

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

author1, author2, author3

Affiliation1, Affiliation2, Affiliation3

Address1, Address2, Address3

author1@zzz.edu, author2@zzz.edu, author3@zzz.edu

{author1, author5, author9}@abc.org

Abstract

In this paper, we show that building a treebank can be used as a way to establish a language. Annotated corpora 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 diverges significantly from treebanks of English in terms of prevalent constructions and frequency of several syntactic units.

Keywords: Naija, Nigerian Pidgin, Treebank

1. The Situation of Naija

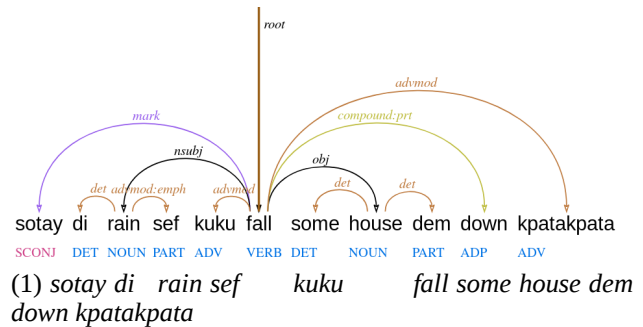
Spoken by educated Nigerians, the Nigerian post-creole has been proven by Deuber (2005) to develop in Lagos as a discrete language, separate from Nigerian English. This language, that we propose to call Naija, is now spoken as a second language by over 100 million speakers, all over Nigeria, a country 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 economic 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 serious 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 strategic 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).



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 presenting 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 characteristics of Naija have called for an ongoing review of the annotating scheme that was initially adopted for the language.

Some of these specific characteristics 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. As this construction does not exist in English, a new subtype had to be introduced to describe it: compound:svc.

Other specific characteristics 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 collocation *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 coverb 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 measure that 1.67 % of all dependency relations¹ 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 0.35% frequency found in English. We also 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 % of dependency links are dislocated (in the UD_French_Spoken, Lacheret et al., 2014). This indicates that dislocation is a major feature of spoken Naija, far more frequent than in English.

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

We might also be interested in the differences in distribution of part-of-speech tags² between English and Naija.



Fig 1. Relative frequency of pos tags in English

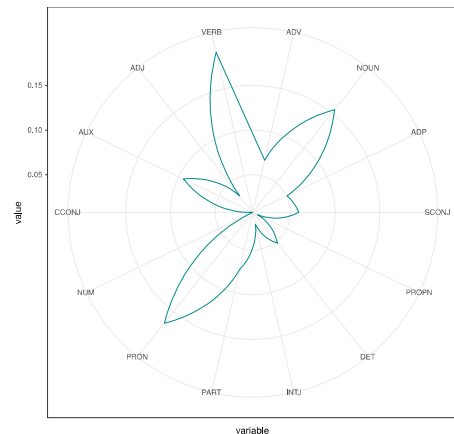


Fig 2. Relative frequency of pos tags in Naija

¹ Other than *punct* links.

² We filtered tokens with PUNCT, X and SYM tags

Naija has a lot 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 it is not enough to explain such a gap. Investigating this gap, we also measure other relations involving verbal dependents such as *ccomp*. We find twice as much 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 differences³ between the treebanks play an important part in this distribution. 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 *dat man im pull over* lit. that man he pulls over, ‘that man pulls over’, or with rhetorical devices which involve repeating the pronoun to emphasize parallelism as in *dem go bring am dem go seize am again* lit. they will bring it they will seize it again, ‘hey will bring it and seize it again’.

5. Conclusion

Annotators who were speakers of Naija reported that throughout the annotation process, their perception 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 to 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 by means of a treebank: 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 corpora. 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 creolization process.

6. Bibliographical References

- Ahrenberg, L. (2007). "LinES: An English-Swedish Parallel Treebank". Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA, 2007).
- 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), Tübingen (Germany).
- 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.
- Lacheret et al. (2014). "Rhapsodie: a prosodic-syntactic treebank for spoken french". LREC.
- 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.

³ 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 only consists of spoken texts (conversations and interviews).