Outils pour l'exploration de l'interface prosodie-syntaxe en pidgin nigérian

Emmett Strickland
MoDyCo, Nanterre, France
emmett.strickland@parisnanterre.fr

MOTS-CLES : prosodie, intonosyntaxe, treebank, stylisation prosodique
KEYWORDS : prosody, intonosyntax, treebank, prosodic stylisation

1 Contexte

Le pidgin nigérian, ou Naijá, est un pidgin-créole parlé par environ 100 millions de locuteurs en Afrique de l'Ouest. Autrefois stigmatisé comme une langue parlée par les personnes sans instruction, le pidgin nigérian est aujourd'hui une lingua franca majeure utilisée par les quelques 200 groupes ethnolinguistiques du Nigeria. Bien que la majeure partie du vocabulaire de la langue provienne de l'anglais, elle a acquis un ensemble distinct de propriétés grammaticales, intonatives et tonales qui la distinguent de celui-ci. La typologie prosodique exacte du pidgin nigérian reste un sujet de débat, notamment en ce qui concerne le rôle du ton et de l'accent lexicaux. Dans cette communication, je présenterai une série d'outils développés dans ma thèse pour contribuer au débat sur la prosodie du pidgin nigérian à l'aide d'une approche basée sur le corpus. Je présenterai tout d'abord SLAM 3, un logiciel de stylisation prosodique qui permet l'étiquetage automatique de contours de F0. Je présenterai ensuite un corpus d'arbres syntaxiques sur lesquelles sont projetés des étiquettes prosodiques décrivant syllabe de chaque token.

2 SLAM 3

Mes recherches reposent en grande partie sur un modèle prosodique appelé SLAM (Liu et al., 2019). Toutes les itérations de SLAM prennent en entrée deux types de fichiers : un fichier son contenant des informations sur la hauteur, et un fichier .TextGrid qui segmente le fichier audio en unités linguistiques que le chercheur souhaite étudier (syntagmes, mots, syllabes...). Pour chaque unité, SLAM produit une étiquette textuelle décrivant les variations de hauteur associées à l'aide d'un alphabet de cinq tons élémentaires : H (très haut), h (haut), m (moyen), l (bas) et L (très bas). Trois informations de base sont encodées par le modèle SLAM : la hauteur du début du segment, la hauteur de la fin du segment, et la hauteur et la position relative du plus grand pic de F0 qui se produit entre les deux. Ainsi, lh décrit un contour montant, tandis que mmh3 décrit un contour qui commence et se termine par une hauteur moyenne, et contient un maximum aiguë dans le dernier tiers du segment.

Dans ma recherche, j'ai décidé d'utiliser la syllabe comme unité primaire d'analyse. Cependant, le fait d'axer ma recherche sur des unités de durée aussi courte comporte des spécificités que les versions précédentes du modèle SLAM n'ont pas pris en compte. Nous verrons, par exemple, pourquoi la perception ou non des changements de hauteur est fortement influencée par la durée de l'unité en question. Les versions précédentes de SLAM produisaient donc des étiquettes qui correspondaient mal aux contours de hauteur tels qu'ils étaient perçus. Ma présentation décrit les nouvelles fonctionnalités introduites pour tenir compte de ce problème et d'autres encore.

3 Un treebank annoté en prosodie

Ma recherche consiste à exploiter une ressource existante produite par le projet ANR NaijaSynCor : un vaste corpus de pidgin nigérian parlé qui a été transcrit et annoté sous forme d'arbres syntaxiques (Manfredi et al., 2021). Ces corpus, connus sous le nom de *treebanks*, facilitent l'analyse quantitative de la grammaire d'une langue en permettant aux utilisateurs d'identifier automatiquement la fréquence de certains mots, parties du discours et structures grammaticales. Ma thèse étend les champs d'application de ce treebank à ceux de la prosodie et de l'intonosyntaxe en appliquant à chaque token du corpus une série d'étiquettes décrivant la hauteur de chaque syllabe. Dans ma communication, je donnerai un aperçu de la structure de ce treebank et de la manière dont elle peut être appliquée à l'étude de la prosodie du pidgin nigérian. Je fournirai diverses démonstrations directement liées à ma thèse, comme l'utilisation du ton pour distinguer les éléments lexicaux tels que les verbes et les éléments grammaticaux tels que les auxiliaires. En outre, les perspectives d'amélioration future de ce treebank seront explorées.

Références

LIU L., LACHERET-DUJOUR A. & OBIN, N. (2019). Automatic modelling and labelling of speech prosody: What's new with SLAM. *International Congress of Phonetic Sciences (ICPhS)*.

MANFREDI S., CARON B., GERDES K. & COURTIN M. (2021), NaijaSynCor: a syntactic treebank, a parser and a wiktionary for Naija. *Summer Conference of the Society of Pidgin and Creole Linguistics*.