A Surface-Syntactic UD Treebank for Naija
B. Caron, M. Courtin, K. Gerdes, S. Kahane

SyntaxFest 2019
Paris, August 26-30 2019
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
• **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 Independence (1960) expanded to most of Nigeria where it is learnt as a 2\textsuperscript{nd} 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

<table>
<thead>
<tr>
<th>Gold</th>
<th>Silver</th>
<th>Deuber (2005)</th>
</tr>
</thead>
<tbody>
<tr>
<td>150k</td>
<td>350k</td>
<td>250k</td>
</tr>
</tbody>
</table>

Current gold (125k):

- Download at [https://github.com/surfacesyntacticud/SUD_Naija-NSC](https://github.com/surfacesyntacticud/SUD_Naija-NSC)
- Query on [http://match.grew.fr/?corpus=SUD_Naija-NSC@dev](http://match.grew.fr/?corpus=SUD_Naija-NSC@dev)
2.1 Corpus

Data collection

Sampling  Recording  Collecting metadata

Editing

Time alignment  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 few first sample texts were 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 languages:
  - 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 {c 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**

<table>
<thead>
<tr>
<th></th>
<th>Percentage of agreement</th>
<th>Percentage of agreement when the annotation differs from the pre-parsed annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A/B⁶</td>
<td>A/C</td>
</tr>
<tr>
<td>UPOS</td>
<td>95</td>
<td>94</td>
</tr>
<tr>
<td>UAS</td>
<td>93</td>
<td>91</td>
</tr>
<tr>
<td>LAS</td>
<td>89</td>
<td>86</td>
</tr>
</tbody>
</table>

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.’

<table>
<thead>
<tr>
<th>Type</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>wey-cleft</td>
<td>na weekend <strong>wey</strong> we dey do am</td>
</tr>
<tr>
<td>bare cleft</td>
<td>na weekend Ø we dey do am</td>
</tr>
<tr>
<td>zero-copula cleft</td>
<td>Ø weekend Ø we dey do am</td>
</tr>
<tr>
<td>double cleft</td>
<td>na weekend <strong>na im</strong> we dey do am</td>
</tr>
</tbody>
</table>
The emergence of double-clefts in Naija

<table>
<thead>
<tr>
<th></th>
<th>Nigeria Pidgin*</th>
<th>Naija</th>
</tr>
</thead>
<tbody>
<tr>
<td>wey-clefts</td>
<td>41%</td>
<td>1%</td>
</tr>
<tr>
<td>bare clefts</td>
<td>39%</td>
<td>89%</td>
</tr>
<tr>
<td>zero-copula clefts</td>
<td>17%</td>
<td>1%</td>
</tr>
<tr>
<td>double clefts</td>
<td>—</td>
<td>9%</td>
</tr>
</tbody>
</table>

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)  
\[
\begin{array}{ccccccc}
\text{na} & \text{who} & >+ & \text{go} & \text{talk} & ?// & \text{who} \\
\text{COP} & \text{who} & & \text{AUX.FUT} & \text{talk} & ?// & \text{who} \\
\text{COP} & & & \text{AUX.FUT} & \text{talk} & & \text{help} \\
\end{array}
\]

‘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)  
\text{de} \quad \text{no} \quad \text{dey} \quad \text{carry} \quad \text{me} \quad \text{travel} \quad \text{go} \quad \text{home} \quad // \\
\text{they} \quad \text{neg} \quad \text{aux.imp} \quad \text{take} \quad \text{me} \quad \text{travel} \quad \text{go} \quad \text{home} \quad //

‘They did not travel home with me.’ [P_ABJ_GWA_03_11]
We used the subtyped relation \texttt{compound:svc} for these constructions. 
\textit{(carry $\rightarrow$ travel; travel $\rightarrow$ go) in sentence (9)}
Conclusion

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

• End of NaijaSynCor project : March 2021.
• Spin-offs of the corpus
  
  o **Dictionary.** Francis Egbokhare has revived an old ongoing project of a Naija dictionary
  
  o **Grammar:** A collaborative online Encyclopaedic Grammar of Naija
  
  o **Orthography:** An online simplified version of the Naija text of corpus, establishing a unified orthography of the language
  
  o Extending the (multilingual, corpus-based) methodology to less documented African languages
WE TANK UNA WELL-WELL