The first Haitian Creole treebank

Sylvain Kahane (Modyco, Paris Nanterre University & CNRS / IUF)
Claudel Pierre-Louis (University of Orléans)
Sandra Jagodzińska (INALCO, Paris)
Agata Savary (LISN, Paris Saclay University & CNRS)

This article presents the first Haitian Creole treebank. It is a dependency treebank annotated in SUD and distributed in UD too. The treebank currently distributed contains 144 sentences and 3278 tokens, with the aim of reaching 10,000 words. We plan to add MWE annotation according to the PARSEME scheme.

Haitian is a creole with a French lexical base. Like most creoles, it is an isolating language. Nouns, verbs and adjectives do not inflect and have no morphological features. Determiners have values for number and definiteness, thus the number of a noun can be explicitly marked via its determiner. Pronouns have the value for number and person.

Syntax remains fairly close to French (SVO order). The main difference is a system of preverbal particles for TAM (Tense, Aspect, Modal). Haitian has its own phonetically-based orthography, far removed from that of French, which prevented the direct use of a French parser to pre-analyze the corpus.

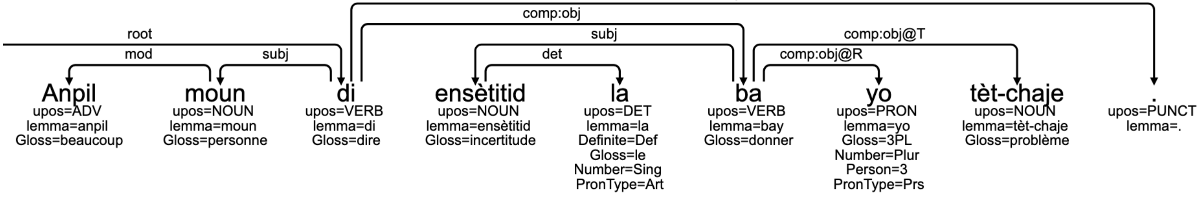
The corpus was annotated according to the SUD annotation scheme (Surface-Syntactic Universal Dependencies, Gerdes et al. 2018, https://surfacesyntacticud.github.io), which is a dependency-based annotation scheme, like UD (Universal Dependencies, de Marneffe et al. 2021, https://universaldependencies.org), but based on distributional criteria, which favor functional heads unlike UD. The treebank was then automatically converted to UD and was first distributed in the UD2.13 release of Nov. 2023. (The analyses presented here differs from this first version and are available on https://universal.grew.fr/?corpus=SUD_Haitian_Creole-Autogramm@latest#.)Annotation was carried out with ArboratorGrew (Guibon et al. 2020; https://arboratorgrew.elizia.net): the tool enables manual annotation, but also trains a parser on the first analyses to pre-annotate the rest of the corpus. It also enables conversion rules to be applied to modify an analysis, search for inconsistencies and harmonize annotation.

The corpus is made up of four types of text: the Creole Bible, a novel entitled *Lanmou titato* (Roy 2021), press articles from the VOA kreyol newspaper and online texts such as the Plateforme Haïtienne de Plaidoyer pour un Développement Alternatif (https://www.papda.org). Over 2,000 words were selected from each text sample, giving us a corpus of over 10,000 words. The corpus is glossed and translated in French.

The treebank is searchable on Grew-Match (Guillaume 2021) in UD and SUD formats (https://universal.grew.fr/?corpus=SUD_Haitian_Creole-Autogramm@latest).

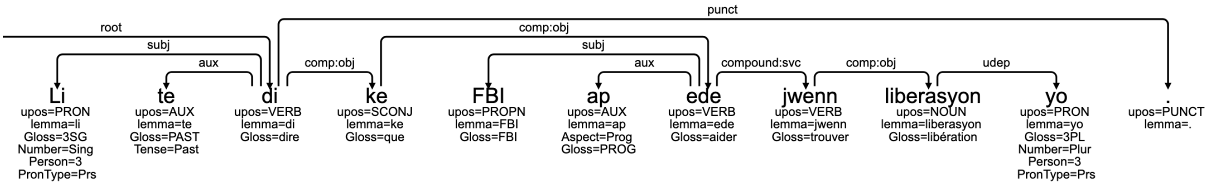
Let us look at three examples of annotated sentences illustrating particular constructions of Haitian.

(1) *Anpil moun di ensètitid la ba yo tèt-chaje.*
‘Many people say that uncertainty is a problem for them.’, lit. gives them head-change



In this sentence, the main verb *di* ‘say’ is the root. The verb ba(y) ‘give’ is the head of the subordinated clause, which is the object complement (comp:obj) of *di*. It subcategorizes a double object construction (which does not exist in French). We have distinguished the Recipient (comp:obj@R) from the Theme (comp:obj@T). The latter forms, together with the verb, a light verb construction of type LVC.full according to the PARSEME typology (https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.3/, Savary et al. 2017). Annotation of such multiword expressions is planned in further steps.

(2) *Li te di ke FBI ap ede jwenn liberasyon yo*.
‘He said the FBI would help with their release.’, lit. help find their release

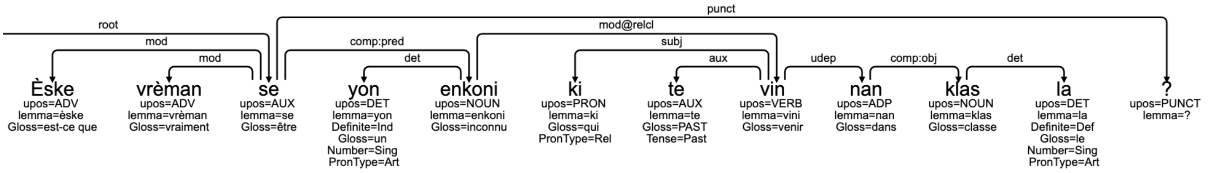


In sentence (2), unlike (1), the subordinate object of *di* ‘say’ is introduced by a subordinating conjunction. We have considered *ede jwenn* ‘help find’ to be a serial verb construction (Glaude 2013). We also note two occurrences of TAM: *te*, which marks the past tense, and *ap*, which we have annotated here as a progressive (Lainy 2010), but which should perhaps be seen as an imperfective (Glaude 2013). TAM markers are, according to the usual analysis in UD, tagged AUX. Four other particles have been considered: the futur *pral* ‘will’, the conditional *ta* ‘would’, and the modal *ka* ‘can’ and *dwe* ‘must’. There is justification in Haitian for distinguishing them from verbs, from which they differ in various properties, including the impossibility of being topicalized:

(3) a. *Jan ka veni.* ‘John can come.’
b. *Se vini Jan ka veni.* ‘John CAN come.’
c. **Se ka Jan ka venir.*

Haitian has noun complements build by simple juxtaposition of the NP on the left of head noun. The same construction is used personal pronouns. For this reason, *liberasyon yo* ‘their liberation’, lit. liberation they, is analyzed as “liberation of them” in (2). In SUD, the relation *udep* (for *underspecified dependent*), which subsumes both *comp* and *mod*, is used for noun complements.

(4) *Èske vrèman se yon enkoni ki te vin nan klas la ?*
‘Was it really a stranger who had come into the classroom?’



Sentence (4) shows the interrogative adverb *èske*, inherited from French *est-ce que*, as well as the copula *se*, inherited from French *c’est* ‘it is’. The relative pronouns are *ki* for subject, *ke* for object, and *kote* ‘where’. The relative pronoun *ke* is omitted in two-thirds of cases (47/69).

Note also that determiners are optional. Two-thirds of the nouns are without determiners (427/703). All the indefinite determiners precede the noun, but the definite determiners follow it: the singular and plural definite article *la* and *yo*, as well as the demonstrative *sa*. Adjectives precede or follow the noun, as in French.

The Grew-match tool makes it easy to retrieve the different constructions we have talked about and to have a nice view on the grammar of the language.

To conclude, we have trained a parser for Haitian using the parser training tool included in ArboratorGrew (Peng et al. 2022). We have fine-tuned a parser pretrained on all SUD treebanks fine-tuned on our 144 annotated sentences and we achieve a LAS of 72.5%, which gives us a reasonable tool to start the annotation of the rest of the corpus.

References

de Marneffe, Marie-Catherine, Christopher D. Manning, Joakim Nivre, & Daniel Zeman. 2021. "Universal Dependencies". *Computational Linguistics* 47 (2): 255 308. https://doi.org/10.1162/coli_a_00402.

Gerdes, Kim, Bruno Guillaume, Sylvain Kahane, & Guy Perrier. 2018. "SUD or Surface-Syntactic Universal Dependencies: An Annotation Scheme near-Isomorphic to UD". In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, 66-74. Brussels, Belgium: Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6008.

Glaude, H. (2013). *Aspects de la syntaxe de l'haïtien*. PhD thesis, Universiteit van Amsterdam, published by Editions Anibwé.

Guibon, Gaël, Marine Courtin, Kim Gerdes, & Bruno Guillaume. 2020. "When Collaborative Treebank Curation Meets Graph Grammars". In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 5291-5300. Marseille, France : European Language Resources Association. https://aclanthology.org/2020.lrec-1.651.

Guillaume, Bruno. "Graph Matching and Graph Rewriting: GREW tools for corpus exploration, maintenance and conversion." *Proceedings of the 16th Conference of the European Chapter (EACL): System Demonstrations*. 2021, 168-175.

Lainy, Rochambeau. 2010. "Temps et aspect dans la structure de l'énonciation rapportée : comparaison entre le français et le créole haïtien". PhD Thesis, Rouen. https://www.theses.fr/2010ROUEL005.

Peng, Z., K. Gerdes, & K. Guiller. 2022. "Pull your treebank up by its own bootstraps". In *Journées Jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain (LIFT) et Traitement Automatique des Langues (TAL)*, CNRS, 139-153.

Savary, A., C. Ramisch, S. R. Cordeiro, F. Sangati, V. Vincze, B. Qasemi Zadeh, M. Candito, F. Cap, V. Giouli, I. Stoyanova, & A. Doucet. 2017. "The PARSEME shared task on automatic identification of verbal multiword expressions". In *The 13th Workshop on Multiword Expression,* EACL, 31-47.

Ressources

Bib La an Kreyol [Haitian Creole Bible]. http://www.fouyebible.com/.

Roy, Méange Sophiana. 2021. *Lamou Titato*. Jacmel, Haïti http://www.papda.org/Savann-Dyann-se-pou-ti-peyizan-li-pa-ni-pou-vann-ni-pou-fe-kado