Methodological and technical challenges of a corpus-based study of Naija

Bernard Caron

Bernard Caron
ORCID 0000-0002-6955-8227
CNRS-LLACAN

Methodological and technical challenges of a corpus-based study of Naija

Abstract

This paper presents early reflections on the NaijaSynCor survey (NSC) financed by the French *Agence Nationale de la Recherche*. The nature of the language surveyed (Naija, a post-creole spoken in Nigeria as a second language by close to 100 million speakers) has induced a specific choice of theoretical framework (variationist sociolinguistics) and methodology (a corpus-based study using Natural Language Processing). Half-way through the 4 year-study, the initial methodological choices are assessed taking into account the nature of the data that has been collected, and the problems that occurred as early as the initial stages of their annotation.

Keywords: Atlantic pidgins and creoles, corpus studies, natural language processing, syntax, prosody

1. Introduction

The NaijaSynCor survey (NSC[1]) is a corpus-based survey of Naija, a pidgincreole (Bakker 2008) spoken in Nigeria as a second language by close to 100 million speakers in Nigeria and in the Nigerian diaspora. The nature and size of the language has compelled us to make the annotation process as automatic as possible, with the help of multiple programmes: PRAAT (Boersma & Weenink 2013) for alignment, Elan-Corpa (Chanard 2014) for transcription-translation and semi-automatic tagging; SPPAS (Bigi & Hirst 2012; Bigi et al. 2017) for phonetisation and syllabification; Analor (Avanzi et al. 2008) for prosodic annotation; Arborator[2] for dependency syntactic annotation; Trameur (Fleury & Zimina 2014) and Grew (Guillaume et al. 2012) for error mining, information retrieval and analysis. The metadata was processed through an application based on Arbil[3], and developed by Christian Chanard (Llacan) to make it more user-friendly.

The parallel use of so many different applications requires precise coordination and constant review to adapt the procedures and ensure a smooth workflow. One of the main

[1] A Corpus-based Macro-Syntactic Study of Naija (Nigerian Pidgin) – NaijaSynCor. *Agence Nationale de la Recherche*. February 2017-July 2020. https://anr.fr/Project-ANR-16-CE27-0007. Principal Investigator: Bernard Caron, CNRS-LLACAN.

[2] https://arborator.ilpga.fr/

[3] http://explorationdecorpus.corpusecrits.huma-num.fr/arbil/

challenges is to make sure that the temporal indexes of the annotations are preserved by the various programmes so that Prosodic, Communicative and Syntactic hierarchies can be projected on each other. Another challenge is linked to the nature of the linguistic object itself: as a rapidly expanding pidgincreole, it is somehow unstable, and we have to deal with innovations for which the annotation system must be revised and revisable without having to redo the annotation and without loss of information. A certain degree of lability must be built into the methodology to allow for this inherent dimension of the research project. Two examples of change of procedure will be given concerning phonetization and syllabification on the one hand, and dependency syntax on the other hand.

This paper concentrates on the methodology, the edition and the annotation of the corpus.

2. Naija and Nigerian Pidgin

Nigeria, with 160 million inhabitants, is a huge and complex multilingual community with over 500 different languages (Lewis et al. 2013) used within the public and private social spaces. Among those, **Nigerian Pidgin**, is spoken as a first language by 5 million people, while over 70 million people use it as a second language or as an interethnic means of communication in Nigeria and in Nigerian Diaspora communities. Since the independence of Nigeria in 1960, this variety of Nigerian Pidgin has been rapidly expanding from its original niche in the Niger delta area to cover two-thirds of the country, up to Kaduna and Jos, and is now deeply rooted in the vast Lagos conurbation of over 20 million people. Apart from its original location and one Lagos district, where it is learnt as a first language and can be used as a single language (Elugbe & Omamor 1991), this emerging variety is learnt alongside and not instead of other Nigerian languages. It has become, over the last 30 years, the most important, most widely spread, and perhaps the most ethnically neutral lingua franca used in the country today.

The origin of Nigerian Pidgin itself (NP) is generally described as a development out of an English-lexified jargon attested in the 18th century in the coastal area of the Niger delta (River State), with lexical and structural influence from Krio through the activities of missionaries from Sierra Leone (Faraclas 1996; Huber 1999). Today, the heartland of NP is the Niger Delta, with Lagos and Calabar as secondary extensions. But a new development has taken place over the last 50 years whereby NP has escaped from its original geographical niche, where it functioned as an auxiliary medium of communication in restricted informal contexts by uneducated people (Deuber 2005), and is now commonly used all over Nigeria by the educated in informal conversations, and in formal domains, *viz* radio, television, politics, advertising, Christian religious activities, etc.

This variety of Nigerian Pidgin is gradually becoming a pidgincreole which we call **Naija**[4] to distinguish it from Nigerian Pidgin, the creole spoken in e.g. Warri, Sapele and the Ajegunle district of Lagos. This paper and the NaijaSynCor project are dealing specifically with Naija.

[4] *Naija*, based on the etymon *niger* which gave its name to the river, is the term used by NP speakers to refer to *Nigeria*.

In terms of functional status, English is Nigeria's official language, and it is dominant in the education system and in written usages (literature, press, etc.). However, Naija has made considerable progress in formal contexts such as information transmission by government and non-government agencies, Christian religious practices, and although it is still excluded from the educational system, it is used unofficially in multilingual schools in southern Nigeria. Naija is a lingua franca in public informal communication in the south and to a certain extent in the north, and it is noticeably popular among university students and among educated speakers in private informal communication (Egbokhare 2004). Recently, Naija has become ubiquitous on local FM radio stations, and has become the single medium of the Wazobia radio and TV and of the Pidgin BBC station since its launching in August 2017. Last but not least, its use is an identifying feature of Nollywood, the prosperous Nigerian film industry now known all over the world.

At the same time as it grows in terms of status and functions, Naija expands geographically, and it is exposed to vernacular languages belonging to different genetic and typological groups (such as Yoruba in the southwest; Igbo, etc. in the southeast; Hausa further north). In the process, does it undergo some degree of contact-induced variation beyond the odd word borrowed from those vernacular languages, or on the contrary, does one standard variety emerge through the influence of modern mass-media such as radio, television and video?

In its functional expansion, Naija is subject to extensive contact and influence from its original lexifier, i.e. English, which is the dominant formal and official language in Nigeria. A question arises as to the extent of this influence today, and what can be deduced of the future of Naija. Does Naija, despite the influence of English (and the indigenous languages of Nigeria) maintain its existence as a discrete language (Deuber 2005) or is it undergoing "decreolization", resulting in what has been described as a post-creole (P/C) continuum (Rickford 1987)? In such a process, a whole range of "mesolectal" varieties create a continuum between the "basilect" (*viz*., in our case, Delta NP) and the acrolectal varieties deeply influenced by the original lexifier and its local variant (*viz*. the Nigerian variety of English). Is Agheyisi (1984) right when she states that "the possibility of a systematic mesolectal variety emerging in the Nigerian situation is rather remote" (p. 230)? Deuber (2005) convincingly argues that the Naija variety spoken by educated speakers in Lagos is a discrete language, distinct and separate from English, and "the more competent a speaker is in both languages, the better he/she is able to keep them apart" (p. 203). However, the question remains whether this situation applies to Naija outside Lagos where it is further influenced by local native languages (e.g. Yoruba, Igbo, Hausa).

The influence of written Nigerian English on Naija needs special consideration. The extension of Naija to formal usages such as the radio news report, political and information podcast blogging, Bible translation, short story writing, exposes the language to the influence of written Nigerian English. News reports on the radio are generally translated from press releases issued in English by news agencies. Podcast blogs are read from written texts. This new dimension is bound to influence the structure of the language. From the structure of oral Naija, where utterances and information units are mainly structured by information structure, the structure of sentences in written Naija tends to be informed by microsyntax.

3. Objectives and hypotheses

The general aim of the NSC project is to take an exhaustive and in-depth look at the **nature and functions of Naija (Nigerian Pidgin) in Nigeria today**, in order to establish the link between change in structure and change in language use and function. It makes use of the most advanced developments in corpus studies and natural language processing, which combines with a sociolinguistic and geographical study of variation according to formal/informal uses, gender and education of speakers. The corpus studies natural (non-elicited) speech in order to evaluate the distance between Naija and Nigerian English through the study of intonation, information structure, morphology, micro- and macro-syntax.

The distinction between micro- and macrosyntax was first proposed by Blanche-Benveniste et al. (1990), Berrendonner (1990), and Cresti (2000) (but see also Andersen & Nølke (2002) for an overview). These studies put forward macrosyntax as a level of linguistic description capable of accounting for a number of cohesion mechanisms particularly frequent in spontaneous spoken language, which cannot be simply regarded as microsyntactic government phenomena, such as, for example, the "paratactic" constructions in (1) where no conjunction expresses the syntactic link between *you carry your children go* and *you go still buy food*:

1. [*you carry your children go*] [*you go still buy food*] (Deuber 2005)
   [you bring your children] [you will still buy food]
   '[Even if] you bring your children, you will still have to buy food.'

Macrosyntactic models characterize some major linguistic units that go beyond government proper and are usually described in the literature from a pragmatic perspective that focuses on their illocutionary or rhetorical values. Macrosyntax, instead, focuses on the span and the form of macrosyntactic units, using syntactic and distributional criteria (such as suppressions, insertions, commutations) to identify and delimit them. For all macrosyntactic models, the main identifying criterion of a macrosyntactic unit is the possibility that this unit has to constitute an autonomous utterance.

The problems facing any programme of an exhaustive and in-depth study of Naija are many. The first one is related to the popular view of Naija as a protean, ever changing, informal medium that has no unity, and varies with every place and situation where it is spoken. The NSC project is based on the assumption that it is a discrete language with a strong unity that accommodates a certain range of variation.

The second one is related to the success of the Bickerton-DeCamp theory of the creolization-decreolization cycle (cf. above) informing the work of researchers such as Faraclas (1996), Elugbe and Omamor (1991) who approach the study of Naija with a purist attitude, for whom the only form of NP worth studying is the Warri-Sapele "creole" variety spoken in Delta State, and who consider other varieties at best as degraded forms working as a lingua franca commonly called Broken, at worst as "pseudo-pidgins" invading the press and media. Their descriptions of Naija are monolectal, based on their intuition as speakers of the language. The NSC project is data driven and multilectal.

The third one is the difficulty of combining a structural approach with a sociological one. The structural approach, best illustrated by Faraclas (1996) concentrates on the

grammar and vocabulary of the language as revealing the inner mechanisms responsible for the birth and evolution of creoles, pidgins, and languages in general. The sociological approach mostly favoured by Nigerian scholars centres on the study of language usage and representation among speakers (e.g., Ajibade et al. 2012), but the link with the nature and structure of the language is often absent. The NSC project bridges the gap between mental representations of Naija and actual linguistic usages. It plans to use the most recent trends of data-driven, corpus-based sociolinguistics, combining qualitative and quantitative corpus methods with functional and structural analyses.

The fourth one is the difficulty attached to the mere size of the language, and the challenge it represents for a study to account for the geographical and functional variations of a language with millions of speakers. This calls for careful corpus planning and well organised team work, but most of all, it requires the use of NLP tools to make corpus annotation as automatic as possible.

To sum up, the objectives of the NSC project are:

1. Building a reference 500,000 words **oral corpus** (the **Reference Naija Corpus**, **RNC**), collected in 10 different points of survey in the country, with a deeply annotated sub-section of 100,000 words (the **Naija prosodic and syntactic Treebank**, **NTB**). Annotated corpora are rare for most of the languages of the world, all the more so if one considers depth of annotation (part-of-speech tagging, syntactic parsing, prosodic annotation). This synchronic picture of Naija, documenting its geographic and demographic variation, is a rare opportunity to study the evolution of a fast emerging, vast new language spoken by tens of millions. This corpus is expected to provide the basis for the standardisation and development of the language.
2. Comparing the RNC with the Nigerian International Corpus of English, ICE Nigeria (Gut 2014), both qualitatively and quantitatively. Naija has been proved to be, in the use of the educated Nigerians living in Lagos, a discrete language that is developing and keeping its own distinctive identity and status separate from English (Deuber, 2005). This study aims to assess whether this holds true in the other parts of Nigeria where it is spoken. This comparison aims at evaluating the discreteness and independence of Naija in relation to Nigerian English, and tests the correlation of potential variations to sociological/functional factors.
3. Achieving a better understanding of the variations of Naija along the formal-informal functional scale through the study of its use on university campuses and in the media, and more specifically on the radio (news reporting, editorials, information, etc.). The following hypotheses will be tested: (i) educated Naija is more standardized possibly due to the geographical and social mobility of its speakers; (ii) it reveals a greater influence from English with more borrowing and more syntactic restructuring; (iii) scripted oral Naija reveals an even stronger influence from English than unscripted oral Naija. The results of the project are expected to provide the basis for the standardisation and development of the language. This assessment of the role and impact of new media in relation with the change of attitude of speakers concerning an emerging language is an unprecedented endeavour. The framework chosen is that of the variationist sociolinguistics framework (Tagliamonte 2012).
4. Understanding the patterns observed in the prosody of emerging languages, and linking the prosodic description of Naija to that of its grammatical and information

structures through the use of NLP tools. The aim is threefold: (i) produce a prosodic description of an underdescribed language in Africa based on instrumental analyses and validated with a resynthesis tools; (ii) provide the Naija Treebank with an in-depth integrated annotation for part-of-speech (POS), intonation, micro- and macro-syntactic structures and information structure, thus producing a gold-standard benchmarking large treebank database, a first for an emerging language; (iii) developing Natural Language Processing (NLP) tools for Naija, namely a POS tagger, an English glosser, and a syntactic parser, integrating a treatment of macrosyntactic constructions (dislocation, clefting …) and phenomena specific to spoken languages (disfluency, reformulation, discourse markers). The integration of macrosyntax into a syntactic parser is a ground-breaking endeavour, where high-gain results are expected for the development of NLP tools.

4. Methodology

This section retraces the workflow of the NSC project from data collection to editing, annotation and sociolinguistic analysis (Figure 1).

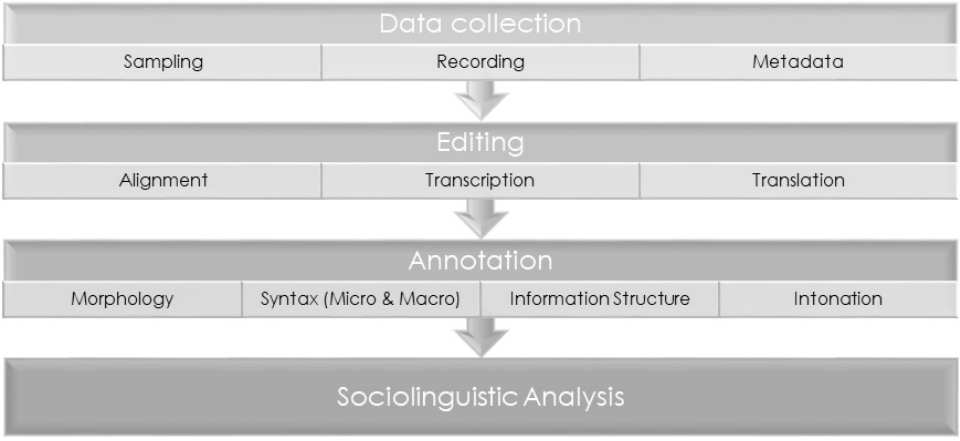


Figure 1. NSC Workflow

4.1. Sampling and data collection

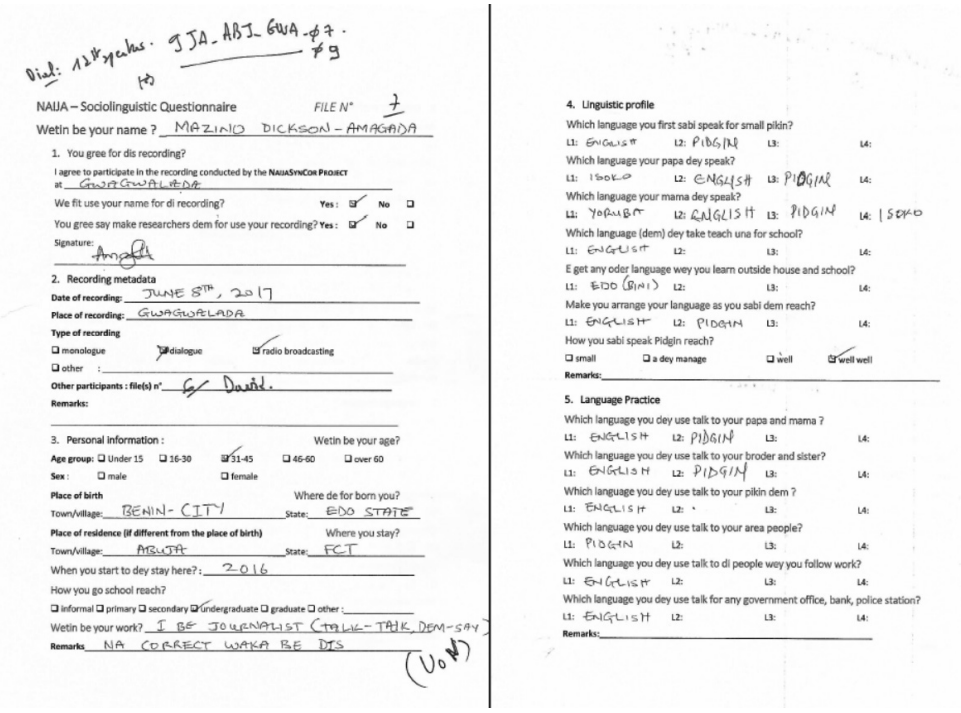
In the absence of sufficient written data, it has been decided to test our hypotheses on an oral corpus to be compared with existing corpora of English, both British and Nigerian, and the Deuber corpora (Deuber 2005) recorded in Lagos almost 20 years ago. The size and nature of the object we want to study imply numerous constraints in collecting the oral data, its annotation, and its analysis.

Due to logistic constraints, our survey focused on urban areas, and more specifically on 10 locations: Lagos, Ibadan, Benin City, Enugu, Onitsha, Port Harcourt, Abuja, Kaduna, Jos, Kano[5].

[5] Calabar, which was originally included in the sampling, was left out for logistic reasons, while Onitsha and Enugu could be included in the survey.

Anyone who volunteered and felt confident in their competence in Naija was recorded speaking in monologues and dialogues on any topic they chose themselves. Out of more than 330 speakers recorded, only 2 requested to remain anonymous. Some life stories were collected, as well as 2 speeches on road safety by government agents and 2 sermons from pastors. Radio call-in programmes, news readings and commentaries were recorded in radio stations in Lagos, Ibadan, Kano, Kaduna, Abuja and Port Harcourt. Two excerpts from a drama representing the Passion of Christ staged in Ibadan for Easter 2017 were read by members of the troupe.

The variationist analysis we want to perform implies collecting samples representing different types of speakers, and different types of functions, with a metadata questionnaire documenting the time and place of the interview and the linguistic biography of the speakers (Figure 2). The questionnaires were processed through Arbil with an interface developed by Christian Chanard (Figure 3). After normalisation of data and georeferencing, the metadata can be analysed and mapped to visualise the sampling. (Figure 4).


NAIJA – Sociolinguistic Questionnaire FILE N° 7

Wetin be your name ? MAZINO DICKSON – AMAGADA

1. You gree for dis recording?

I agree to participate in the recording conducted by the NAIJASYNCOR PROJECT at GWAGWALADA

We fit use your name for di recording? Yes : ☑ No ☐

You gree say make researchers dem for use your recording? Yes : ☑ No ☐

Signature:

2. Recording metadata

Date of recording: JUNE 8TH, 2017

Place of recording: GWAGWALADA

Type of recording

☐ monologue ☑ dialogue ☑ radio broadcasting

☐ other :

Other participants : file(s) n° 6/ David.

Remarks:

3. Personal information : Wetin be your age?

Age group: ☐ Under 15 ☐ 16-30 ☑ 31-45 ☐ 46-60 ☐ over 60

Sex : ☐ male ☐ female

Place of birth Where de for born you?

Town/village: BENIN-CITY State: EDO STATE

Place of residence (if different from the place of birth) Where you stay?

Town/village: ABUJA State: FCT

When you start to dey stay here? : 2016

How you go school reach?

☐ informal ☐ primary ☐ secondary ☑ undergraduate ☐ graduate ☐ other :

Wetin be your work? I BE JOURNALIST (TALK-TALK DEM-SAY)

Remarks NA CORRECT WAKA BE DIS (VOA)

4. Linguistic profile

Which language you first sabi speak for small pikin?

L1: ENGLISH L2: PIDGIN L3: L4:

Which language your papa dey speak?

L1: IGBO L2: ENGLISH L3: PIDGIN L4:

Which language your mama dey speak?

L1: YORUBA L2: ENGLISH L3: PIDGIN L4: IGBO

Which language (dem) dey take teach una for school?

L1: ENGLISH L2: L3: L4:

E get any oder language wey you learn outside house and school?

L1: EDO (BINI) L2: L3: L4:

Make you arrange your language as you sabi dem reach?

L1: ENGLISH L2: PIDGIN L3: L4:

How you sabi speak Pidgin reach?

☐ small ☐ a dey manage ☐ well ☑ well well

Remarks:

5. Language Practice

Which language you dey use talk to your papa and mama ?

L1: ENGLISH L2: PIDGIN L3: L4:

Which language you dey use talk to your broder and sister?

L1: ENGLISH L2: PIDGIN L3: L4:

Which language you dey use talk to your pikin dem ?

L1: ENGLISH L2: L3: L4:

Which language you dey use talk to your area people?

L1: PIDGIN L2: L3: L4:

Which language you dey use talk to di people wey you follow work?

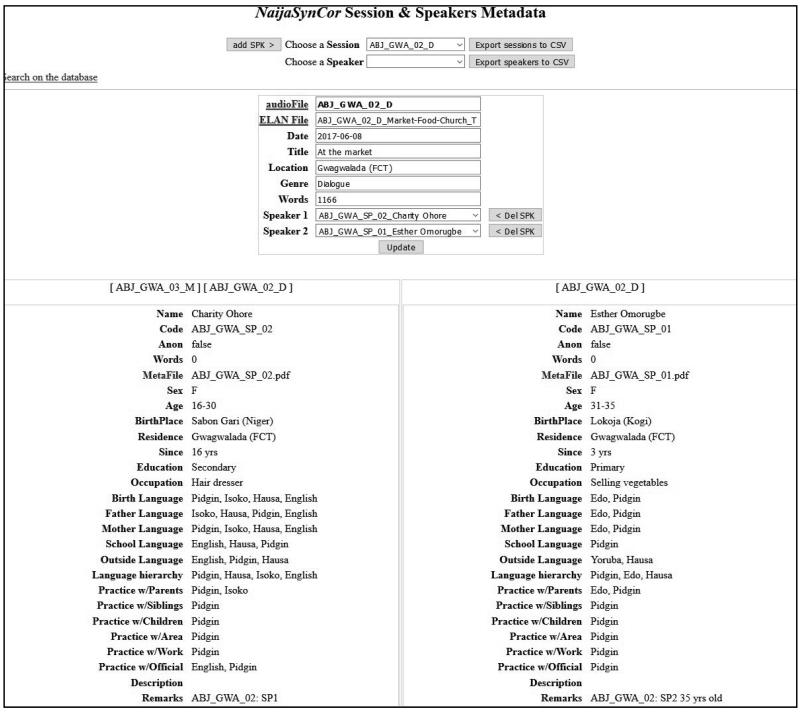
L1: ENGLISH L2: L3: L4:

Which language you dey use talk for any government office, bank, police station?

L1: ENGLISH L2: L3: L4:

Remarks:

Figure 2. Sample of a metadata questionnaire



NaijaSynCor Session & Speakers Metadata

add SPK > Choose a Session ABJ_GWA_02_D Export sessions to CSV
Choose a Speaker Export speakers to CSV

earch on the database

audioFile	ABJ_GWA_02_D
ELAN File	ABJ_GWA_02_D_Market-Food-Church_T
Date	2017-06-08
Title	At the market
Location	Gwagwalada (FCT)
Genre	Dialogue
Words	1166
Speaker 1	ABJ_GWA_SP_02_Charity Ohore < Del SPK
Speaker 2	ABJ_GWA_SP_01_Esther Omorugbe < Del SPK

Update

[ABJ_GWA_03_M] [ABJ_GWA_02_D]		[ABJ_GWA_02_D]	
Name	Charity Ohore	Name	Esther Omorugbe
Code	ABJ_GWA_SP_02	Code	ABJ_GWA_SP_01
Anon	false	Anon	false
Words	0	Words	0
MetaFile	ABJ_GWA_SP_02.pdf	MetaFile	ABJ_GWA_SP_01.pdf
Sex	F	Sex	F
Age	16-30	Age	31-35
BirthPlace	Sabon Gari (Niger)	BirthPlace	Lokoja (Kogi)
Residence	Gwagwalada (FCT)	Residence	Gwagwalada (FCT)
Since	16 yrs	Since	3 yrs
Education	Secondary	Education	Primary
Occupation	Hair dresser	Occupation	Selling vegetables
Birth Language	Pidgin, Isoko, Hausa, English	Birth Language	Edo, Pidgin
Father Language	Isoko, Hausa, Pidgin, English	Father Language	Edo, Pidgin
Mother Language	Pidgin, Isoko, Hausa, English	Mother Language	Edo, Pidgin
School Language	English, Hausa, Pidgin	School Language	Pidgin
Outside Language	English, Pidgin, Hausa	Outside Language	Yoruba, Hausa
Language hierarchy	Pidgin, Hausa, Isoko, English	Language hierarchy	Pidgin, Edo, Hausa
Practice w/Parents	Pidgin, Isoko	Practice w/Parents	Edo, Pidgin
Practice w/Siblings	Pidgin	Practice w/Siblings	Pidgin
Practice w/Children	Pidgin	Practice w/Children	Pidgin
Practice w/Area	Pidgin	Practice w/Area	Pidgin
Practice w/Work	Pidgin	Practice w/Work	Pidgin
Practice w/Official	English, Pidgin	Practice w/Official	Pidgin
Description		Description	
Remarks	ABJ_GWA_02: SP1	Remarks	ABJ_GWA_02: SP2 35 yrs old

Figure 3. A metadata questionnaire processed through Arbil

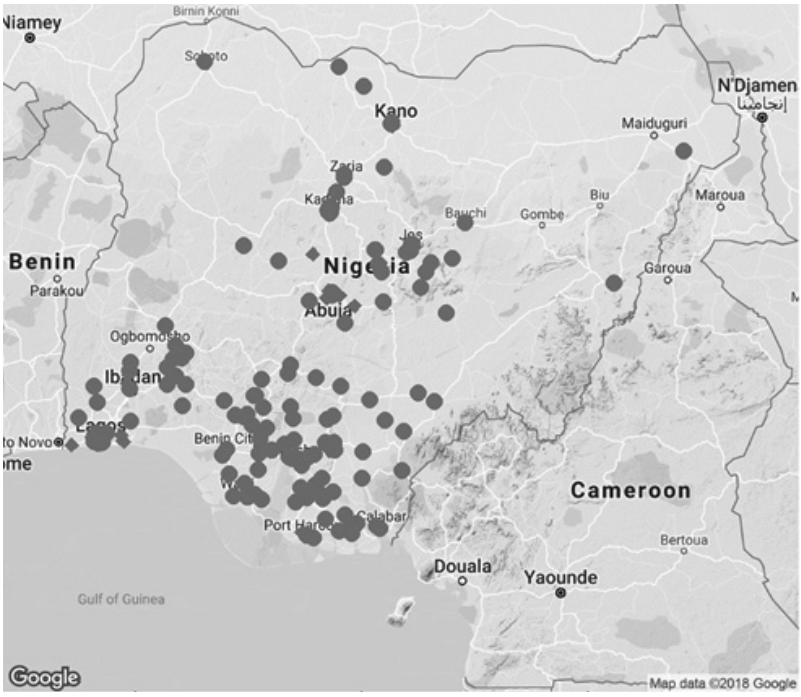


Figure 4. Speakers' birthplaces (Čéplö & Manfredi 2019)

4.2. Annotation, querying and NLP tools

NLP tools are used for corpus editing, morphosyntactic tagging (Elan; §4.3.1); phoneticising (SPPAS) and prosodic modelling (Analor; §4.3.2); syntactic annotation (MATE and Arborator; §4.3.3); error mining, querying and sociolinguistic analysis (Grew and Trameur; §4.4.4).

Figure 5. NLP tools in NSC

4.2.1. Editing the data

Elan (https://tla.mpi.nl/tools/tla-tools/elan/) (Sloetjes 2014) was used in annotate the files by providing time alignment, transcription, tokenization into words, semantic word-level glosses and translation into Standard English. Compatibility of transcription has been ensured by using the orthography developed in (Deuber 2005). This etymological orthography (adapted from the lexifier language orthography, i.e. English) has been chosen by Deuber preferably to the phonological script used by linguists (e.g. Faraclas, Elugbe, etc.) as it is spontaneously used by educated Nigerians, and thus easier to teach to transcribers. Codeswitched sections were identified by dedicated boundaries. Transcriptions and translations are double checked for the sake of consistency. A macrosyntactic punctuation marks macro-syntactic boundaries (i.e. illocutionary units and their main components: nucleus, prenuclei and post nuclei, including discourse markers) and limits between pile layers (disfluencies, reformulation, coordination). All these boundaries are marked by punctuations in written texts.

Our segmentation is based on a long tradition of the study of syntax of spoken production in Romance languages (Blanche-Benveniste et al. 1990; Cresti 2000; Simon & Degand 2011). Our maximal syntactic units are illocutionary units, that is, assertions, questions, and demands. We use the markup developed in the Rhapsodie project of annotation of spoken French (Deulofeu et al. 2010; Kahane & Pietrandrea 2012), which is a kind of formalized punctuation. The delimiter for illocutionary units is //. Consider this extract (2) from a sample illustrating the markup:

2. *den you go dey wrap dat food { small |r small } // cut cocoyam //= cut dat uh & // take { cocoyam |c and yam } wey you don grind //= […]*
   'then you will wrap that food in small pieces, cut the cocoyam, cut that er… take the cocoyam and yam which you have ground […]' [DEU_A05]

The notation { X | Y } indicates that the phrase Y occupies the same syntactic position as X and piles up on X (Gerdes & Kahane, 2009). Four types of lists are considered: "|c" marks coordination (*cocoyam* |c *and yam*); "|r" marks (syntactic) reduplication (*small* |r *small* 'very small'); "|a" marks appositions (*John* |*a my friend*); and "||" marks disfluencies and reformulation (*some* || *some people dey ask* 'some… some people are asking').

Inserting the macrosyntactic annotation into the text is part of the segmentation of the transcription and constitutes a first coarse-grained syntactic analysis. The macrosyntactic annotation can be studied as such to quantify phenomena that are more typical for spoken language such as left and right dislocations and disfluencies. It is also geared for the direct study of the prosody-syntax interface. The macrosyntactic annotation improves parsing results and it can easily be simplified into a standard punctuation.

4.2.2. Prosodic analysis

The main objective of this part of NSC is to include a prosodic level in the description of Naija. It will produce (i) an analysis of its prosodic units and their nature, with a description of their precise acoustic correlates, based on an instrumental analysis and validated with a speech synthesis tool; (ii) a version of the 100Kw Golden Corpus annotated for prosody, adapting schemes developed in the treebank Rhapsodie for French, based on perceptual and acoustic cues, developed independently from the micro/macro-syntactic parsing and labelling, which will serve in the final functional analyses. It will answer questions pertaining to the prosodic system of Naija, including phenomena such as speech rhythm, tonal structure, intonation and stress, e.g. (i) Is Naija a 'tone language', 'pitch-accent language' or 'stress language' (Hyman 2006)? (ii) What is the interplay between putative tone and intonation?

Methodology: from the point of view of phonetic instrumentation and tools, the corpus has been aligned and segmented into phonemes and syllables using SPPAS (Bigi & Hirst 2012). A total of 20 files were manually annotated in prominences and semi-automatically segmented into major prosodic units using the Analor software (http://www.lattice.cnrs.fr/ressources/logiciels/analor/) (Avanzi et al. 2008). In addition, SLAM+, a tool that generates intonative contours automatically (https://github.com/vieenrose/SLAMplus) was developed to process intonative contours, particularly syntactic units, on a large scale (Liu et al. 2019). There arose here an interoperability problem and a tool was developed as part of the project, to retrieve under PRAAT the data encoded in formats dedicated exclusively to syntax processing (Arborator and CONLL). Concerning the functional analysis, a first set of data was used to make hypotheses on the phonetic marking of the focus in Naija (Simard et al. 2019). Another has been processed since July 2019 to study the intonative contours of pre-kernels and macrosyntactic nuclei. This new step should allow us to answer a set of questions related to the encoding of the informational structure of the message in Naija, including: in the initial position of statement, are there specific intonative markers of the pre-kernels and are the observed intonative variations correlated to the informational status of the element included in the pre-kernel (topic

vs. frame; topic active, vs. accessible, introduced, reactivated, etc.) These studies on the intonational markers of framing, topicalization and focusing operations, are still ongoing.

4.2.3. Morphosyntactic analysis

The main objective is to tag, gloss, and parse the 500 Kw corpus, using state of the art NLP tools. In the process, a 100 Kw gold-standard, manually corrected treebank (NTB), has been produced. A new parser, using word embedding and neuronal technology will soon be trained on these files and used to annotate the remaining part of the corpus (Reference Naija Corpus and Deuber Corpus). The evaluation of the automatically annotated corpus (NRC) will start early 2020. Finally, the remaining 400 Kw data will be analysed by the parser trained on the gold-standard treebank, providing the 500 Kw treebank to be compared with ICE-Nigeria and more generally available for use by the sociolinguistic work package.

Glossing and POS tagging. To start the annotation process, a first sample text was tagged with a model trained on English. Insofar as most of the lexicon of Naija is borrowed from English, and its meaning is transparent, the glossing was kept to a minimum. Function words do not have glosses beyond their morphological features, and only Naija lexical innovations were glossed (e.g. *pikin* 'child', *patapata* 'full'). The POS annotation was manually corrected and a first dictionary of the function words and most common lexical items of Naija was created, containing the form, some orthographic variants, the POS tag, and an English gloss if necessary. This dictionary was then used on a dozen text samples inside the Elan-Corpa tool (Chanard 2014), an extended version of the Elan tool3 (Sloetjes 2014), which proposes the dictionary's POS for each token for validation by the annotator. Through this semi-automatic process, the dictionary was enriched and later on used by the automatic tagger that was developed for the project. The POS tags follow the UD conventions (Nivre et al. 2016) with the caveat that some changes were made to accommodate the specificities of the Naija system. For example, Naija has three copulas, *be*, *dey* and *na*, among which two are tagged as VERB (and 'be') and one, also used as a focus particle, is tagged as PART ('it is'). Regularly, the POS tagger is trained again on the corrected tags and thus improved in a bootstrapping loop.

Annotation guidelines. The annotation process for the samples was organized collectively, where each file was assigned to one of the three annotators. They were allowed to discuss the difficult cases among each other. At the end of this process, the annotation was consolidated through the use of a dictionary that was controlled independently and applied to the corpus. The final adjudication was done by an expert adjudicator on every single file. In this process some amendments had to be discussed more widely in the UD community. The annotators are asked to verify their annotations by means of an annotation guide and to report directly into the guide any decision that is not directly derived from it. We thus have an annotation guide that undergoes constant refinement.

From UD to SUD. A preliminary result of the syntactic annotation was published on the UD project website as a mini-corpus pilot of 4 files with a grammatical sketch[6]. After this publication, we made a paradigm shift in the annotation of the corpus. It was decided to develop an enriched syntax annotation for Naija (Syntactic Universal Dependency, SUD) (Gerdes et al. 2018) compatible with Universal Dependency's "classical" model.

[6] http://universaldependencies.org/treebanks/pcm_nsc/index.html.

5. Half-way assessment

5.1. Planned results

The NSC project plans to produce a 100 Kw gold standard treebank for Naija (manual correction); a 400 Kw treebank for Naija (automatic annotation); syntactic annotation guidelines for Naija; a tagger and a glosser for Naija; a dependency parser for Naija (MATE trained on our gold standard treebank); a 500 Kw treebank for Nigerian English (ICE Nigeria analysed with the English Stanford parser). The whole process is expected to deliver the annotated corpora for analysis in early 2020.

5.2. Corpus editing and annotation

The early stages of the NSC corpus construction (fieldwork, alignment, transcription, translation) went quite smoothly and ran ahead of schedule. The corpus construction was completed in December 2017 in Nigeria by the annotation team at the University of Ibadan, under the supervision of the Principal Investigator. In the 10 survey points, more than 380 files were initially edited (representing 31 hours of recording) with corresponding metadata (approx. 350 speakers). After reviewing the audio quality and contents, the corpus was sized down to 321 files and 343 speakers, constituting the NSC reference corpus. By mid-2019, the size of our corpus is as follows (Table 1):

Golden Treebank	Reference Corpus	Deuber (2005)
80 files	241 files	50 files
96 kw	302 kw	100 kw
7 hours	24 hours	10 hours

Table 1. NSC Corpus

Inconsistencies in the morphosyntactic annotation forced us to run systematic tests and extra manual corrections on the Golden Corpus, which have entailed some delays. Then, in order to improve the quality of the automatic parsing, an extra revision of the macro punctuation and sentence alignment was done for the remaining 241 files of the Reference Corpus.

5.3. General problems faced by the project

5.3.1. Sampling

If the geographical sampling is acceptable, the result is not balanced in terms of sex (women represent only 1/3 of the sample; cf. Figure 6), age (Figure 7) and education (Figure 8).

Very few young and elderly speakers were recorded and more than 50% of the speakers were highly educated: more than 50% were graduates and more than 20% were higher education students.

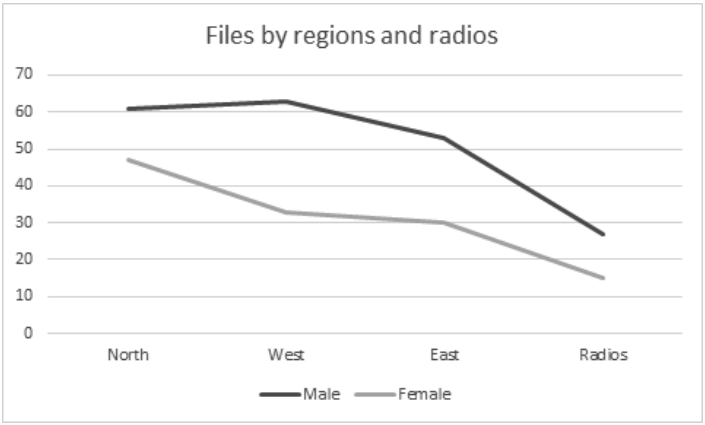


Figure 6

Figure 7

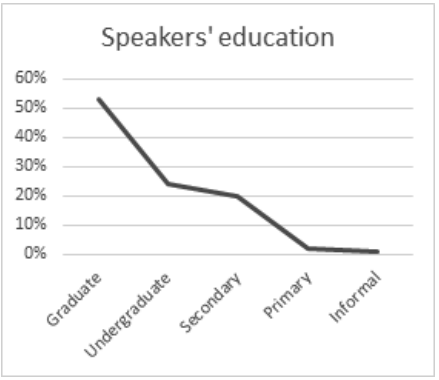


Figure 8

This paints a picture of Naija as a language proudly spoken by highly educated people, who codeswitch confidently from Naija to English, far from the usually prevalent picture of a limited code (a pidgin) used by illiterates who are linguistically impaired.

As a conclusion, we can state that the corpus is expected to give a good representation of the language; not an exhaustive representation of its status or its speakers. However, it contradicts the widely spread prejudice that Naija is a language associated with illiterates, and a danger for Nigerian education and English in Nigeria.

5.3.2. Transcription

Orthography: due to unstable orthography, annotators have not been consistent. A lemma was selected in the tagging process, and variants were kept in the transcription, e.g. 'thing': *thing*, *ting*, *tin* (lemma: *ting*); 'their': *their*, *deir*, *dier* (lemma : *deir*); 'there': *there*, *dere*, *dier* (lemma: *dere*); 'him': *him*, *im*, *in* (lemma : *im*)

Phonetics: The transcription of phonetic variants has not been consistent, which means that this aspect of variation cannot be studied in the corpus, e.g. 'make': [*mek*, *me*, *ma*, *mo*]; 'him': [him, *im*, *ĩ*]; 'them': [*đem*, *dem*, *de*].

Morphology: the speakers' habits, some orthographic decisions have been taken to eliminate lexical ambiguities that have strong morphological and syntactic implications. When annotators did not respect these decisions, the variants were kept in the text and the NSC orthography indicated as a lemma. This is the case with morphemes derived from serial verb constructions, in which the second verb has been grammaticalised, e.g. the complementizer which we have chosen to write *sey* instead of *say* (as it is usually spelt by Naija speakers) to disambiguate it from the verb *say*. Likewise, we have chosen to write the future auxiliary *con* instead of *come*.

As a first contribution to the establishment of a standard orthography of Naija, it has been decided to publish online a normalised version of the corpus aimed at a wider Nigerian audience.

5.3.3. Morphosyntactic analysis

Decisions had to be taken concerning the tagging of various items, e.g. the copulas *be*, *dey* and *na*, which can be tagged as auxiliaries following the UD guidelines, verbs or particles. Since *na* (a focus marker which can be also used as a predicative copula) cannot be combined with TAM markers or negation, we have analysed it a as particle, contrary to *be* and *dey*, which have been analysed as verbs.

Another problem concerns a category that has been labelled "property items" by Mazzoli (2013) in a bid to avoid calling them either verbs or adjectives. Faraclas clearly argues for the inexistence of adjectives in Nigerian Pidgin: "*there is no category 'adjective' in Nigerian Pidgin. Most of the items which convey the same meanings as do adjectives in other languages are stative verbs in NP. Stative verbs take the same arguments and modifiers in the same combinations and the same order as do other verbs."* (Faraclas 1989: 132)

Indeed, our Naija corpus corroborates Faraclas's analysis of Nigerian Pidgin, in that numerous examples of "property items" are used as stative verbs, both intransitive, as in (3) and (4) and transitive as in (5):

3. *Women sef, we bad*. 'We, women, we [are] bad.' [P_JOS_14]
4. *Di meat sweet o!* 'The meat [is] quite tasty!' [P_JOS_20]
   and transitive, as in (5)
5. *Di weather dey sweet us.* 'We enjoy the weather (lit. the weather [is] nice to us)'. [P_JOS_20]

They combine with TAM auxiliaries, e.g. the Future AUX *go* in (6) and the Imperfective *dey* in (8):

6. *Di fruit go sweet*. 'The fruit will [be] nice.'[P_IBA_31]

They combine with the negative particle *no*, as in (7):

7. *Belle no sweet am at all* 'He is not happy at all. (lit. the stomach does not satisfy him at all.)' [P_WAZK_07]

In (8), the term *sweet* is part of the comparative serial verb construction *sweet … pass*, lit. 'be sweeter':

8. *Na dat one dey sweet me pass.* 'It's that one I like best. (*lit*: It's that one that satisfies me most.)' [P_IBA_02]

This leads Faraclas to analyze nouns modified by property items (e.g. big money, small work, bad name) as relative clause constructions: *"Since the category 'adjective' does not exist in NP […] and because of the fact that the only type of clause in the language which may serve to modify nominal elements is the relative clause, the label 'adjective clause' is not employed here, 'relative clause' being used instead."* (Faraclas 1989: 75)

However, if terms like *sweet*, *smooth*, etc. function mainly as a verbs in our Naija corpus, others like *next*, *last*, *waye* 'dubious', etc. only function as adjectives.

In conclusion, in our Naija corpus, there is no clear case in favor of choosing either adjective (ADJ) or verb (VERB) as a POS tag for those items. For comparative purposes, we have decided to tag as ADJ items that can function both as adjectives and verbs, while we keep the tag VERB for those that can only function as verbs. Likewise, we tag NOUN items that can function both as nouns and verbs.

5.4. Exploratory results of the corpus-based analysis

Several papers have been published exploring some striking properties of Naija, as they already appear in the corpus.

(Čéplö & Manfredi 2019) have made a first attempt at assessing morpho-syntactic variation in Naija. (Simard et al. 2019) make a first presentation of the prosody of the language. They deal with focus and prominence types and more specifically, with the prosodic encoding of narrow focus. They were able to show that narrow focus (*na*+ focused element construction) is conveyed mainly with duration prominence (See fig. X)

9. *De fit say de dey strike.* 'they can declare industrial action' [P_IBA_21]

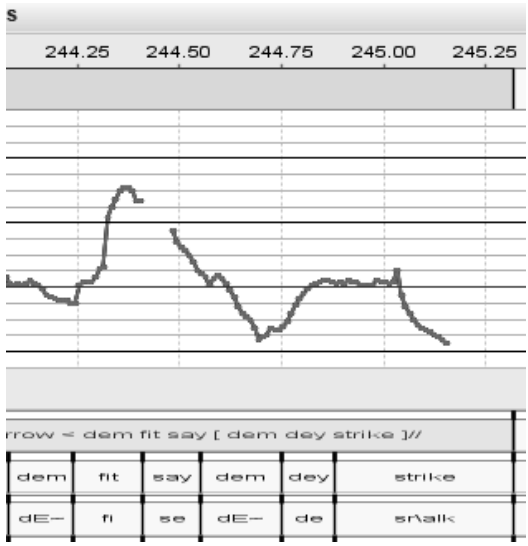


Figure 9. Pitch track of Example 9, illustrating the longer duration on the focused element, strike.

Several papers deal with the syntax of Naija, and, among other things, with serial verb constructions (Caron et al. 2019) and clefts (Caron 2020). The basic element in the structure of clefts in Naija is the focus particle *na*, as in (6):

10. *na nineteen eighty four >+ wey de born me //* 'it's in nineteen eighty-four that I was born.' [P_KAD_09]

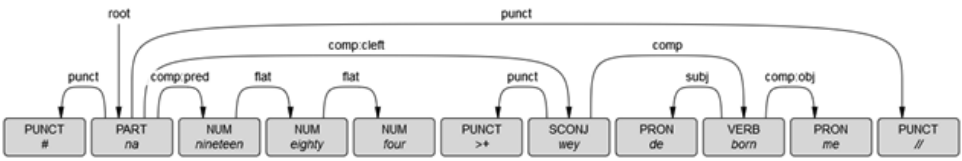


Figure 10. Dependency structure of Example 10.

Clefts in Naija show a great variety of structures and clear innovations from Nigerian Pidgin. Naija clefts have four variants, illustrated in Table 2: *wey*-clefts (a), with a relative clause introduced by the relativizer *wey*; bare clefts (b), where the relativizer is omitted, resulting in a bare relative clause; zero clefts (c) where both the copula and the relativizer are omitted; and double clefts (d), where the relativizer *wey* is replaced by a repetition of the copula *na* followed by an expletive invariable 3sg pronoun: *im*.

a	***wey***-cleft	***na*** *1984* ***wey*** *de born me*	'(It's) in 1984 (that) I was born.'
b	bare cleft	***na*** *1984* ***Ø*** *de born me*	
c	zero cleft	*1984 de born me*	
d	double cleft	***na*** *1984* ***na im*** *de born me*	

Table 2. The four structures of Naija clefts

We have quantified the relative use of these structures in Naija in a sub-section of 9621 sentences (almost 150 000 tokens) that constitute the syntactic treebank mirroring the social and geographic sampling of the full corpus, and compared those figures with Faraclas (2013), a presentation of the structures of Nigerian Pidgin with good data analysis. Using our own terminology, Faraclas's figures highlight 3 main patterns representing fairly evenly cleft constructions in NP: *wey*-clefts (41%); bare clefts (39%) and zero-copula clefts (17%). Our own figures are respectively 1%, 89%, and 1%, with the rest of cleft patterns taken up by double clefts (9%). This shows a tendency in Naija, over the past 30 years, to marginalize *wey*- and zero-copula clefts, in favor of bare clefts, and give birth to a new pattern absent in Faraclas's description, called double cleft, which seems to replace *wey-clefts*. In the double cleft construction, an emerging relative pronoun (*na im* → [nãĩ/nã] 'who, which') which is used only in this construction, replaces the relativizer *wey*, which is becoming specialized in modifying relative clauses.

This shows that Naija is changing fast while the Nigerian Pidgin (NP) described in (Faraclas 2013) has not changed much from the one described in (Faraclas 1989), based on a large corpus gathered by the author in Rivers State thirty years ago. However, there is a stunning difference between NP, a creole spoken in the Niger delta area, and the Naija documented in our corpus, a pidgincreole spoken as L2 by a vast majority of Nigerian.

6. Conclusion

To conclude this half-way assessment of the NaijaSynCor project, it should be noticed that it has already produced a large fine-grained annotated corpus, the first for creoles and pidgins studies[7]. This corpus will serve as a basis for the development of Naija, through the publication of a large corpus in stabilized orthography, a dictionary, and a grammar.

The project has exceeded the time allotted for annotation, which leaves a limited time for analysis. However, we anticipate that the quality of annotation and the powerful tools currently developed or improved by the research team (Trameur, Grew) will help us compensate for the time limitation. 2020 will be devoted to the exploration and evaluation of the corpus, and we intend to finish the NaijaSynCor research project by the end of the year.

Last but not least, the NSC project has developed and improved a series of NLP tools that can be extended and adapted to the study of lesser-described African languages. However, since the efficiency of the new generation of NLP tools relies more and more on access to very large quantities of raw data, the limited resources available for minority languages will impact the quality of the results. The output of such tools will probably be used as annotation propositions that will have to be manually validated by researchers. This nevertheless opens the way to a new methodology for the documentation, description and development of minority languages.

Abbreviations

ADJ	Adjective	POS	Part of Speech
AUX	Auxiliary	TAM	Tense-Aspect-Mood
NP	Nigerian Pidgin	UD	Universal Dependencies
NSC	NaijaSynCor	VERB	Verb

References

Agheyisi, Rebecca Nogieru. 1984. “Linguistic implications of the changing role of Nigerian Pidgin English”. *English World-Wide* 5, 211–233.

Ajibade, Yetunde A., Beatrice Bunmi Adeyemi & Emmanuel Olajide Awopetu. 2012. “Unity in Diversity: The Nigerian Youth, Nigerian Pidgin English and the Nigerian Language Policy”. *Journal of Educational and Social Research* 2(3), 289-295.

Andersen, Hanne Leth & Henning Nølke (eds.). 2002. *Macro-syntaxe et macro-sémantique: actes du colloque international d’Århus, 17-19 mai 2001*. Berne: Peter Lang.

Avanzi, Mathieu, Anne Lacheret & Bernard Victorri. 2008. Analor, un outil d’aide pour la modélisation de l’interface prosodie-grammaire. CERLICO, 27-46. https://halshs.archives-ouvertes.fr/halshs-00636544.

Ayafor, Miriam & Melanie Green. 2017. *Cameroon Pidgin English: a comprehensive grammar*. Amsterdam: John Benjamins.

Bakker, Peter. 2008. “Pidgins versus Creoles and Pidgincreoles”. In: Silvia Kouwenberg & John Victor Singler (eds.). *The handbook of Pidgin and Creole studies* [Blackwell Handbooks in Linguistics]. Chichester, West Sussex/Malden, MA: Wiley-Blackwell Pub, 130–157.

[7] See however a corpus of Cameroon Pidgin (Ozón et al. 2017), which has been automatically tagged for POS and used as a basis for a grammar of the language (Ayafor & Green 2017).

Berrendonner, Alain. 1990. “Pour une macro-syntaxe”. *Travaux Linguistiques de Gand* (21), 25–36.

Bigi, Brigitte, Bernard Caron & Abiola S. Oyelere. 2017. “Developing Resources for Automated Speech Processing of the African Language Naija (Nigerian Pidgin)”. *8th Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*. Poznań, Poland. https://hal.archives-ouvertes.fr/hal-01705707, 441–445.

Bigi, Brigitte & Daniel Hirst. 2012. “SPeech Phonetization Alignment and Syllabification (SPPAS): a tool for the automatic analysis of speech prosody”. *Speech Prosody*, Shanghai (China): Tongji University Press, 19–22.

Blanche-Benveniste, Claire, Mireille Bilger, Christine Rouget, Karel van den Eynde & Piet Mertens. 1990. *Le français parlé: études grammaticales*. Paris: CNRS.

Boersma, Paul & David Weenink. 2013. *PRAAT: Doing phonetics by computer*. http://www.fon.hum.uva.nl/praat/ (30 November, 2013).

Caron, Bernard. 2020. “Clefts in Naija, a Nigerian pidgincreole”. *Linguistic Discovery* 17(1). 149–174.

Caron, Bernard, Marine Courtin, Kim Gerdes & Sylvain Kahane. 2019. “A Surface-Syntactic UD Treebank for Naija”. In: *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*. Paris, France: Association for Computational Linguistics, 13–24. https://doi.org10.18653/v1/W19-7803.

Čéplö, Slavomír & Stefano Manfredi. 2019. Assessing morpho-syntactic variation in Naija (Nigerian Pidgin): a corpus-driven study. Presented at the 2019 SPCL summer meeting, University of Lisbon, June 17-19, 2019, Lisbon, Portugal.

Chanard, Christian. 2014. *ELAN-CorpA-V4.7.3*. http://llacan.vjf.cnrs.fr/res_ELAN-CorpA.php.

Cresti, Emanuela. 2000. *Corpus di italiano parlato*. 2 vols. Firenze, Italie: Accademia della Crusca.

Deuber, Dagmar. 2005. *Nigerian Pidgin in Lagos. Language contact, variation and change in an African urban setting*. London: Battlebridge Publications.

Deulofeu, Jose, Kim Gerdes, Sylvain Kahane & Paola Pietrandrea. 2010. Depends on what the French say: Spoken corpus annotation with and beyond syntactic function, 1–8. https://halshs.archives-ouvertes.fr/halshs-00649791 (1 December, 2015).

Egbokhare, Francis O. 2004. “Language and politics in Nigeria”. In: Kọ́lá Owólabí & A. O Dasylva (eds.). *Forms and functions of English and indigenous languages in Nigeria: a festschrift in honour of Ayọ Banjọ.* Ibadan: Group Publishers, 507–22.

Elugbe, Ben Ohiọmamhẹ & Augusta Phil Omamor. 1991. *Nigerian Pidgin: (background and prospects)*. Ibadan: Heinemann Educational Books Nigeria PLC.

Faraclas, Nicholas. 1996. *Nigerian Pidgin*. London ; New York: Routledge.

Faraclas, Nicholas. 2013. Survey chapter: “Nigerian Pidgin”. In: Susanne Maria Michaelis, Philippe Maurer, Martin Haspelmath & Magnus Huber (eds.). *Atlas of Pidgin and Creole Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://apics-online.info/surveys/17.

Faraclas, Nicholas Gregory. 1989. *A grammar of Nigerian Pidgin*. University of California at Berkeley PhD.

Fleury, Serge & Maria Zimina. 2014. “Trameur: A Framework for Annotated Text Corpora Exploration”. In: Tsujii, Junichi & Jan Hajic (eds.). *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations. August 2014, Dublin, Ireland*. Dublin City University and Association for Computational Linguistics, 57–61.

Gerdes, Kim, Bruno Guillaume, Sylvain Kahane & Guy Perrier. 2018. “SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD”. *Universal Dependencies Workshop 2018*. Brussels, Belgium. https://hal.inria.fr/hal-01930614.

Guillaume, Bruno, Guillaume Bonfante, Paul Masson, Mathieu Morey & Guy Perrier. 2012. “Grew : un outil de réécriture de graphes pour le TAL”. https://hal.inria.fr/hal-00760637 (17 November, 2019).

Gut, Ulrike. 2014. “ICE Nigeria”. *SourceForge*. http://sourceforge.net/projects/ice-nigeria/ (6 August, 2014).

Huber, Magnus. 1999. *Ghanaian Pidgin English in Its West African Context: A Sociohistorical and Structural Analysis*. John Benjamins Publishing.

Kahane, Sylvain & Paola Pietrandrea. 2012. *La typologie des entassements en français*. Vol. 1. SHS Web of Conferences. http://www.shs-conferences.org/.

Lewis, M. Paul, Gary F. Simons & Charles D. Fennig. 2013. Nigeria. *Ethnologue: Languages of the World, Seventeenth edition*. Dallas, Tex.: SIL. International. (22 February, 2014).

Liu, Luigi (Yu-Chen), Lacheret-Dujour & Nicolas Obin. 2019. "Automatic modelling and labelling of speech prosody: what's new with SLAM+ ?". *International Congress of Phonetic Sciences (ICPhS)*. Melbourne, Australia. https://hal.sorbonne-universite.fr/hal-02119926.

Mazzoli, Maria. 2013. *Copulas in Nigerian Pidgin*. Pavia: University of Pavia, PhD dissertation.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, et al. 2016. "Universal Dependencies v1: A Multilingual Treebank Collection". In: Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Paris, France: European Language Resources Association (ELRA).

Ozón, Gabriel, Miriam Ayafor, Melanie Green & Sarah Fitzgerald. 2017. "The spoken corpus of Cameroon Pidgin English". *World Englishes* 36(3), 427–447. doi:10.1111/weng.12280.

Rickford, John R. 1987. *Dimensions of a Creole continuum: history, texts & linguistic analysis of Guyanese Creole*. Stanford, Calif.: Stanford University Press.

Simard, Candide, Anne Lacheret-Dujour & 'Biola S. Oyelere. 2019. "Broad and narrow focus marking in Naija (Nigerian Pidgin): the role of prosody". In: Sasha Calhoun, Marija Tabain Escudero & Paul Warren (eds.). *Proceedings of the 19th International Congress of Phonetic Sciences.* Canberra, Australia: Australasian Speech Science and Technology Association Inc. http://intro2psycholing.net/ICPhS/papers/ICPhS_3956.pdf, 3907–3911.

Simon, Anne Catherine & Liesbeth Degand. 2011. "L'analyse en unités discursives de base : pourquoi et comment ?". *Langue française* 170(2), 45–59.

Sloetjes, Han. 2014. "ELAN: Multimedia Annotation Application". In: Jacques Durand, Ulrike Gut & Gjert Kristoffersen (eds.). *The Oxford handbook of corpus phonology*. Oxford: Oxford University Press 305–320.

Tagliamonte, Sali. 2012. *Variationist sociolinguistics: change, observation, interpretation* [Language in Society 40]. Malden, MA: Wiley-Blackwell.