Establishing a Language by Annotating a Corpus

Marine Courtin, Bernard Caron, Kim Gerdes, Sylvain Kahane

Establishing a Language by Annotating a Corpus: the Case of Naija, a Post-creole Spoken in Nigeria

Marine Courtin[1], Bernard Caron[2], Kim Gerdes[3], Sylvain Kahane[1]
[1]Modyco, Université Paris Nanterre & CNRS
[2]Llacan, CNRS / IFRA Ibadan, CNRS
[3]LPP, Université Sorbonne Nouvelle & CNRS
marine.courtin@sorbonne-nouvelle.fr, bernard.caron@cnrs.fr,
kim@gerdes.fr, skahane@parisnanterre.fr

Abstract

In this paper, we show that building a treebank can be used as a way to establish a language. Annotated corpus can be used as tools when arguing that some linguistic data belongs to a separate language (rather than a dialect or variety of another established language). We provide here a case study on a treebank of Naija, a Post-creole spoken in Nigeria which presents us with significant differences from treebanks of English in terms of existing constructions and frequency of several syntactic units.

Keywords: Naija, Nigerian Pidgin, Treebank, Quantitative Linguistics, Typology

1. The Situation of Naija

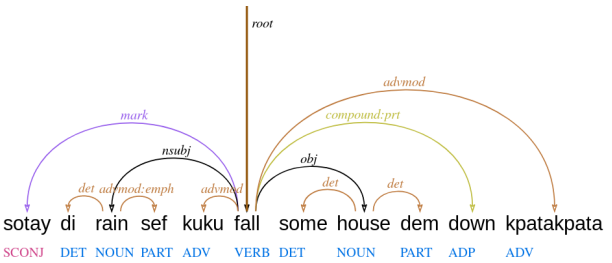
Spoken by educated Nigerians, the Nigerian post-creole has been shown by Deuber (2005) to develop in Lagos as a discrete language, separate from Nigerian English. This language, that we propose to call Naija, is now spoken as a second language by over 100 million speakers, all over Nigeria, a country of 180 millions people, where about 450 native languages are spoken with three dominating languages (Igbo, Yoruba, and Hausa). This new language has taken a considerable economical and cultural importance in Nigeria. Nevertheless, for its speakers, this language is often considered as an inferior version of English (they call it "Broken") with a negative influence on Nigerian education. Most speakers are not conscious that, as a separate language with its own grammar and lexicon, it has a outstanding potential in favor of national cohesion, since it is perceived as ethnically neutral, and for regional integration, due to its intercomprehension with Ghanaian and Cameroonian pidgins.

Considering the particular situation of this language, building a syntactic treebank takes a particular significance. Of course, as for any language, a treebank can be useful for many applications, such as the training of a syntactic parser. But here the treebank helps us to establish the existence of Naija as a language separate from (Nigerian) English, by showing constructions that are specific to Naija (qualitative analysis) and constructions that are over-represented in Naija (quantitative analysis).

2. Tools and Workflow

The study is based on a 750,000 word corpus collected all around Nigeria. The transcription is a scientific and political challenge by itself because most words stem from English, but some of them have grammaticalized and are pronounced differently. We follow what is done in the (mostly informal) writing of Naija: keep the English spelling for lexical words, with exceptions for very frequent words such as *broda* 'brother'; and a more phonetic spelling for grammatical terms (*dem* 'them', *im* 'him', *sey* complementizer lit. say).

Naija also borrowed lexical items from other local languages, in particular ideophones such as *kpatakpata* 'completely'.



(1) *sotay di rain sef kuku fall some house dem down kpatakpata*
so_that the rain EMPH commonly fall some house PL down completely
'So that, often, the rain completely destroys houses.'

We use the Arborator (Gerdes 2013) as the online annotation tool for POS and dependency annotation. The Arborator's exercise mode allows to present pre-annotated sentences as exercises to newly recruited annotators. The Arborator integrates the Mate parser (Bohnet 2010) that can be trained at any time which allows for quick and easy bootstrapping of the annotation process.

In order to allow for typological comparison and distance measures on Naija, we use a surface-syntactic dependency annotation scheme that is compliant with standard dependency annotation (e.g. prepositions as governors) and thus easy to learn and to apply, but which allows for a lossless transformation into Universal Dependencies (UD) by means of a graph rewriting process (Guillaume 2012). Each treebank for the 75 languages of the UD database must conform to the universal tagset for POS and dependency relation names. Language idiosyncrasies have to be encoded as additional features next to the POS or as subtypes of dependency relation names, e.g. in English the noun modifier (nmod) receives a subtype to describe the Saxon genitive: "*John*[*'s*] <-nmod:poss- *book*".

Currently the treebank has 12,000 tokens and is available on the UD webpage. We intend to manually annotate

100,000 tokens and then to automatically parse the whole corpus.

3. Qualitative Analysis

A good number of morphosyntactic specificities of Naija have called for an ongoing review of the annotating scheme that was initially adopted for the language.

Some of these specificities are linked to the influence of adstrate vernacular languages belonging mainly to the Niger-Congo family. This is the case of emphatic adverbial particles (e.g. *sha*, *o*) tagged with the ADV POS label, but whose function is characterized by the mod:emph dependency link. The influence of adstrate vernacular languages 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) Such constructions appear in languages of Nigeria, such as Yoruba (Stahlke 1970) (see (2)), and it has already been shown that they are present in creoles languages.

(2) *mo mú ìwé wá ilé* (Yoruba, Aubry 2010)
1SG take book come home'
'I brought a book to my home'

We used the subtyped relation compound:svc for these constructions, which do not exist in English (see (3)).

Other specificities are linked to the emergence of up to here undescribed structures which the corpus has enabled us to identify. One of them is a focus structure where the focus particle *na* (which identifies the clefted constituent) is doubled by the morpheme *naim* (which introduces the cleft clause). This morpheme originates in the grammaticalization of the colocation *na* + *im*, lit. 'it is' + 'him/it/her'. This discovery of a new structure is the result of a collaborative analysis done by the team of annotators during the production of the corpus.

The same ongoing grammaticalization process is observed in the formation of TAM auxiliaries where full lexical verbs (e.g. *go* 'go'; *come* 'come' ; *dey* 'exist') coexist with their grammaticalized equivalents (*go*, future; *come*, realis; *dey*, imperfective). Likewise, the verb *make*, which already appears in Serial Verb Constructions to express the equivalent of the comitative case, is used as an auxiliary for converb forms (e.g. *dem want make e go church* 'they want him to go to church'). This flourishing multifunctionality, typical of creole languages, creates challenges for the recognition of government.

4. Quantitative analysis

In creoles, it is usually assumed that there is a division of labor between the lexifier language which provides the majority of the lexicon (in our case English) and substrate languages in areal contact with the creole (in the case of Naija these might be Yoruba, Igbo and Hausa for example). We attempt to show quantitative evidence of structural similarities and differences between Naija and English.

One of our hypothesis concerning these differences is that information packaging (or communicative structure) plays a larger role in Naija than in English. To explore this hypothesis it is necessary that we dispose of an annotated corpus, as we need to measure the frequency of some structures (for example dislocations and cleft sentences), rather than their strict presence or absence in the language. For this purpose, we use all available treebanks of English in UD v2.1: UD_English-ParTUT (Bosco and Sanguinetti, 2014), UD_English-LinES (Ahrenberg 2007), UD_English-EWT (Silveira et al., 2014), and v2.2 version of UD_Naija-NSC. We also parsed the Santa Barbara Corpus of Spoken American English (Du Bois et al. 2000-2005) to get a reference of what spoken English might look like in terms of syntactic relations' distribution.

The table below presents some of the interesting differences between (1) written English, (2) spoken English and (3) spoken Naija :

	det	**case**	**obl**	**dislo-cated**	**ccomp**	**aux**	**cc**
(1)	9.4 %	10.6 %	5.8 %	0.0 %	1.1%	4.2 %	3.7 %
(2)	6.7 %	6.6 %	4.2 %	?	2.0 %	4.5%	4.3 %
(3)	5.7 %	4.2 %	3.7 %	1.7 %	2.1 %	9.3 %	1.4 %

To test the significance of the observed differences in frequency counts, we applied a Fisher's Exact Test for Count Data with simulated p-value (based on 2000 replicates), giving us an overall p-value of 0.0004998.

Some differences such as the lower frequency of determiners are easily explained. A Naija sentence such as *no dey stay for middle of road* would not require definite determiners in front of *middle* or *road*, while its English counterpart, *don't stay in the middle of the road*, would.

Another variation concerns the frequency of auxiliaries, which are more than twice as frequent in Naija than in English, regardless of the distinction written/spoken. We then looked at the ratio of verb on auxiliaries to see which language had more complex verbal constructions and found that Naija had the highest score (which means less auxiliaries per verb on average).

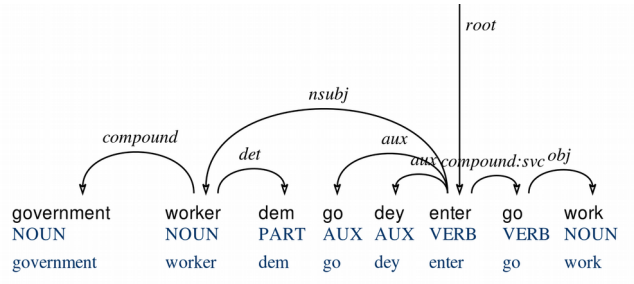
	Verb / Auxiliaries ratio
(1)	1.9
(2)	1.8
(3)	2.0

Taking into account the fact that Naija also has the highest frequency of auxiliaries (9.3% against 4.2% for written English and 4.6% for spoken English) we observe that Naija must compensate by having a high frequency of verbs which can be accounted for by the compound:svc, ccomp, acl:relcl and root relations. If we look more precisely at the distribution of these auxiliaries, it appears that it is the auxiliaries which are not shared with English

(*dey*, *come*, *go*, *don*, *fit*, *for* and *neva*) which are more frequent, while there is only one occurrence of the shared auxiliary *will*.

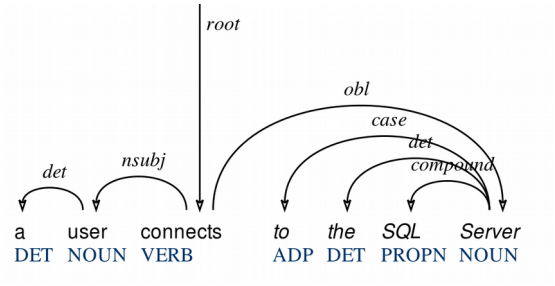
The lower frequencies for both oblique and case relations are correlated: Naija seems to use less oblique complement in favor of more direct objects. Locative complements can be expressed through Serial Verb Constructions with the place as direct object of the second verb as in (3).

(3) *government worker dem go dey enter go work*
government worker PL FUT PROG get_on go work
'government workers will be getting on to go to work'



This role would be filled by an oblique complement introduced by an adposition in English, as in the example below:

(4)



Other differences do not show such clear-cut contrasts between English and Naija, but are still interesting as they indicate areas which might need to be investigated further. We measure that 1.7 % of all dependency relations[1] in the Naija treebank are labeled dislocated. The mean length of sentences being around 10 tokens, this means that on average there is a dislocation in 1 sentence out of 6, which is very significant, even more so when compared to the 0.0004% frequency found in written English.

Unfortunately our parser performs poorly on this relation (due to the lack of training data) and no reliable frequency count of this relation type can be extracted from the spoken English corpus. We therefore look at spoken French (which has the reputation of being particularly prone to dislocations) to get a better sense of the significance of our findings, and find that 1.0 % of dependency links are dislocated (in the UD_French_Spoken, Lacheret and al., 2014). This indicates that dislocation is a major feature of spoken Naija. However, the variation in frequency of this dislocated link is not significantly more important between written English and spoken Naija than it is between written and spoken French, which seems to suggest that this might very well be a product of the genre rather than a characteristic of the language.[2]

This over-representation seems to apply to cleft sentences as well. The subtype :cleft, which we used in the annotation of both UD_Naija and UD_French_Spoken, can be found on 1.1 % of all relations in Naija, while it is considerably less frequent in spoken French (0.2%).

Another interesting findings is that Naija also shows three times less coordinating conjunctions than English does (1.4% for Naija against 3.7% and 4.3% for written and spoken English). This is interesting as we would expect a higher frequency of coordinations in spoken texts, to accommodate for lists and reformulations which are more common. In Naija it is not uncommon to have several coordinations without any coordinating conjunction as in (5) [conjuncts are underlined].

(5) *Lagos don follow see dis kind rain o wey <u>uproot tree take am block road spoil dose big billboard dem</u> […] <u>comot di roof of plenty house dem</u>.*
'Lagos has experienced the kind of rain where trees were uprooted and blocked the road, destroyed those big billboards […] and removed the roofing of lots of houses.'

This suggests that Naija might favor other strategies such as juxtaposition rather than coordinated constituents linked with coordinating conjunctions.

We might also be interested in the differences in distribution of part-of-speech tags[3] between English and Naija.

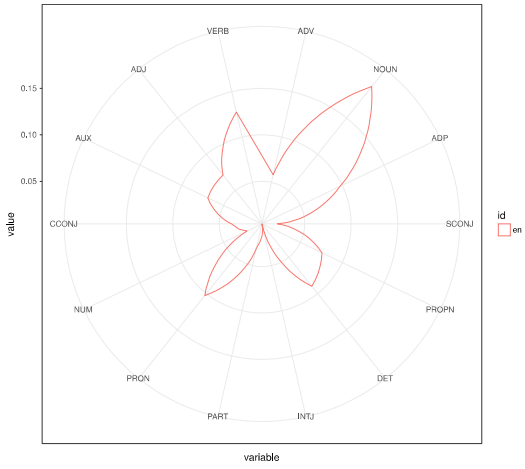


Fig 1. Relative frequency of pos tags in English

[1] *punct* links excepted

[2] One reviewer also noted that some of the English corpora such as EWT were automatically converted from constituent treebanks using rule-based systems which often fail to identify dislocated constructions.

[3] We filtered tokens with PUNCT, X and SYM tags

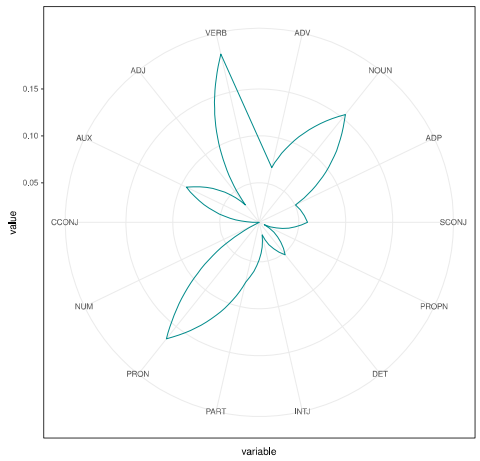


Fig 2. Relative frequency of pos tags in Naija

Naija has significantly more verbs while the English corpus is a lot richer in nouns. Part of the over-representation of verbs in Naija can be attributed to Serial Verb Constructions, with verbs in the second position representing 1.48 % of all tokens, but this account does not suffice to explain such a gap. Investigating this disparity, we also measured other relations involving verbal dependents such as *ccomp*. We find twice as many clausal complements with respectively 1.64 % and 0.82 % *ccomp* links in Naija and English. This indicates that looking at complex sentences in more details might provide us with additional examples of differences between the two languages.

We also expect that genre differences[4] between the treebanks play an important part in this repartition. Future work using a Nigerian English corpus of both spoken and written texts should allow us to better determine the extent of differences due to genre and the variety of English being considered.

Interestingly enough, even though Naija allows the dropping of pronouns they are still very frequent in our corpus. One possible explanation is that pronouns are highly susceptible to repetition and reformulation in spoken language. But it might also have to do with the frequent topicalization of subjects through dislocation in Naija, as in (6), or with rhetorical devices which involve repeating the pronoun to emphasize parallelism as in (7).

(6) dat man im pull over
that man he pulls over
'that man pulls over'

(7) dem go bring am dem go seize am again.
they will bring it they will seize it again
'they will bring it and seize it again'

[4] There is a small portion of spoken English in UD_English-LinES, but apart from this the corpus we used is all written texts, with variations in terms of genres (news, wiki, nonfiction, blog, emails, legal texts..). The Naija treebank is all spoken texts (conversations and interviews).

5. Conclusion

Annotators who were speakers of Naija reported that throughout the annotation process, their vision of Naija had changed. They noticed more readily that some syntactic phenomena were specific to Naija and that there were complex rules which governed the Naija grammar. We believe this to be an interesting pedagogical experiment where student annotators re-discover their language through the annotation of a corpus, and are confronted with regularities and patterns that sometimes went unnoticed in their day to day life (particularly so since speaking Naija is mostly depreciated).

We think that claims of Naija being a separate language can better be supported using a treebank. Indeed, while lexical differences are certainly noticeable between Naija and English, we believe that the identity of the language lies in its syntactic structure which is not as easily accessible from raw text or even tagged corpus. Having a treebank of Naija enables us to quantify the frequency of some syntactic structures, which in turns helps us to evaluate the complexity and idiosyncracies of the Naija grammar, and to measure the distance the language has taken from English. Comparisons between the two languages could also yield interesting insights concerning the ungoing creolization process of Naija.

Acknowledgments

We thank our reviewers for valuable remarks and corrections. This work is supported by the French National Research Agency (ANR) with the project NaijaSynCor

References

Ahrenberg, L. (2007). "LinES: An English-Swedish Parallel Treebank". Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA, 2007).

Aubry, N. (2010) Changements syntaxiques dans le Yorùbá de la presse (1930-2010) : traitement automatique d'un corpus diachronique et analyse des résultats, PhD thesis, Inalco.

Bohnet, B. (2010). "Very high accuracy and fast dependency parsing is not a contradiction." Proceedings of the 23rd international conference on computational linguistics. Association for Computational Linguistics.

Bosco, C. and Sanguinetti, M. (2014). "Towards a Universal Stanford Dependencies parallel treebank". In Proceedings of the 13th Workshop on Treebanks and Linguistic Theories (TLT-13), Tubingen (Germany).

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

Du Bois, John W., Wallace L. Chafe, Charles Meyer, Sandra A. Thompson, Robert Englebretson, and Nii Martey. (2000-2005). *Santa Barbara corpus of spoken American English, Parts 1-4*. Philadelphia: Linguistic Data Consortium.

Gerdes, K. (2013). "Collaborative dependency annotation." Proceedings of the second international conference on dependency linguistics (DepLing 2013).

Guillaume, B., Bonfante, G., Masson, P., Morey, M. and Perrier, G. (2012). "Grew: un outil de réécriture de graphes pour le TAL (Grew: a Graph Rewriting Tool for NLP)[in French]." Proceedings of JEP-TALN-RECITAL.

Haspelmath, M. (2016). The serial verb construction: Comparative concept and cross-linguistic generalizations. Language and Linguistics, 17(3), 291-319.

Jansen, B., Koopman, H., Muysken, P. (1978). Serial verbs in the creole languages. Amsterdam Creole Studies 2. 125–159.

Lacheret, A., Kahane, S., Beliao, J., Dister, A., Gerdes, K., Goldman, J. P., Tchobanov, A. (2014). Rhapsodie: a prosodic-syntactic treebank for spoken french. In Language Resources and Evaluation Conference.

Silveira, N., Dozat, T., de Marneffe, M., Bowman, S., Connor, M., Bauer, J., and Manning, C. (2014)."A Gold Standard Dependency Corpus for English." LREC.