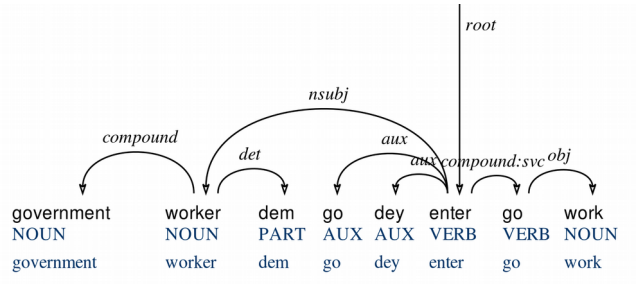
(*dey*, *come*, *go*, *don*, *fit*, *for* and *neva*) which are more frequent, while there is only one occurrence of the shared auxiliary *will*.

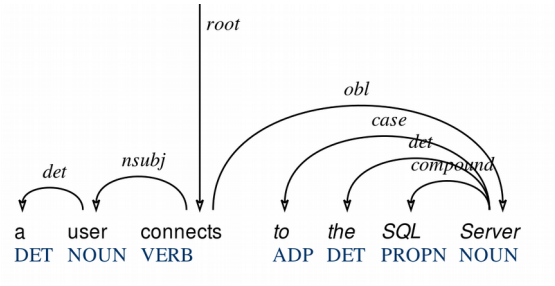
The lower frequencies for both oblique and case relations are correlated: Naija seems to use less oblique complement in favor of more direct objects. Locative complements can be expressed through Serial Verb Constructions with the place as direct object of the second verb as in (3).

(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</u> <u>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.

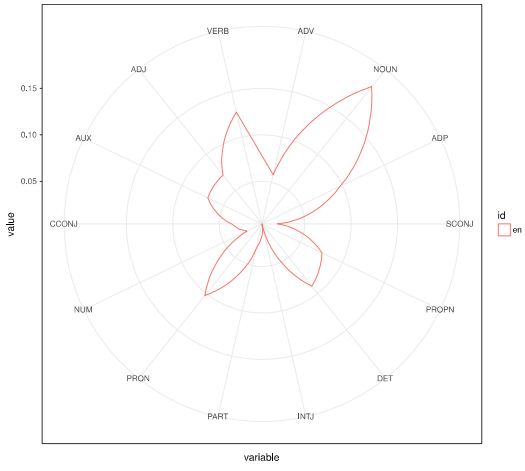


Fig 1. Relative frequency of pos tags in English

[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

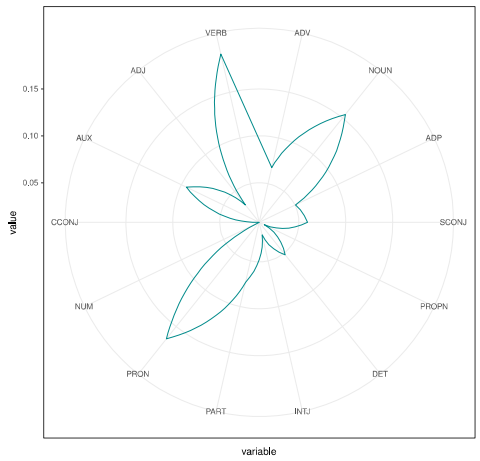


Fig 2. Relative frequency of pos tags in Naija

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.