A Surface-Syntactic UD Treebank for Naija

Bernard Caron, Marine Courtin, Kim Gerdes, Sylvain Kahane

A Surface-Syntactic UD Treebank for Naija

Bernard Caron
IFRA, CNRS
bernard.caron@cnrs.fr

Marine Courtin **Kim Gerdes**
LPP, Sorbonne Nouvelle & CNRS
kim@gerdes.fr,
marine.courtin@sorbonne-nouvelle.fr

Sylvain Kahane
Modyco, Université
Paris Nanterre & CNRS
sylvain@kahane.fr

Abstract

This paper presents a syntactic treebank for spoken Naija, an English pidgincreole, which is rapidly spreading across Nigeria. The syntactic annotation is developed in the Surface-Syntactic Universal Dependency annotation scheme (SUD) (Gerdes et al., 2018) and automatically converted into UD. We present the workflow of the treebank development for this under-resourced language. A crucial step in the syntactic analysis of a spoken language consists in manually adding a markup onto the transcription, indicating the segmentation into major syntactic units and their internal structure. We show that this so-called "macrosyntactic" markup improves parsing results. We also study some iconic syntactic phenomena that clearly distinguish Naija from English.

1 Introduction

Naija is an English pidgincreole (Bakker, 2009) spoken by an estimated 100 million speakers in Nigeria and the Nigerian diaspora in Africa, the UK and the USA. Its origin lies in Nigerian Pidgin, a creole spoken in the Niger delta (Faraclas, 1989; Elugbe and Omamor, 1991). As the creole escaped its ecological niche and spread all over Nigeria since the national independence (1960), it has acquired new functions and is spoken as a second language by speakers whose first language belongs to the four genetic phyla represented by the 500 or so languages spoken in Nigeria. Although it has no official status or standard orthography, it has been adopted for private and informal communication by the educated youth and the Nigerian elite. The Wazobia radio and TV network founded in 2007, uses Naija as its only medium, and the BBC opened a "Pidgin" station in Lagos in 2017, "Pidgin" being the common name used by the locals to name what we call "Naija".

In the process, Naija has developed new structures, a new vocabulary, and probably a new prosody, that differentiate it from Nigerian Pidgin. Despite the ever-growing importance of the language in Nigeria, little attention has been paid to Naija as such, and most of what can be read in the literature concerning the language is based on impressionistic intuitions influenced by previous descriptions of Nigerian Pidgin. This has driven us to start the NaijaSynCor project (NSC), a corpus-based survey of Naija, financed by the French research agency ANR (Caron, 2017). The size of the language, and its geographical span has induced a specific choice of variationist sociolinguistics (Tagliamonte, 2012) as a theoretical framework, and an extensive use of Natural Language Processing tools for our corpus annotation and interpretation.

As it stands now, the NSC corpus counts 321 audio files averaging 5 minutes each, and 319 speakers, which represents a total of 500,000 words collected in 11 locations (see Figure 1). The genres recorded cover life stories, speeches, radio programs, free conversations, cooking recipes, comments on current state of affairs, etc. The sampling of speakers aims at balancing age, sex, education, linguistic and geographic background. Our aim is to annotate each file as finely as possible and prepare queries that cross the linguistic annotation with demographic information collected from each of the 319 speakers. The audio files are annotated with time-aligned transcription and translation into English, morphological tagging, macrosyntactic segmentation, dependency syntax, and prosodic annotation.

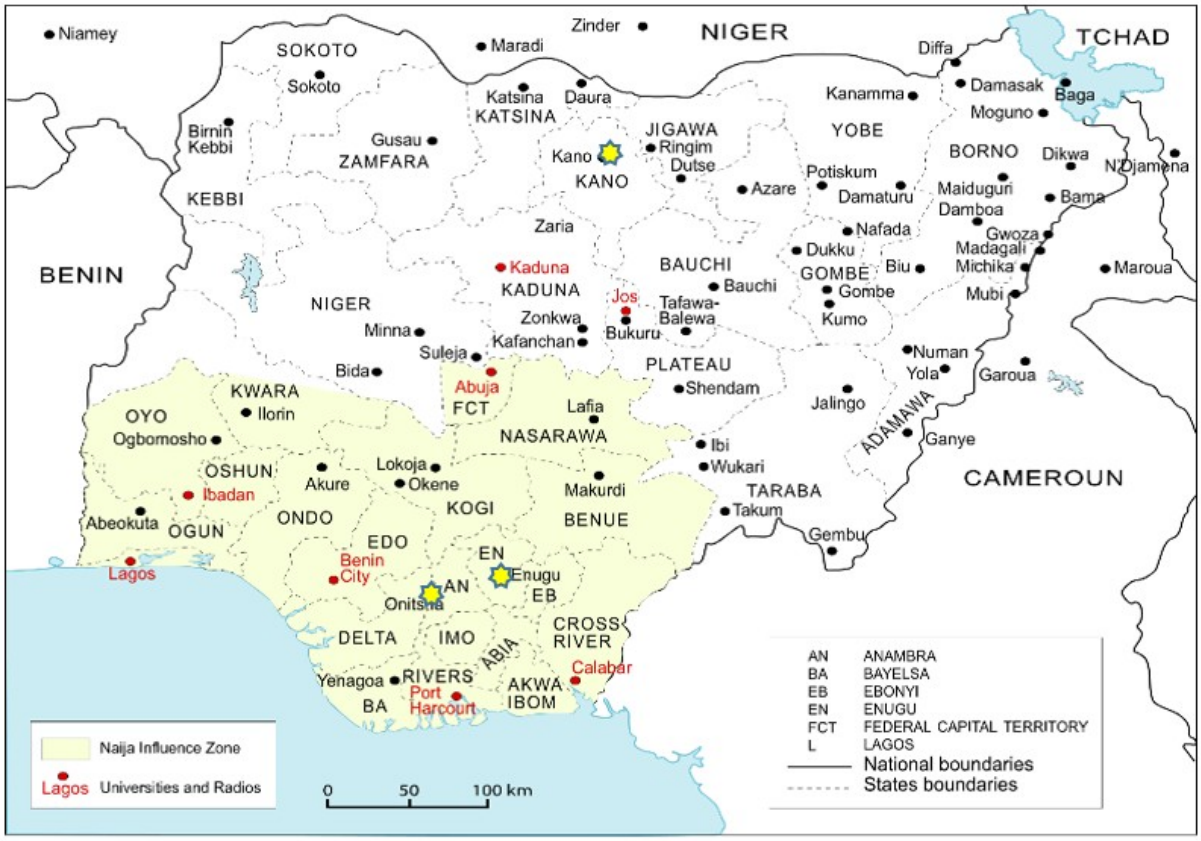


Figure 1. Map of the 11 survey locations

This paper focuses on the syntactic annotation of the corpus and the constitution of a 150,000 words gold standard treebank. The current state of the treebank is accessible at SurfaceSyntacticUD.github.io and can be queried directly at http://match.grew.fr/?corpus=SUD_Naija-NSC@dev. The treebank is currently still undergoing manual and automatic validation. An automatically converted UD version of our treebank will extend the current UD Naija-NSC (available since UD 2.2 (Nivre et al., 2018)) with the upcoming release.

Once the 150k words gold section will be completed, the rest of the corpus will be parsed automatically, together with the 250k words of spoken (historic) Nigerian Pidgin data from Deuber (2005), resulting in a treebank of about 750k words.

Our workflow is explained in section 2, especially the choice of Surface-Syntactic UD, rather than UD. Section 3 presents some interesting constructions in Naija.

2 Treebank development

Section 2.1 concerns the corpus itself (metadata, transcription, translation, and glossing). The particularities of the morphosyntactic annotation, due to the fact that Naija is an English lexifier pidgincreole, are described in Section 2.2. Section 2.3 presents the theoretical choice of our segmentation into maximal syntactic units for this spoken corpus. SUD annotation is developed in Section 2.4. Evaluation is presented in Section 2.5.

2.1 Corpus

Metadata. The variationist analysis we have chosen implies collecting samples representing different types of speakers, and different types of functions. A questionnaire was administered and recorded to provide the relevant metadata about the speakers: time, place and conditions of recording; sex, age, education, professional activity, geographic origin, linguistic background and history. The information was entered into an IMDI[1] database produced using the metadata editor Arbil[2] (Withers, 2012).

[1] ISLE Meta Data Initiative (IMDI) is a metadata standard to describe multi-media and multi-modal language resources.
[2] https://tla.mpi.nl/tools/tla-tools/arbil/

Transcription. Naija is commonly written, in particular on the internet, in forums, but also for example on the BBC website. Although an official orthography or normalization has not taken place, the speakers of Naija have strong opinions on how most words have to be spelled, and we decided to follow these evolving conventions. Mostly, the speakers prefer etymological orthography (i.e. inspired by the Standard English) modified for some emblematic Naija words for which specific spellings have developed, e.g. ***wetin*** 'what', ***moda*** 'mother', ***fada*** 'father', ***dem*** 'they/them/plural marker'. We have used a specific orthography to disambiguate certain function words, e.g. ***de*** (a variant of ***dem***) vs. ***dey*** (the imperfective auxiliary); ***come*** 'to come' and ***con*** (the consecutive auxiliary), ***say*** 'to say', and ***sey*** (the reported speech complementizer). As this emerging orthography is not stabilized, in order to avoid promoting an artificially authoritative norm, we have maintained all the variants in the transcriptions. An example is the word 'thing', which can be written ***ting***, ***tin***, ***thing*** by the annotators. These variants are associated to a common lemma ***ting***, which could be changed later following statistical tendencies that will emerge.

Translation. The translation of all the sentences into English has been done by a team of native speakers of Naija, once the macrosyntactic analysis had been stabilized. It aims at remaining as faithful as possible to the structure and style of the original oral data, keeping the hesitations, repetitions, and general disfluencies. However, the translators have had to strike a balance between a tendency common in Nigerian academics to use erudite and abstruse vocabulary, and on the other hand the risk of using Nigerian English expressions and grammar that would not be understood by non-Nigerians (e.g. a general tendency to use ***would*** instead of ***will*** as a future auxiliary).

2.2 Morphosyntactic analysis

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 tool[3] (Sloetjes and Wittenburg, 2008), 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[4]. The POS tags follow the UD conventions (Nivre et al., 2018) with the caveat that some changes were made to accommodate the specificities of the Naija system. For example, Naija has three copulas, among which two are tagged as VERB (***be*** and ***dey*** 'be') and one is tagged as PART (***na*** 'it is')[5]. 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 SUD 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. The same process was used for the dependency annotation, see Section 2.4.
To assess the quality of the annotation we verified the inter-annotator agreement on three samples composed altogether of 121 sentences. The pre-parsed sample was annotated independently by our three annotators without communication among them and then validated by the expert to obtain the gold annotation. This allows us to compare the inter-annotator agreement based on the pre-parsed structure and measure the difference on the tags and relations that have to be changed to obtain the gold annotation.

[3] https://tla.mpi.nl/tools/tla-tools/elan/
[4] The POS tags where provided by a model of the Mate parser (Bohnet, 2010), other morpho-syntactic features were added by means of a Wapiti-based CRF tagger (Tellier et al., 2010).
[5] The copulas are converted in the POS AUX in UD according to the UD guidelines.

	Percentage of agreement			Percentage of agreement when the annotation differs from the pre-parsed annotation		
	A/B[6]	A/C	B/C	A/B	A/C	B/C
UPOS	95	94	95	46	41	37
UAS	93	91	91	68	60	58
LAS	89	86	87	60	51	50

Table 1. Inter-annotator agreement scores

The agreement scores are then improved by the final adjudication, and our semi-automatic query of the corpus to look for inconsistencies using the grew tool.

2.3 Macrosyntactic segmentation

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; Degand and Simon, 2009). 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; Pietrandrea and Kahane, 2019), which is a kind of formalized punctuation. The delimiter for illocutionary units is //. Consider this extract from a sample illustrating the markup:

(1) 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]

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

(2) { some || some } people dey ask [e good make man { get || go } test im children ?//] //
'some, some people were asking: "Is it good for a man to get… go and test his children ?"'
[ABJ_GWA_09_Journalism_48]

An illocutionary unit is organized around a nucleus that bears the illocutionary force and some optional and non-autonomous components we call ad-nuclei. The nucleus is separated from pre- and post-nuclei by the delimiters "<" and ">":

(3) and many of dem wey vote dat time < na because of internet //
'and many of those who voted at the time, it was because of the internet' [ABJ_GWA_09_Journalism_27]

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 (Liu et al., 2019). The macrosyntactic annotation improves parsing results (see Section 2.5) and it can easily be simplified into a standard punctuation.

2.4 SUD

Two different strands of thought, one rather practical, the other more theoretical, have led us to annotating the corpus not in the standard UD dependency annotation scheme but rather in the Surface-Syntactic UD scheme (SUD) (Gerdes et al., 2018).

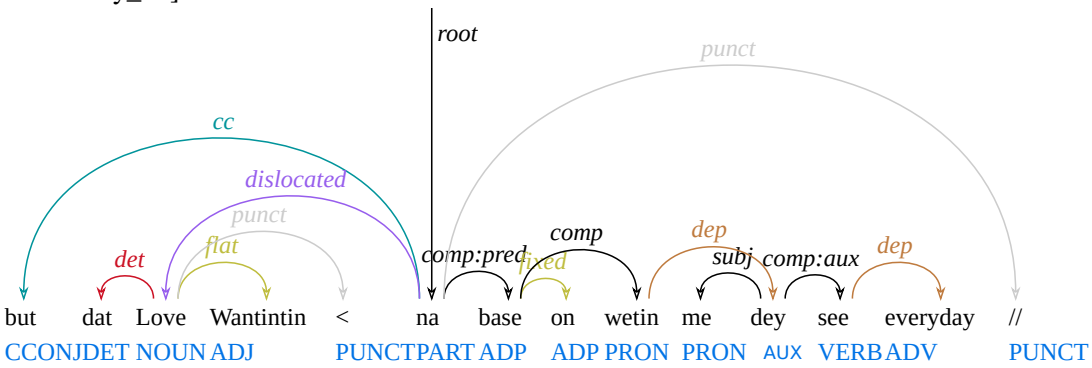
Firstly, the Nigerian annotators have been trained in a standard syntactic X-bar sentence structure, where, for example, a PP is headed by a preposition (Osborne and Gerdes, 2019). In this context, SUD is much easier to acquire than UD dependencies (Gerdes et al., 2019).

[6] We look at agreement between pairs of annotators, A/B means we are looking at the agreement between the annotator A and annotator B

Secondly, the NaijaSynCor project has a central typological component, and language comparisons should be possible, based on syntactic differences, which is easier in a scheme based purely on distributional criteria, such as SUD, than on the rather semantic function word vs content word distinction that constitutes the basis of UD.

We can add that UD is particularly problematic for multi-words expression (MWEs) working as functional items (complex adpositions or complex conjunctions), especially when they are syntactically quite regular (Kahane et al., 2018). In SUD, MWEs such as the Naija adposition *base on* '(based) on'[7] are connected and in the dependency tree they occupy the same syntactic position as a simple word, see (4).

(4) but dat Love Wantintin[8] < na base on wetin me dey see everyday //
'but that Love Wantintin, it is based on what I see everyday' [WAZK_11_M_Chiagozies-Life-Story_21]



The Naija treebank uses the SUD version proposed in (Gerdes et al., 2018), which is automatically converted into UD. Contrary to UD, two elements that are mutually exclusive and thus occupy the same syntactic position are linked to their governor by the same relation: For instance, *the problem* and *you're wrong* are both **comp:obj** in *I know the problem* and *I know you're wrong*, while the first is **obj** and the second is **ccomp** in UD. Considering that most of our readers are more or less familiar with UD, we choose to explain the specific SUD relations and how they are converted into UD. Adpositions (ADP) and subordinating conjunctions (SCONJ) govern the complement they introduce by the relation **comp**. These relations are reversed in UD: ADP **comp>** NOUN becomes NOUN **case>** ADP in UD and SCONJ **comp>** VERB becomes VERB **mark>** SCONJ.

For dependents of verbs, we distinguish between subjects (**subj**), complements (**comp**) and modifiers (**mod**). The relation **subj** becomes **nsubj** or **csubj** in UD according to the POS of the dependent. The relation **mod** becomes **advmod** for adverbs, **obl** for prepositional phrases, or **advcl** for clauses in UD. For verb complements, we distinguish the following sub-relations:

- **comp:obj**, for direct objects, which, in UD, becomes **obj** for a nominal dependent and **ccomp** for a clausal dependent;
- **comp:obl**, for oblique complements, which becomes **ccomp** for a verbal or clausal dependent, **iobj** for a nominal (or pronominal) dependent, and **obl** in other cases;
- **comp:pred,** for relations between two predicates that share an argument. This relations generally corresponds to UD's **xcomp**[9] but is reversed when the governor is a copula (AUX): AUX **comp:pred>** VERB becomes VERB **cop>** AUX in UD.
- **comp:aux,** for relations between a TAM (Tense–Aspect–Mood) auxiliary and the full verb, which is also reversed in UD.
- **compound:svc** is used for serial verb constructions, which are typical for Naija (see Section 3.3).

The difference between UD and SUD annotations is exemplified in Figure 2.

(5) dem go seize am //

[7] "base on" is not a passive construction in Naija as there is not morphological passive.
[8] *Love Wantintin* is the name of a radio programme.
[9] As remarked by (Przepiórkowski and Patejuk, 2018) and (Gerdes et al., 2018), raising is orthogonal to the syntactic function and it would be better to add **…:pred** to the syntactic function in case of raising, which would give us **comp:obj:pred** for objects with raising, **comp:obl:pred** for obliques with raising and **mod:pred** for modifiers with raising (such as *without talking* in *She explained it without talking*).

'They will seize it.' [DEU_C01_D_6]

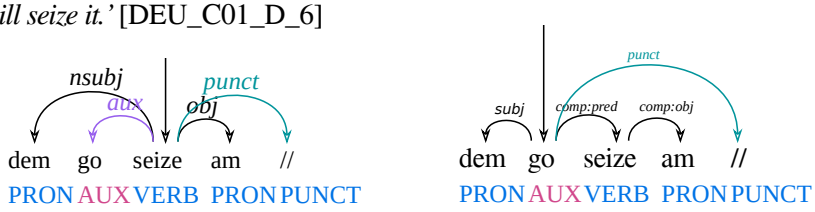


Figure 2. UD analysis vs. SUD analysis of (5)

All samples are first annotated by a trained annotator and the resulting trees together with the POS tags are then validated by an expert. Difficult cases are discussed among the annotators and the shared annotation guide is constantly updated. We apply simple error mining techniques such as looking for inconsistencies between the dictionary and the treebank. The SUD annotation scheme is a still ongoing process, and some special needs for the annotation of Naija have provided input for improvement of SUD.

2.5 Evaluation of treebank coherence and the impact of macrosyntactic annotation

In this section we will present the results of the annotation by evaluating parser performance on the current state of the treebank. In particular, we evaluate the relevance of macrosyntactic markup for syntactic parsing. We expect the macrosyntactic annotation to have a positive influence on dependency parsing, in particular for constructions such as coordination and dislocation, which have specific macrosyntactic markups resulting in specific dependency relations.

In order to verify this claim, we trained the Mate tagger and parser by Bohnet (2010), first on a version of the treebank with these markups, and then on a version of the treebank where they have been removed except for "//" (the segmentation into illocutionary units ≃ sentences). This type of experiment is also important to set a baseline for the development of a Naija parser, to be used for parsing the rest of the NSC corpus (which is transcribed and macrosyntactically annotated) as well as for parsing other spoken and written data without markup.

We used a sample of 52k words, with 90% training and 10% test data on the Mate parser. While the POS tagging scores are as expected very similar whether macrosyntactic annotation is present or not, we obtain an LAS error reduction of 11% and a UAS error reduction of 18% through macrosyntactic annotation, see Table 1.[10]

	Macro-syntax +	Macro-syntax -	Error reduction
UPOS	92.44	92.23	*
UAS	90.76	89.23	18%
LAS	84.45	82.02	11%

Table 2. Parsing results with and without macro-syntactic annotation.

Unsurprisingly, the syntactic functions which most benefit from this type of markup are those that are targeted by the annotation, such as piles (paradigmatic relations like **conj:coord**, **conj:dicto**, **compound:redup**) and coordinators (**cc**). We also observe an improvement for relations that connect a nucleus and adnuclei, such as clefts, dislocations, and peripheric modifiers.

The parser scores are promising,[11] in particular for spoken texts, and we hope to further improve the parser performance by the ongoing process of semi-automatic rule-based enhancement of the treebank coherence. In particular, we address this problem of annotation inconsistencies by a systematic comparison of parsing results with the gold annotation and the double SUD-UD-SUD conversion, and by different error mining tools such as the relation table proposed by the grew tool available on match.grew.fr (Bonfante et al., 2018), which shows the number of dependency relation types between any pair of categories. Also, the move to a neural network-based parser can be expected to result in better scores.

[10] **punct** relations are excluded from the evaluation as they are exclusively used for macro-syntactic markers.

[11] Although Naija was part of the CoNLL 2018 Shared Task (Multilingual Parsing from Raw Text to Universal Dependencies), it is difficult to compare the results as Naija was one of the low-resource languages. The best score for UPOS, UAS, and LAS are 67.93, 38.62 and 30.07 (Zeman et al., 2018).

3 Some idiosyncratic syntactic constructions of Naija

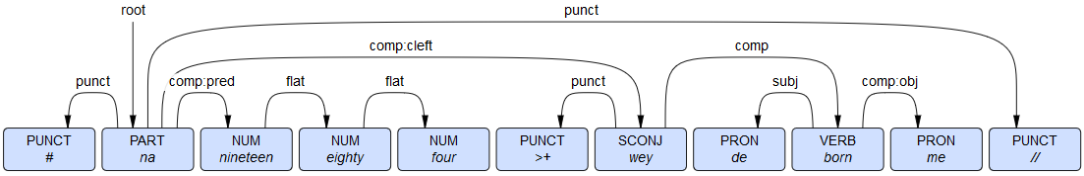
In this section we present three interesting constructions of Naija that show clear structural differences with English, Naija's lexifying language .

3.1 Na-clefts and modifying relative clauses

Surface-syntax UD nicely captures the complexity of clefts[12] in Naija and the way they contrast with modifying relative clauses. We will restrict our presentation to one of the three cleft structures of Naija, *wey*-clefts.

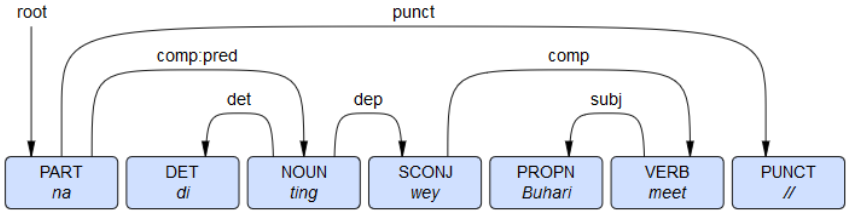
In Naija, clefts use the copula *na*,[13] that has two complements: First, a *predicative complement*, linked by the relation **comp:pred** (*na → nineteen eighty-four* in (6)) and second, a clause introduced by *wey*, that we link to *na* with the **comp:cleft** relation (*na → wey de born me* in (6)). In the macrosyntactic markup, we use ">+" before the cleft clause.

(6) # na nineteen eighty four >+ wey de born me //
it's 1984 >+ that they bare me
'it's in nineteen eighty-four that I was born' [KAD_09_Kabir-Gymnasium_P_6][14]



Cleft sentences are superficially similar to copular predications in which the relative clause modifies the predicative complement. Yet, in clefts, the relation between the antecedent and the cleft clause is mediated by the copula, and the cleft clause is not dependent on the predicative complement but is raised and attached to the copula; whereas copular predications are thetic sentences that have a copula, a nominal, and a relative clause, but no syntactic restructuring and no backgrounding of the relative clause. The thetic meaning is clear when *na* has a presentational and not an identificational meaning, see example (7). In the syntactic representation of modifying relative clauses in copular non-identifying clauses, the copula takes only one complement: **comp:pred** (*na → di ting wey Buhari meet*):

(7) na di ting wey Buhari meet //
'this is the thing that Buhari found' [IBA_25_Buying-Indomi_159]



3.2 Interrogatives

In the NSC corpus, content questions are analyzed as clefts. This is corroborated by examples where the question word of a content question can be preceded by the copular particle *na* without changing the behaviour or meaning of the sentence. The following two questions (8) occur in direct sequence and show how the copula *na* can be present or not without semantic consequence:

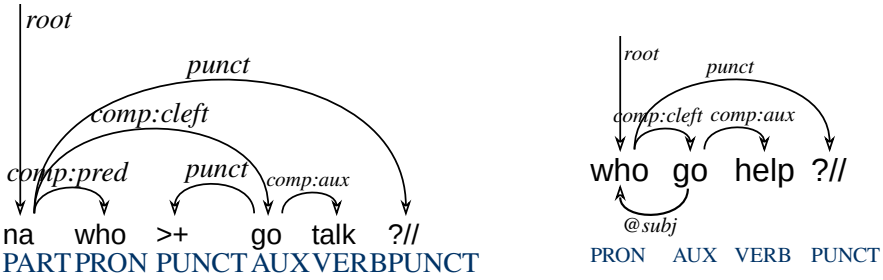
(8) **na** who >+ go talk ?// who go help ?//

[12] Clefts are defined by Lambrecht (2001) as "a complex sentence structure consisting of a matrix clause headed by a copula and a relative or relative-like clause whose relativized argument is co-indexed with the predicative argument of the copula. Taken together, the matrix and the relative express a logically simple proposition, which can also be expressed in the form of a single clause without a change in truth conditions".

[13] ***na*** is classified as a particle and not a verb or an auxiliary in Naija because it cannot be negated or combined with TAM markers, two of the defining features of (auxiliary) verbs.

[14] The NaijaSynCor project, entirely based on oral data, intends to study, among others, the interface between prosody and syntax. The # stands for a pause in the speech unit, a major cue for the study of prosody. The # is identified as a punctuation mark (PUNCT) in the syntactic representation.

'Who will speak? Who will help?' [ENU_33_A-Beg_P_56, 57]



This leads us to interpret question-words as focused, and the rest of the sentence as the focus-frame. In the absence of the focus particle *na*, the question word becomes promoted to **root** of the sentence through deletion of its previous head. In this analysis, the question word has a double function: It is the root of the sentence and a dependent of the verb.

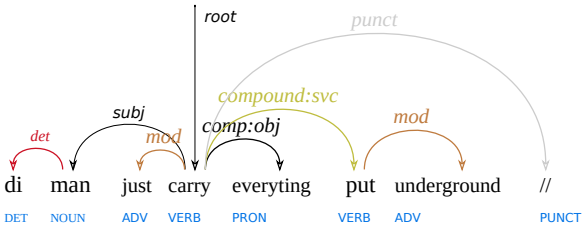
The complexity of the cleft structure of content questions cannot be captured by UD which treats the sentence verb as a root. Moreover, the parallelism between the two questions will not be kept by UD, as the one with *na* will be treated as a cleft, with the cleft phrase as the head, contrary to the other question without *na*. As a compromise between surface syntax and convertibility to UD, a second link has been added to the root, which annotates explicitly the dependency of the question word (this second relation is preceded by a "@", see **@subj** in (8)). In our Surface-syntactic representation, both cases are represented by means of a cleft structure, see the above analysis. This is congruent with many analyses of wh-words which consider that they occupy two syntactic positions, one as a complementizer and another as a pronoun inside the clause they complementize (see, in particular, (Tesnière, 1959[2015]: ch. 246)).

During the conversion into UD we can only keep one of the relations, we have to keep the second relation as this follows the UD analysis of relative clauses. This leads to a "catastrophe" between the two syntactically related interrogative constructions (Gerdes and Kahane, 2016).

3.3 Serial Verb Constructions

The influence of adstrate vernacular languages, belonging mainly to the Niger-Congo family, is observed in the use of Serial Verb Constructions, that is "monoclausal construction[s] consisting of multiple independent verbs with no element linking them and with no predicate-argument relation between the verbs." (Haspelmath, 2016). We used the subtyped relation **compound:svc** for these constructions. Sentence (8) contains an example of a serial verb construction (*carry → put*).

(9) di man [...] just carry everyting put underground //
'The man just carried everything and put it underground.' [ABJ_INF_12_Evictions_P_13]



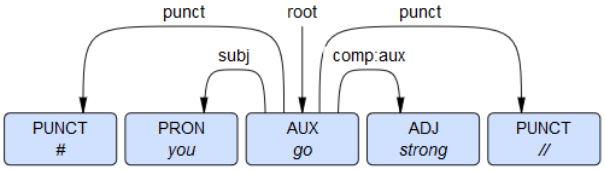
3.4 Polycategoriality and polyfunctionality

Following the UD guidelines[15], the morphological specification of a (syntactic) word in the UD scheme consists of three levels of representation: a lemma, a POS tag, and a set of features representing lexical and grammatical properties. In order to reduce polycategoriality and its consequent multiplication of syntactic words, the annotation process has been guided by the principle of separation of the morphological tagging of a word from its syntactic function: A single lexeme can be polyfunctional, but it cannot be polycategorial. This principle applies in all languages, e.g. to adpositions (ADP) which can take a nominal, clausal or zero complement without changing their abstract lexical category (Huddleston and Pullum, 2008).

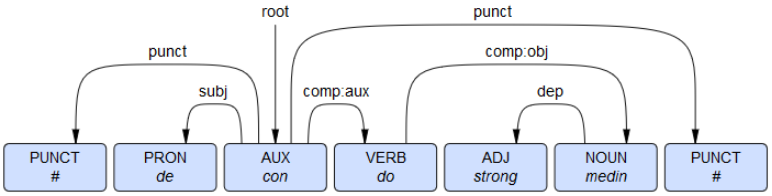
[15] https://universaldependencies.org/u/overview/morphology.html

This principle has been applied to adjectives in Naija, which can function as predicates without any copula (10) or noun modifiers (11). In both cases, the words keep their morphological assignment: they are ADJ.

(10) # you go **strong** //
'you will (be) strong' [PRT_05_Ghetto-life_P_24]

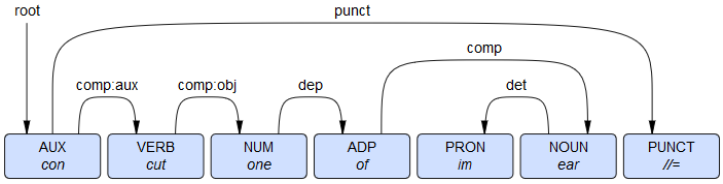


(11) # de con do **strong** medin #
*'they then did **strong** magic'* [IBA_04_Alaska-Pepe_P_95]

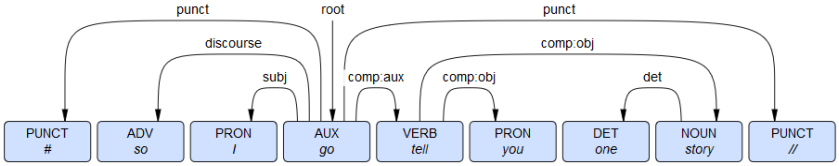


However, some lexical words are grammaticalized into new function words. An example is given by the numeral *one*, tagged NUM (12), which has grammaticalized into the determiner *one* 'some, a certain' (13), tagged DET, and the pronoun *one*, tagged PRON (14).

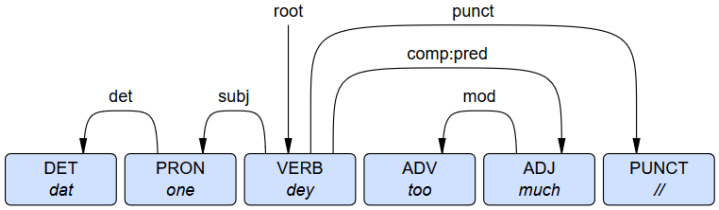
(12) con cut **one** of im ear //=
*'he then cut **one** of its ears'* [IBA_04_Alaska-Pepe_P_168]



(13) # so I go tell you one story //
'so I will tell you a story' [IBA_04_Alaska-Pepe_P_5]



(14) dat **one** dey too much //
*'that **one** is too much'* [ABJ_INF_08_Impatience_106]



3.5 Naija Grammar

The preliminary assessment of the NSC corpus has proved two things. First, despite the diversity of its speakers in terms of geographic origin and mother tongues, the corpus is remarkably homogeneous. Second, this homogeneity takes place while distancing the language from Nigerian Pidgin. Not only is new vo-

cabulary acquired through the necessity to cope with new functions and new cultures, but new grammatical structures are emerging and a new stability is found in the use of competing structures.

The study of Naija clefts is a good indicator. Naija clefts have three variants: ***wey*-clefts**, with a relative clause introduced by the relativizer *wey* (Section 3.1, example (5)), **bare clefts**, where the relativizer is omitted, resulting in a bare relative clause and **double clefts**, where the relativizer *wey* is replaced by a repetition of the copula followed by an expletive invariable 3sg pronoun: *na im*. They are exemplified in Table 1.[16]

1b'	***wey*-cleft**	***na*** *weekend* ***wey*** *we dey do am*	
1b"	**bare** cleft	***na*** *weekend* ***Ø*** *we dey do am*	'It's in the weekend that we do it.'
1b"'	**double** cleft	***na*** *weekend* ***na im*** *we dey do am*	

Table 3. The three 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.

4 Conclusion

We have described the workflow for the development of the gold section of the NSC treebank in the SUD annotation scheme, and we have shown the SUD analysis of some interesting syntactic constructions of Naija.

In parallel with the treebank construction we develop various interfaces to access the audio corpus, the transcription, and the different annotations. For example the most recent version of the SUD syntactic annotation is accessible at match.grew.fr/?corpus=SUD_Naija-NSC@dev.

In order to be part of the UD treebank family, an automatically converted UD version of the treebank will also be distributed, although the current UD platform does not foresee the joint distribution of the audio data. The increasing interest in spoken data will surely bring the UD community to discuss the format that will best allow for phonosyntactic studies. We will also distribute two text versions, one with macrosyntactic markup and a second version without the markup that can be used to train parsers on bare texts.

The perspective of this treebank creation goes beyond purely linguistic interest. It has deep sociolinguistic implications through the creation of a Naija dictionary. In order to create this treebank, we had to create an inventory of spelling variants, and we propose systematic distinctions of function and content words. The tools and resources of the NSC treebank enhances the interest in the specificity of Naija grammar, and the project can be seen as a step in the further establishment of Naija as a language (Courtin et al., 2018).

Acknowledgments

This research is financed by the French National Research Agency via the project ANR NaijaSynCor. We would like to express our gratitude to the Nigerian annotators: Onwueqbuzia Emeka Felix, Ajede Chika Kennedy, and Tella Sansom Adekunle. The quality of the treebank has been largely enhanced by the constant interaction with Bruno Guillaume, the developer of the Grew platform. Thanks also to the anonymous reviewers of the SyntaxFest 2019 that helped us to clarify this paper.

References

Peter Bakker. 2009. Pidgins versus Creoles and Pidgincreoles. *The Handbook of Pidgin and Creole Studies*, John Wiley & Sons, 130–157.

[16] See (Caron, 2019) for a complete presentation of clefts in Naija.

Bernd Bohnet. 2010. Very High Accuracy and Fast Dependency Parsing is not a Contradiction. *Proceedings of the 23rd International Conference on Computational Linguistics* (*COLING*), Beijing, China.

Guillaume Bonfante, Bruno Guillaume, Guy Perrier. 2018. *Application of Graph Rewriting to Natural Language Processing*. John Wiley & Sons.

Bernard Caron. 2017. N*aijaSynCor : A corpus-based macro-syntactic study of Naija (Nigerian Pidgin)*. naijasyncor.huma-num.fr (September 2017).

Bernard Caron. 2019. Clefts in Naija, a Nigerian pidgincreole. *Linguistics Discovery,* 41p.

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

Marine Courtin, Bernard Caron, Kim Gerdes, Sylvain Kahane. 2018. Establishing a Language by Annotating a Corpus. *Proceedings of the Workshop on Annotation in Digital Humanities* (*annDH*), Sofia, ceur-ws.org/Vol-2155, 7-11.

Emanuela Cresti. 2000. *Corpus di italiano parlato*. Accademia della Crusca, Florence.

Lisbeth Degand, Anne-Catherine Simon. 2009. On identifying basic discourse units in speech: theoretical and empirical issues. *Discours,* 4, discours.revues.org.

Henri-José Deulofeu, Lucie Dufort, Kim Gerdes, Sylvain Kahane, Paola Pietrandrea. 2010. Depends on what the french say: Spoken corpus annotation with and beyond syntactic function, *Proceedings of the 4th Linguistic Annotation Workshop* (*LAW IV*).

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

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

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

Nicholas Faraclas. 2013. Nigerian Pidgin structure dataset. 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/contributions/17.

Kim Gerdes, Sylvain Kahane. 2009. Speaking in piles: Paradigmatic annotation of french spoken corpus. *Proceedings of the Fifth Corpus Linguistics Conference,* Liverpool.

Kim Gerdes, Sylvain Kahane. 2016. Dependency annotation choices: Assessing theoretical and practical issues of universal dependencies. *Proceedings of the 10th Linguistic Annotation Workshop* (*LAW-X*), ACL, 131-140.

Kim Gerdes, Bruno Guillaume, Guy Perrier, Sylvain Kahane. 2018. SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD, *Proceedings of the Universal Dependencies Workshop* (*UDW*), EMNLP, Bruxelles.

Kim Gerdes, Bruno Guillaume, Guy Perrier, Sylvain Kahane. 2019. Improving Surface-syntactic Universal Dependencies (SUD): surface-syntactic relations and deep syntactic features, *Proceedings of the Universal Dependencies Workshop* (*UDW*), SyntaxFest, Paris.

Rodney Huddleston, Geoffrey K. Pullum. 2008. *The Cambridge Grammar of the English Language (*2nd ed.), Cambridge: Cambridge University Press.

Sylvain Kahane, Marine Courtin, Kim Gerdes. 2018. Multi-word annotation in syntactic treebanks: Propositions for Universal Dependencies, *Proceedings of the 16th international conference on Treebanks and Linguistic Theories* (*TLT*), Prague.

Knud Lambrecht. 2001. A framework for the analysis of cleft constructions. *Linguistics* 39(3). 463–516.

Luigi (Yu-Cheng) Liu, Anne Lacheret-Dujour, Nicolas Obin. 2019. Automatic Modelling and labelling off speech prosody: What's new with SLAM+?. *Proceeding of the International Congress of Phonetic Sciences* (*ICPhS*), Melbourne.

Joakim Nivre, Mitchell Abrams, Željko Agić, Lars Ahrenberg, Lene Antonsen, Maria Jesus Aranzabe, Gashaw Arutie, et al. 2018. *Universal Dependencies 2.2*. https://hal.archives-ouvertes.fr/hal-01930733.

Timothy Osborne, Kim Gerdes. 2019. The status of function words in dependency grammar: A critique of Universal Dependencies (UD). *Glossa: a journal of general linguistics*, *4*(1).

Paola Pietrandrea, Sylvain Kahane. 2019. The macrosyntactic annotation. In Anne Lacheret-Dujour, Sylvain Kahane , Paola Pietrandrea (eds.), *Rhapsodie: A Prosodic and Syntactic Treebank for Spoken French*, John Benjamins, Amsterdam.97-126.

Adam Przepiórkowski, Agnieszka Patejuk. 2018. Arguments and adjuncts in Universal Dependencies. In *Proceedings of the 27th International Conference on Computational Linguistics* (*COLING*), Santa Fe, NM, 3837–3852.

Han Sloetjes, Peter Wittenburg. (2008). Annotation by category – ELAN and ISO DCR. *Proceedings of the 6th International Conference on Language Resources and Evaluation* (*LREC*)

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

Isabelle Tellier, Iris Eshkol, Samer Taalab, Jean-Philippe Prost. 2010. POS-tagging for oral texts with CRF and category decomposition. *Research in Computing Science,* 46, 79-90.

Lucien Tesnière. 1959. *Éléments de syntaxe structurale*. Paris: Klincksieck. [Transl. by Timothy Osborne, Sylvain Kahane, 2015. *Elements of structural syntax,* Amsterdam: John Benjamins.]

Peter Withers. 2012. Metadata Management with Arbil. In Victoria Arrantz, Dan Broeder, Bertrand Gaiffe, Maria Gavrilidou & Monica Monachini (eds.), *Proceedings of the workshop Describing LRs with Metadata: Towards Flexibility and Interoperability in the Documentation of LR,* LREC, 72–75.

Daniel Zeman et al. CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies. *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*.