perceive and produce language; only then will be gain the most crucial insights into the impetus for language use.

An avenue for a possible follow-up study might be to search for instances of word-initial velar plosives from participants who are identified as speaking in multilingual contexts, since this relies somewhat less heavily on the transcriber's identification of the language in question as it does the fact that there is an identi- fied 'switch' in language. For instance, a given participant's use of stem-initial velars in a narrative provided in a formal setting to one of the linguistic researchers could be compared with those of the same participant playing cards with friends. The difficulty of this type of comparison thus far has been the availability of finding the same participant speaking across different types of communicative genres; as noted in the previous section, few participants were found across the various genres and if so, were often speaking languages outside of the three-targeted area languages, such as Wolof, French, or other varieties of Jóola, making in depth comparative study difficult.

References

Achard, M. and S. Lee. 2016. "Toward a model of multilingual usage". In: Ortega, L., A.E. Tyler, H.I. Park and M. Uno (eds.), *The usage-based study of language learning and multilingualism*. Washington, DC: Georgetown University Press. 255–275.

Ayala, A. 2011. *Phonetic convergence: A case study of a Puerto Rican Spanish speaker (senior essay).* New Haven: Yale University.

Beyer, K. 2015. "Multilingual speakers in a West-African contact zone: An integrated approach to contact-induced language change". In: Stell, G. and K. Yakpo (eds.), *Code-switching between structural and sociolinguistic perspectives*. Berlin: De Gruyter Mouton. 237–258.

Biber, D. 1993. "Representativeness in corpus design". *Literary and Linguistic Computing 8*(4). 243–257.

Boersma, P. and D. Weenink. 2017. *Praat: doing phonetics by computer [computer program]*. Retrieved from <http://www.praat.org>. (Version 6.0.29.)

Chafe, W. (ed.). 1980. *The pear stories: Cognitive, cultural, and linguistic aspects of narrative production*. Norwood, NJ: Ablex.

Chang, C. 2013. " A novelty effect in phonetic drift of the native language". *Journal of Phonetics* 41. 520–533.

Cobbinah, A. 2010. " Casamance as an area of intense language contact". *Journal of language contact THEMA* 3. 175–201.

Cobbinah, A. (2013). Nominal classification and verbal nouns in Baïnounk Gubëeher (PhD dissertation, SOAS, London.)

Cobbinah, A., A. Hantgan, F. Lüpke and R. Watson. 2017. "Carrefour des langues, carrefour des paradigmes". In: Auzeanneau, M. (ed.), *Pratiques plurilingues, mobilités et éducation*. Edition des Archives Contemporaines.

Dreyfus, M. and C. Juillard, C. 2005. *Le plurilinguisme au Sénégal: langues et identités en devenir*. Paris: Karthala.

Durand, J., U. Gut and G. Kristoffersen. 2014. *The Oxford handbook of corpus phonology*. Oxford: Oxford University Press.

Flege, J. and W. Eefting. 1987. " Cross-language switching in stop consonant perception and production by Dutch speakers of English". *Speech Communication* 6(3). 185–202.

Fowler, C., V. Sramko, D. Ostry, S. Rowland, and P. Hallé. 2008. " Cross language phonetic influences on the speech of French–English bilinguals". *Journal of Phonetics* 36. 649–663.

Gaved, T. and S. Salffner. 2014. " Working with ELAN and FLEx together: an ELAN-FLEx-ELAN teaching set". <http://www.mpi.nl/tools/elan/tp/how-to/ELAN-FLEx-ELAN_2015-11-06.zip>

Gibbon, D., R. Moore and R. Winski (eds.). 1997. *Handbook of standards and resources for spoken language systems*. Berlin: de Gruyter Mouton.

Goodchild, S. 2016. "'Which language(s) are you for?' ' I am for all the languages.' Reflections on breaking through the ancestral code: Trials of sociolinguistic documentation". *SOAS Working Papers in Linguistics* 18. 75–91.

Goodchild, S., M.P.S. Cooper, R. Watson and A. Cobbinah. 2013. *New methods in the field and new data in the lab: Research methods in multilingualism.* London: SOAS, University of London.

Goodchild, S. and M. Weidl. 2016a. *Documentation of speakers' linguistic practices in two sociolinguistically diverse settings in the Casamance, Senegal.* (Language Documentation and Linguistic Theory 5.)

Goodchild, S. and M. Weidl. 2016b. "Translanguaging practices in the Casamance, Senegal". Paper presented at the joint KPAAM-CAM and Crossroads workshop. SOAS, London.

Green, D.W. and J. Abutalebi. 2013. "Language control in bilinguals: The adaptive control hypothesis". *Journal of Cognitive Psychology* 25(5). 515–530.

Gries, S. and A. Berez. (To apprear.) "Linguistic annotation in/for corpus linguistics". In: Ide, N. and J. Pustejovsky (eds.), *Handbook of linguistic annotation.* Berlin: Springer.

Hantgan, A. 2016. " How foreign is accent? Expressions of peace in Casamance". In: *Voices from around the world, Special issue on multilingualism in the Global South.* Cologne: University of Cologne: Global South Studies Center.

Hantgan, A. 2017. "Choices in language accommodation at the Crossroads: convergence, divergence, and mixing". *Journal of the Anthropological Society of Oxford* IX(1). 102–118.

Himmelmann, N.P. 1998. "Documentary and descriptive linguistics". *Linguistics* 36. 161–195.

IBM SPSS statistics for Windows. 2016. Armonk, NY: IBM Corp. <https://www-01.ibm.com/support/docview.wss?uid=swg21476197> (Ver. 24.0.)

Kennedy, G. 1998. *An introduction to corpus linguistics*. London: Longman.

Lüpke, F. 2005. "Small is beautiful: contributions of field-based corpora to different linguistic disciplines, illustrated by Jalonke". *Language Documentation and Description* 3. 75–105.

Lüpke, F. 2016a. "Multiple choice: Language use and cultural practice in rural Casamance between convergence and divergence". In: Knörr, J. and W.T. Filho (eds.), *Creole languages and postcolonial diversity*. Berghahn: Oxford.

Lüpke, F. 2016b. "Perspectives on small-scale multilingualism". Paper presented at the joint KPAAM-CAM and Crossroads workshop. SOAS, London.

Lüpke, F. 2016c. "Towards a typology of small-scale multilingualism". *Critical Multilingualism Studies* 4(2). 35–74.

Lüpke, F. and A. Storch. 2013. *Repertoires and choices in African languages*. Berlin: De Gruyter Mouton.

Mikhailov, M. and R. Cooper. 2016. *Corpus linguistics for translation and contrastive studies: A guide for research*. London: Routledge.

Mosel, U. 2015. "Searches with regular expressions in ELAN corpora". <https://tla.mpi.nl/wp-content/uploads/2011/12/Searches_in_ELAN_with_regular_expressions.pdf>

Newman, P. 2013. "The law of unintended consequences: How the endangered languages movement undermines field linguistics as a scientific enterprise". Paper presented at the Linguistics Departmental Seminar Series. SOAS, University of London.

O'Keeffe, A. and M. McCarthy (eds.). 2008. *The Routledge handbook of corpus linguistics*. London: Routledge.
Ortega, L., A.E. Tyler, H.I. Park and M. Uno (eds.). 2016. *The usage-based study of language learning and multilingualism*. Washington, DC: Georgetown University Press.
Pozdniakov, K. and G. Segerer. (In press.) "A new classification of Atlantic languages". In: Lüpke, F. (ed.), *The Oxford guide to the Atlantic languages of West Africa.* Oxford: Oxford University Press.
Rosch, E. 1973. "Natural categories". *Cognitive Psychology* 4. 328–350.
Sagna, S. 2008. Formal and semantic properties of the Gújjolaay Eegimaa (a.k.a Banjal) nominal classification system. (PhD dissertation, SOAS, London.)
Sagna, S. 2016. "'Research Impact' and how it can help endangered languages". *Ogmios* 59. 5–8.
Sancier, M. and C. Fowler. 1997. "Gestural drift in a bilingual speaker of Brazilian Portuguese and English". *Journal of Phonetics* 25. 421–436.
Schmidt, T. and K. Wörner (eds.). 2012. *Multilingual corpora and multilingual corpus analysis* (Vol. 14). Amsterdam: John Benjamins.
Segerer, G. and S. Flavier, S. 2011–2016. *Reflex: Reference lexicon of Africa.* Paris, Lyon. <http://reflex.cnrs.fr/>. (Version 1.1.)
Silverstein, M. 2003. " Indexical order and the dialectics of sociolinguistic life". *Language and Communication* 23. 193–229.
Simons, G.F. and C.D. Fennig (eds.). 2017. *Ethnologue: Languages of the world* (20th edn.). Dallas, TX: SIL International. <http://www.ethnologue.com>.
Voormann, H. and U. Gut. 2008. " Agile corpus creation". *Corpus Linguistics and Linguistic Theory* 4(2). 235–251.
Watson, R. 2015. Verbal nouns in Joola Kujirerai. (PhD dissertation, SOAS, London.)
Watson, R. 2017. "Deviation from the norm". Paper presented at the Fourth International Conference on Language Contact in Times of Globalization (LCTG4) workshop. Greifswald, Germany.
Watson, R. (In press.) *Languages as categories: using prototype theory to create reference points for the study of multilingual data.*

Address for correspondence:
Abbie Hantgan-Sonko
Independent research
910 Carolina Avenue
Winston-Salem, NC 27101
United States
ahantgan@gmail.com

Appendix A. Language Abbreviations

WOL	Wolof
BAN	Bandial
FR	French
KUJ	Kujireray
GUB	Gubëeher
HUL	Huluf
KAS	Kaasa
JOL	Jóola
KY	Kuwaatay
MAN	Mandinka
MANQ	Manjack
FOG	Fogny
ENG	English
CPG	Guinea-Bissau Creole
BAY	Bayot
X	Sign language
YOU	Youtou
BAL	Balante

Appendix B. Gujireray dictionary excerpt

MULTILINGUAL CROSSROADS DICTIONARY
Semantic Domain: Palm Tree

A

alen$_{KUJ}$ fualen [fualɛn] *n* palm wine market; marché du vin de palme

B

bes$_{KUJ}$ kabes [kabɛs] *n* palm leaf; feuille de rônier
bes$_{GUB}$ gubes [gʊbɛs] *n* palm leaf; feuille de rônier
bes$_{BAN}$ gabes [gaβɛs] *n* palm leaf; feuille de rônier

C

combo$_{KUJ}$ [cɔmbɔm] *n* palm oil dish; plat du huile de palme
conkom$_{WOL}$ [coŋkom] *n* palm wine (fresh); vin de palme (frais)

D

dang$_{GUB}$ randang [randaŋg] *n* big oil palm; grand rônier à huile
dem$_{GUB}$ budem [bodɛm] *n* palm wine vessel; récipient de vin de palme

F

feet$_{GUB}$ kafeet [kafɛ:t] *v* clear palm tree; nettoyer le rônier
firol$_{KUJ}$ efirol [ɛfirol] *n* fan, palm flower; fleure du rônier

G

góób$_{GUB}$ bugóób [bugo:b] *n* palm grove; palmeraie
góób$_{GUB}$ jëgóób [jəgo:b] *v* harvest; récolter
góób$_{GUB}$ góóbulum [go:bulum] *n* instrument for wine harvest; instrument pour récolter du vin
góóm$_{GUB}$ [bugo:m] *n* central branch of palm tree; rameau central de rônier

H

hobot$_{KUJ}$ kahobot [kahɔbɔt] *n* palm wine spoon; cuillère de vin du palme

I

it$_{BAN}$ ñiit [ɲi:t] *n* palm tree; rônier *Borassus aethiopum*
it$_{KUJ}$ jiit [ji:t] *n* palm tree; rônier *Borassus aethiopum*
it$_{KUJ}$ këit [kəit] *n* palm leaf; feuille de rônier
it$_{KUJ}$ siit [si:t] *n* palm fruit; fruit de rônier

J

jal$_{KUJ}$ kajalen [kajalɛn] *v* prune palm tree; tailler le rônier
jóñ$_{GUB}$ sijoñ [sijɔɲ] *n* type palm tree; espèce de rônier
jund$_{KUJ}$ ejund [ɛjund] *n* palm wine vessel; récipient de vin de palme

K

kaŋ$_{GUB}$ sinkaŋ [siŋkaŋ] *n* fibre of palm leaf; partie de feuille de rônier
kóoni$_{WOL}$ kóoni [ko:ni] palm tree fruit; fruit du rônier
kunno$_{GUB}$ kunno [kʊn:ɔ] *n* palm wine; vin de palme
kuup$_{GUB}$ bukuup [bokɔ:p] *v* cut palm fronds; couper feuilles de ronier

L

lëër$_{GUB}$ guleer [gulə:r] *n* stem of palm leaf; tige de feuille de rônier
let$_{GUB}$ gulet [gulɛt] *n* nail for palm wine harvest; clou pour récolter du vin

M

miita$_{KUJ}$ miita [mi:ta] *n* palm oil; huile de palme
miita$_{GUB}$ miita [mi:ta] *n* palm oil; huile de palme

N

ñaak$_{GUB}$ gunaak [gunaːk] 1) *n* plug for palm wine harvest; bouchon 2) *der.v.* weave funnel; tisser entonnoir
niip$_{GUB}$ buniip [bunɪ:p] *n* male flowers of palm tree; fleures mâles du rônier
nipat$_{KUJ}$ funipat [funiɸat] *n* palm flower; fleures du rônier
no$_{GUB}$ kunoo [kʊnɔ:] *n* palm wine; vin de palme
nooh$_{GUB}$ bunooh [bʊnɔ:x] *n* palm wine shack; cabaret
nuh$_{BAN}$ bunuh [bonox] *n* palm wine; vin de palme
nuh$_{KUJ}$ bunuh [bunux] *n* palm wine; vin de palme
nuk$_{GUB}$ bunuk [bonok] *v* collect palm wine; descendre le vin de palme
ñuxaat$_{GUB}$ tiñuxaat [tiɲoxaːt] *n* palm juice; jus de palme

R

reeja$_{GUB}$ gureeja [gurɛ:ja] *n* part of palm tree (petiole?); partie de rônier
risend$_{GUB}$ gurisend [gurisɛnd] *n* palm kernel; noix de palme
ron$_{WOL}$ ron bi [rɔn bi] *n* palm tree; rônier *Borassus aethiopum*
rukand$_{KUJ}$ furukand [furukand] *n* palm rat; rat palmé *Rattus palmarum*

S

sau$_{GUB}$ kosaw [kɔsaw] *n* little oil palm; petit palme à huile

T

tijit$_{KUJ}$ futijit [futijit] *n* bunch of palm fruit
tokond$_{KUJ}$ katokond [katɔkɔnd] *n* palm wine spoon; cuillère de vin du palme
tos$_{GUB}$ butos [bʊtɔs] *n* palm kernel sauce; sauce de palmiste

U

uc$_{GUB}$ ruuc [ruʊc] *n* palm tree; rônier *Borassus aethiopum*

V

va$_{BAN}$ gava [gava], *v* palm wine tapping; récolter le vin de palme
vér$_{KUJ}$ evér [ɛvɛr] *n* palm nut; noix de palme

W

wa$_{KUJ}$ kawa [kawa] *v* harvest palm wine; récolter du vin de palme
wa$_{KUJ}$ awaa [awaa] *n* palm wine harvester; récolteur du vin de palme