NAIJA SYNCOR

cnrs

A Surface-Syntactic UD Treebank for Naija

B.Caron, M.Courtin, K.Gerdes, S.Kahane

SyntaxFest 2019
Paris, August 26-30 2019

Langage, Langues et Cultures d'Afrique Noire

LPP Laboratoire de Phonétique et Phonologie

modyco Modèles, Dynamiques, Corpus UMR 7114

NaijaSynCor (ANR)

- Sociolinguistic snapshot of Naija (Nigeria)
- Corpus-based
- Variationist
- Syntax, Morphology, Lexicon, Intonation

- Syntax = (S)UD

1. Introduction: background information on Naija and challenges implied
2. Corpus and treebank development
3. Some idiosyncratic grammatical constructions in Naija
4. Conclusion

1. Naija

Ogini Bernard:
Oga Pikin
(2018)

- **Naija** (Common Nigerian Pidgin)
- 100 million speakers
- No official status
- Under-resourced
- Nigeria: 200 million inhabitants

- Syntactic Treebank
- Surface-Syntactic Universal Dependency annotation scheme (SUD)

  (Gerdes et al., 2018)
- Part of an ANR project
- Sociolinguistic snapshot of Naija
- 500k word corpus

Map of the 11 survey locations

The emergence of Common Nigerian Pidgin

- Nigerian Pidgin

  - Has creolised in the Niger Delta (2 to 10 million speakers) and in Lagos where it is a 1st language
  - **But**: has since the National Independance (1960) expanded to most of Nigeria where it is learnt as a 2nd language.
  - 100 million speakers. Intercomprehension with other languages (e.g. Cameroon, Ghana, Sierra Leone, Equatorial Guinea, etc.)

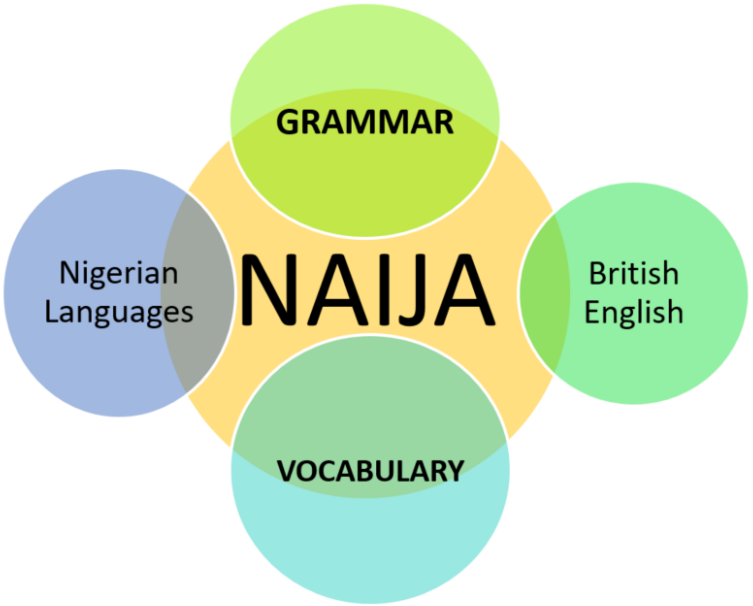
  - **One of the largest languages in the world**.

Nigerian Pidgin: a multitude of definitions

- An expanded pidgin (Mufwene)
- A postcreole continuum
- A pidgincreole in the process of becoming a vernacular language

- **But most of all** : a language that is fast expanding (both in geography and function) and rapidly changing, and is emerging under a new form: **Common Nigerian Pidgin**

The structure of Naija



- The majority of Nigerian languages are Benue-Congo of Niger Congo.
- There is a basic substrate structure and grammatical frame, no matter the original language of contact.
- The process of language learning involves the insertion of lexical frames into the common grammatical frame.
- There is a common core of popular vocabulary that defines the Naija lexicon.

2. Treebank development

1. Corpus
2. Morphosyntactic analysis
3. Macrosyntactic segmentation
4. SUD
5. Evaluation of treebank coherence

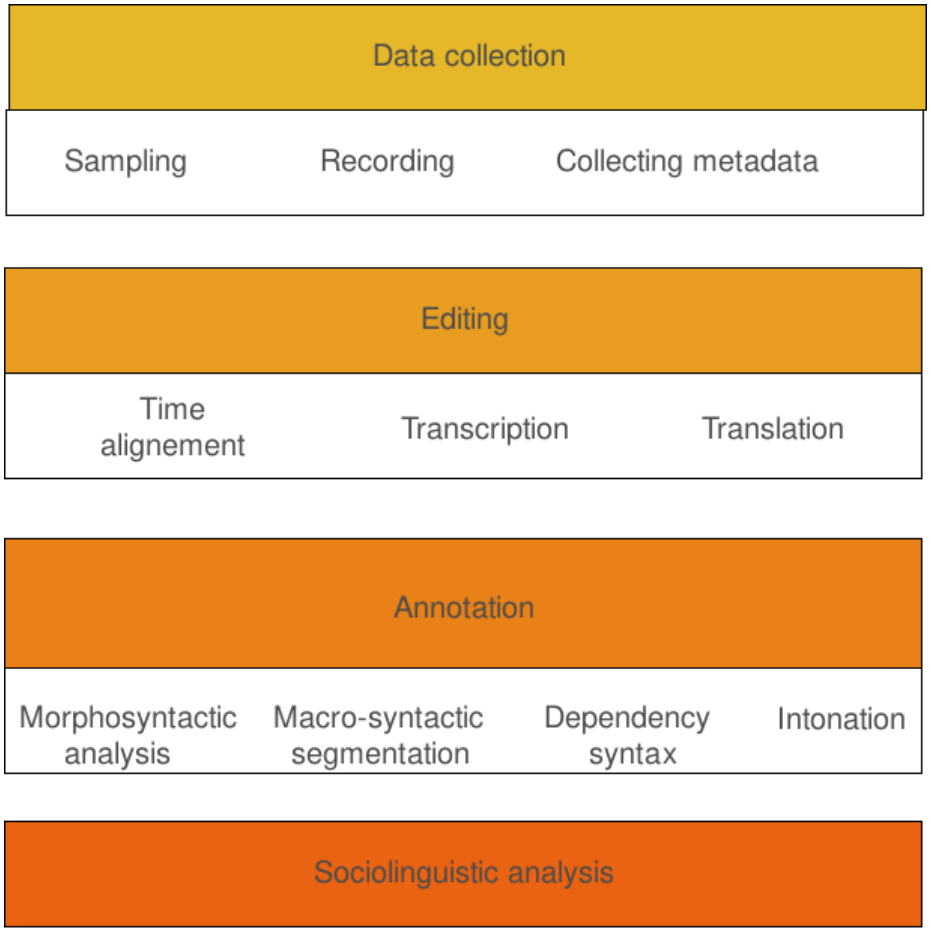
2.1 Corpus

Gold	Silver	Deuber (2005)
150k	350k	250k

Current gold (125k) :

- Download at https://github.com/surfacesyntacticud/SUD_Naija-NSC
- Query on http://match.grew.fr/?corpus=SUD_Naija-NSC@dev

2.1 Corpus


Data collection
Sampling
Recording
Collecting metadata
Editing
Time alignement
Transcription
Translation
Annotation
Morphosyntactic analysis
Macro-syntactic segmentation
Dependency syntax
Intonation
Sociolinguistic analysis

2.1 Morphosyntactic analysis

- We follow UD guidelines for POS and morphological features.

- Workflow:
  - A fewn first sample texts were was tagged and parsed with a model trained on English + manual corrections
  - Dictionary of the function words and most common lexical items of Naija containing
    - Form and orthographic variants
    - POS tag
    - frequency
    - English gloss (if necessary)

• 2.3 Macrosyntactic segmentation

- Spoken data -> we need a segmentation step to define the maximal units of syntax: the illocutionary units (Blanche-Benveniste et al. 1990, Cresti 2000, Degand & Simon 2009).

- Markup developed in the Rhapsodie project (Deulofeu et al., 2010; Pietrandrea and Kahane, 2019), represents a kind of formalized punctuation.

• 2.3 Macrosyntactic segmentation

- Encodes information that is particularly relevant for spoken langages :
  - Sentence segmentation
    - Illocutionary Units
    - Pre and post-nuclei
  - Lists
    - Coordination
    - Disfluencies
    - Reformulations
- 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]
- 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]

• 2.3 Macrosyntactic segmentation

- Also used to indicate code-switching :

- { di suspect |a twenty two years old Stephen Otuyi } < dem say

  [di guy nko < e go [yor ledi apo po yor] //] //

[IBA_33_News-Comments]

• 2.5 Evaluation of treebank coherence

	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

Still a lot of disagreements when annotators deviate from the pre-parsed annotation :

- High inter-annotator agreement due to pre-parse ?
- The annotators disagree on more difficult cases ?

Some idiosyncratic syntactic constructions of Naija

The preliminary assessment of the NSC corpus has proved two things.

- The corpus is remarkably homogeneous.

- Distancing the language from Nigerian Pidgin.
  - new vocabulary
  - new grammatical structures
  - new stability in the use of competing structures.

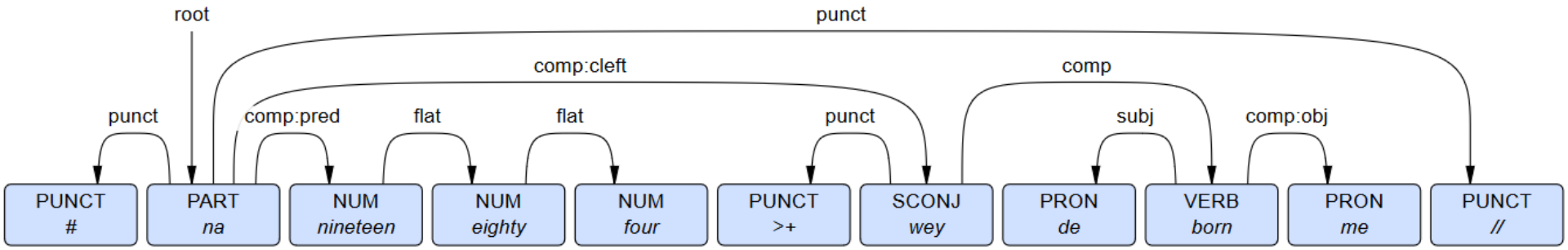
1. Na-clefts and modifying relative clauses
2. Interrogatives
3. Serial Verb Constructions

• 3.1 Na-clefts

(6)

	#	*na*	*nineteen*	*eighty*	*four*	>+	*wey*	*de*	*born*	*me*	//
	#	COP	1984			>+	that	they	bear	me	//

'it is in nineteen eighty-four that I was born' [P_KAD_09_6]



Innovation in Naija Clefts

- 4 types of clefts

'It's in the weekend that we do it.'

wey-cleft	**na** weekend **wey** we dey do am
bare cleft	**na** weekend **Ø** we dey do am
zero-copula cleft	**Ø** weekend **Ø** we dey do am
double cleft	**na** weekend **na im** we dey do am

The emergence of double-clefts in Naija

	Nigeria Pidgin*	Naija
wey-clefts	41%	1%
bare clefts	39%	89%
zero-copula clefts	17%	1%
double clefts	—	9%

Faraclas, Nicholas. 2013. Nigerian Pidgin structure dataset. In Michaelis *et al.* (eds.), *Atlas of Pidgin and Creole Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

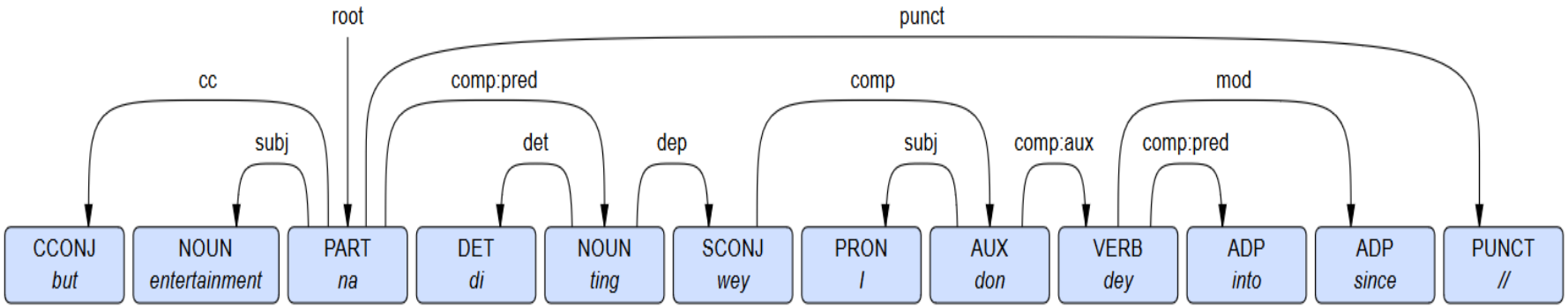
Modifying relative clause

- The relative clause is directly dependent on the predicative complement (*ting*)

(7)

but	*entertainment*	*na*	*di*	*ting*	*wey*	*I*	*don*	*dey*	*into*	*since*	*//*
but	entertainment	COP	the	thing	that	I	PAST	am	into	for_long	//

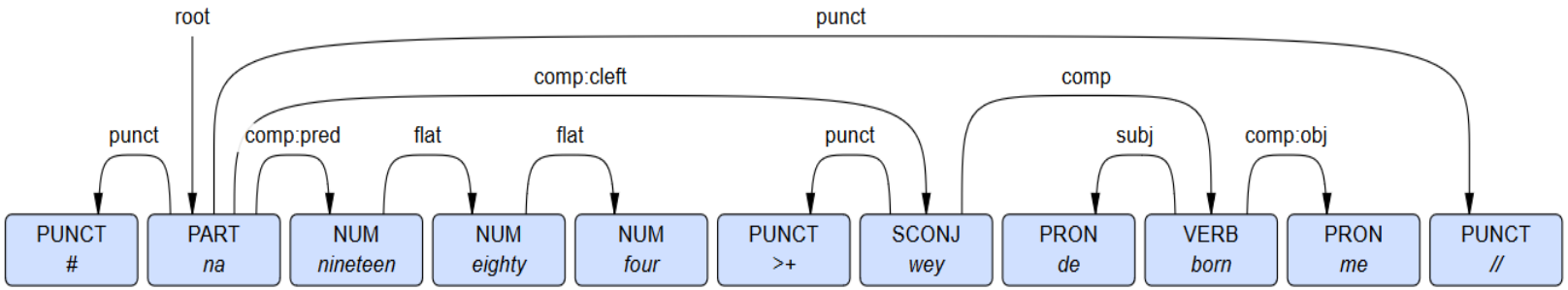
'But *entertainment* is the thing that I have been into for a long time.' [P_WAZA_10_90]



NB: Clefts

- the relation between the antecedent (1984) and the cleft (relative) clause is mediated by the copula
- the cleft clause is not dependent on the predicative complement (1984) but is raised and attached to the copula

‘It is in 1984 that I was born’

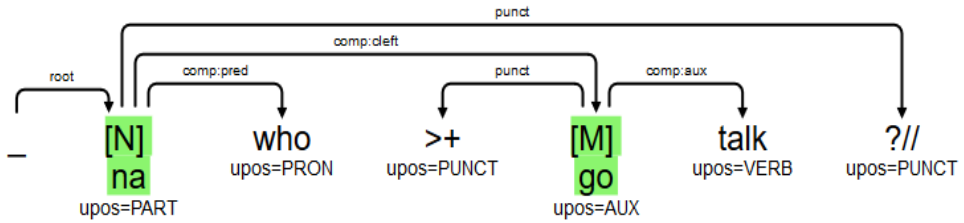


• 3.2 Interrogatives

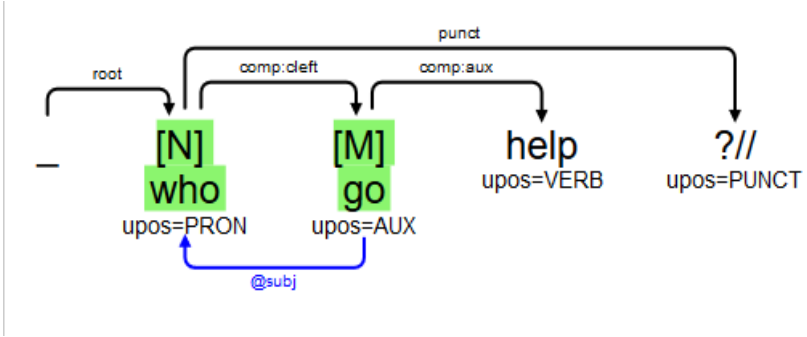
- In the NSC corpus, content questions are analyzed as clefts.

(8)	*na*	*who*	*>+*	*go*	*talk*	*?//*	*who*	*go*	*help*	*?//*
	COP	who		AUX.FUT	talk		who	AUX.FUT	help	

'Who will speak? Who will help?' [ENU_33_56, 57]



- The question-word is focused, and the rest of the sentence is the focus-frame
- In the absence of the focus particle *na*, the question word becomes promoted to **root** of the sentence.
- The question word has a double function: It is the root of the sentence and a dependent of the verb.



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 “@”

• 3.3 Serial Verb Constructions

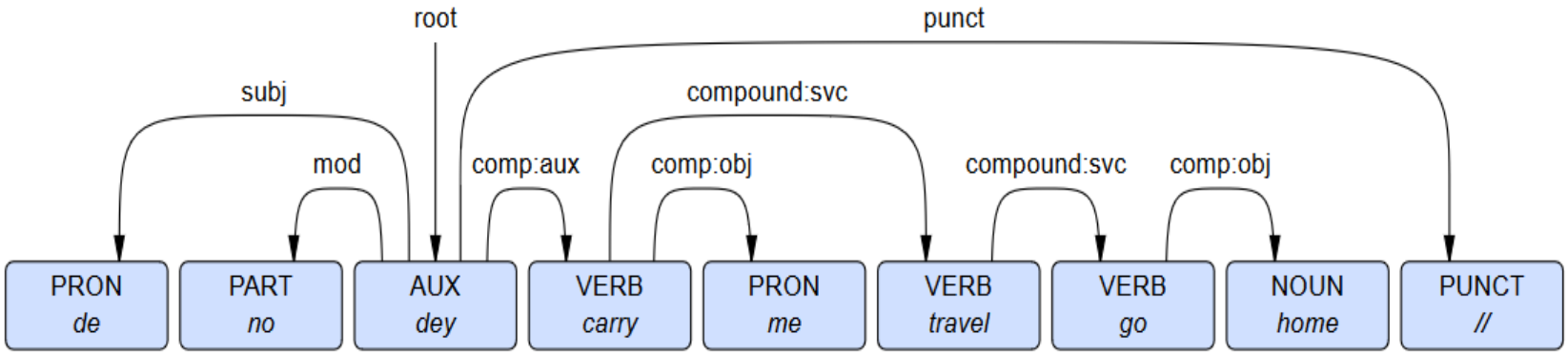
- “monoclausal construction[s] consisting of multiple independent verbs with no element linking them and with no predicate-argument relation between the verbs.” (Haspelmath, 2016).

(9)

de	*no*	*dey*	*carry*	*me*	*travel*	*go*	*home*	*//*
they	neg	aux.imp	take	me	travel	go	home	//

‘They did not travel home with me.’ [P_ABJ_GWA_03_11]

We used the subtyped relation **compound:svc** for these constructions. (*carry* → *travel*; *travel* → *go* in sentence (9)



Conclusion

Ongoing work

- Development of a 500k syntactically annotated corpus of spoken Naija
    - Elaboration of a SUD native annotation scheme
    - Conversion of the resulting SUD treebank into UD
    - Error mining and consistency checking using the Grew querying tool
    - Merging the annotation and querying tools to facilitate error-mining

- End of NaijaSynCor project : March 2021.

- **Spin-offs of the corpus**
  - **Dictionary**. Francis Egbokhare has revived an old ongoing project of a Naija dictionary
  - **Grammar**: A collaborative online Encyclopaedic Grammar of Naija
  - **Orthography** : An online simplified version of the Naija text of corpus, establishing a unified orthography of the language
  - Extending the **(multilingual, corpus-based) methodology** to less documented African languages

WE TANK UNA WELL-WELL