Université Paris Nanterre

Mémoire de Master 2 Traitement Automatique des Langues

Parcours Recherche et Développement

Interface entre prosodie et structure informationnelle

une approche statistique appliquée à un corpus de naija

Un mémoire de recherche rédigé par Emmett STRICKLAND sous la direction des Professeurs Anne LACHERET-DUJOUR et Sylvain KAHANE

Défendu en septembre 2021 et complété ultérieurement selon les commentaires du jury

Attestation de non-plagiat

Je, soussigné Emmett STRICKLAND, déclare avoir rédigé ce travail sans aides extérieures ni sources autres que celles qui sont citées. Toutes les utilisations de textes préexistants, publiés ou non, y compris en version électronique, sont signalées comme telles. Ce travail n'a été soumis à aucun autre jury d'examen sous une forme identique ou similaire, que ce soit en France ou à l'étranger, à l'université ou dans une autre institution, par moi-même ou par autrui.

Le 3 septembre 2021

Remerciements

Je tiens à remercier les membres du jury ainsi que mes encadrants, Anne Lacheret-Dujour et Sylvain Kahane, pour avoir supervisé mon travail et pour avoir partagé leur expertise respective en prosodie et en syntaxe pendant mon master. Je remercie également les membres actuels et anciens de l'équipe NaijaSynCor, notamment Kim Gerdes et les stagiaires Solveig Poder, Lila Kim et Mariam Nakhlé, dont la participation a contribué directement ou indirectement à la réalisation de ce mémoire. Je tiens également à saluer les chercheuses Candide Simard et Lucie Vercruyssen pour leur étroite collaboration avec Anne et moi afin d'assurer le succès de ce projet de recherche. Je remercie enfin le directeur du projet NaijaSynCor, Bernard Caron, pour m’avoir initié au monde de la linguistique africaine.

Résumé

Ce projet vise à explorer le rôle de la prosodie dans la structure informationnelle d'une langue d'Afrique de l'Ouest appelée naija en se concentrant sur le contenu de structures syntaxiques nommés prénoyaux. L'objectif de ce projet était de déterminer si différentes fonctions pragmatiques étaient associées ou non à des contours prosodiques distincts, ce qui suggérerait que les fonctions pragmatiques sont, dans une certaine mesure, marquées prosodiquement. Dans cette perspective, nous avons exploité un ensemble d'outils pour extraire des informations prosodiques d'un corpus de naija parlé et établir des corrélations avec un ensemble parallèle d'annotations pragmatiques. Cette approche nous a permis de démontrer ce qui semble être un lien clair entre prosodie et fonction pragmatique.

Table des matières

Figures et tableaux

1 Introduction et hypothèses

Ce mémoire résume un travail de recherche effectué dans le but de mieux comprendre la relation entre la prosodie et la structure informationnelle en appliquant une appliquant une approche statistique au naija, un pidgin-créole parlé par environ cent millions de personnes au Nigeria. Précisément, ce projet vise à utiliser des outils existants pour modéliser de façon automatique les caractéristiques prosodiques d'un ensemble de prénoyaux, une catégorie syntaxique située au début de certains énoncés, dans un corpus oral de 23 monologues. En parallèle, des annotateurs humains ont effectué une annotation pragmatique rigoureuse de chaque prénoyau utilisé en attribuant à chacun une ou plusieurs étiquettes spécifiant ses fonctions pragmatiques, telles que l'introduction d'un nouveau thème ou la localisation d'un événement dans le temps ou l'espace. Ensuite, nous avons effectué des analyses en composantes multiples (ACM) pour identifier des corrélations statistiques entre ces étiquettes pragmatiques et les différents types de contours prosodiques.

La question de recherche principale qui a conduit ce projet était de savoir si les différentes fonctions pragmatiques sont ou non démarquées les unes des autres sur le plan prosodique. Au cours de ce mémoire de recherche, nous tenterons de valider trois hypothèses :

1. qu'il existe des différences prosodiques statistiquement observables entre les différents types pragmatiques

2. qu'il est possible d'identifier les contours prosodiques qui sont le plus fortement associés à certains types pragmatiques, et

3. que les locuteurs sont plus susceptibles de marquer prosodiquement les topics qu'ils supposent moins accessibles dans la représentation mentale de leurs interlocuteurs.

Après cette tâche initiale, nous avons appliqué la même technique statistique à un ensemble de données modifié dans lequel chaque contour a été remplacé par un ensemble de caractéristiques de base telles que sa hauteur moyenne ou sa direction (c'est-à-dire, descendante, montante ou plate), dans le but de déterminer si certains types pragmatiques sont associés à certaines caractéristiques prosodiques fondamentales. Comme ce type d'analyse est largement basé sur les impressions, nous avons également appliqué un deuxième test statistique, le test exact de Fisher, au même ensemble de données. Notre objectif dans ce cas était de fournir un moyen plus objectif de quantifier la corrélation entre les variables différentes.

La suite de ce mémoire commencera par un contexte scientifique où l'on présentera le naija, ainsi que les différents concepts sur lesquels ce travail de recherche est basé. Ensuite, nous

décrirons la méthodologie utilisée pour préparer nos données et mener notre recherche. Dans la section suivante, nous présenterons les résultats de notre recherche et reviendrons sur nos hypothèses. Nous décrirons ensuite une série de tests appliqués postérieurement à notre ensemble de données et décrirons les observations supplémentaires effectuées. Enfin, nous conclurons en discutant des limites de notre recherche et en présentant différentes directions de recherche future.

2 Contexte scientifique

2.1 Le naija

Le naija est une langue parlée par environ 100 millions de personnes au Nigeria, soit la moitié de la population du pays. Les origines de la langue remontent aux avant-postes portugais, hollandais et britanniques établis le long de la Gold Coast africaine à partir du XVIIe siècle. Le contact entre les Européens et les peuples indigènes a entraîné le développement de divers jargons commerciaux, ou pidgins, utilisés exclusivement dans la communication entre ces groupes. Les XIXe et XXe siècles ont vu l'émergence d'un pidgin à base lexicale anglaise qui était utilisé par les peuples indigènes au-delà de la communication avec les Européens (Huber 1999:57). Cette langue peut être considérée comme l'ancêtre du naija.

Bien qu'il soit difficile de savoir dans quelle mesure les pidgins antérieurs ont influencé cette langue (Huber 1999:57), les traces du portugais sont rendues évidentes par un ensemble restreint de lexèmes communs tels que *pikin* 'enfant' et *sabi* 'savoir'. En plus des emprunts anglais qui constituent la majorité de son lexique, le naija compte également des mots empruntés à diverses langues régionales comme le haoussa et le yoruba. Le naija partage également un certain nombre de similitudes frappantes avec d'autres langues qui ont émergé pendant cette période de contact afro-européen, tels que le pidgin camerounais, le pidgin ghanéen, et le krio de Sierra Leone. Les caractéristiques grammaticales communes comprennent notamment la présence des copules *dey* et *na* ainsi que le pronom de la deuxième personne du pluriel *una* (Huber 1999:90-97).

Depuis les années 1960, le naija s'est rapidement répandu dans tout le Nigeria et est devenu la langue auxiliaire primaire du pays, bien que l'anglais reste la langue officielle de l'administration. Le naija a traditionnellement servi de langue véhiculaire pour faciliter la communication entre les quelque 500 groupes linguistiques du pays, mais est de plus en plus utilisé dans la plupart des contextes informels, y compris entre les membres d'un groupe linguistique commun. Au cours de la dernière décennie, la langue a bénéficié d'une présence

médiatique croissante. La station de radio populaire Wazobia FM propose de l'actualité, des débats et d'autres contenus exclusivement en naija depuis 2007, tandis que la *British Broadcasting Channel* (BBC) a lancé son propre service en naija en 2017 (BBC 2018). Ce dernier a rapidement atteint une audience hebdomadaire de 7,5 millions de personnes (Edionhon 2018).

Les 100 millions de locuteurs du naija le placent au rang du turc, du tamoul et du japonais comme l'une des 20 plus grandes langues du monde, le distinguant comme la plus grande langue émergente d'Afrique.[1] Et compte tenu de la forte croissance démographique du Nigeria, il est presque certain que l'influence de cette langue va s'accroître avec le temps. La taille et la portée interethnique de cette langue en font un candidat de choix pour l'application des outils de traitement du langage naturel et les études linguistiques en général, deux domaines sous-représentés sur le continent africain.

2.2 Le projet NaijaSynCor

La recherche contenue dans ce mémoire se base sur des travaux réalisés dans le cadre du projet NaijaSynCor, une initiative conjointe rassemblant les unités de recherche Langage, Langues et Cultures d'Afrique (LLACAN) et Modèles, Dynamiques, Corpus (MoDyCo). Piloté depuis 2017 par le chercheur Bernard Caron, le projet NaijaSynCor travaille à la mise en place d'un ensemble d'outils informatiques afin de faciliter l'étude de la structure syntaxique et prosodique du naija. Ce projet a notamment permis la création d'un corpus de 500 000 mots composé de 384 échantillons de langue orale enregistrés à travers le Nigeria (Bigi *et al.* 2018). Chaque enregistrement a été transcrit et segmenté en énoncés, qui ont été systématiquement traduits. Des alignements temporels ont également été calculés au niveau de chaque mot, ce qui signifie que chaque mot s'est vu attribuer une paire de valeurs numériques correspondant à son temps de début et de fin dans l'enregistrement correspondant. Un alignement syllabique encore plus précis a également été produit, mais il n'a pas été utilisé dans ce projet.

Ces informations ont été encodées dans un ensemble d'arbres syntaxiques produits pour chaque énoncé. Ces arbres contiennent diverses informations morphosyntaxiques, telles que la partie du discours de chaque mot, et la nature des relations syntaxiques qui les relient.[2] Un sous-ensemble de 88 de ces enregistrements (80 monologues et 8 dialogues) a été utilisé pour produire des arbres soigneusement annotés par des linguistes entraînés, et a été revu

[1] Suivi par le haoussa, une langue également parlée au Nigéria. La langue la plus parlée en Afrique reste néanmoins l'arabe, un lointain cousin du haoussa qui a été introduit sur le continent depuis l'Asie occidentale lors de l'expansion de l'Islam.

[2] Le terme **treebank** est désormais utilisé pour décrire les corpus contenant un ensemble d'arbres syntaxiques.

et corrigé au cours du projet. Ce sous-ensemble de données est appelé le corpus *Gold* dans le cadre du projet. Ce corpus de référence a ensuite été utilisé pour entraîner un analyseur syntaxique qui a généré automatiquement des arbres syntaxiques à partir des fichiers restants, un travail effectué dans le cadre d'un mémoire de recherche rédigé par Guiller (2020).

La création de ce treebank a pour but de contribuer à la description linguistique du naija en permettant aux chercheurs de mener diverses analyses quantitatives de ses propriétés lexicales, morphologiques et syntaxiques, ce qui peut aider à la formulation de grammaires descriptives (Song 2020, Courtin *et al*. 2018). Dans le cadre du présent mémoire de recherche, nous nous sommes basé exclusivement sur les données contenues dans le corpus *Gold*.

Un deuxième objectif du projet NaijaSynCor est d'étudier la structure prosodique du naija, i.e. les caractéristiques intonatives du naija parlé. Avant que le présent mémoire de recherche ne soit commencé, une ancienne stagiaire travaillant sur le projet, Solveig Poder, a préparé un script permettant d'exploiter les valeurs temporelles associées à chaque mot dans le treebank pour extraire les positions de certaines structures syntaxiques afin de faciliter l'étude de leurs propriétés intonatives. Ce script a permis de jeter les bases du présent projet de recherche.

2.3 La syntaxe

La syntaxe est traditionnellement définie comme l'étude des règles qui déterminent l'ordre des mots dans un énoncé. Dans le cadre de cette étude, la syntaxe est utilisée comme un terme général pour couvrir deux niveaux de description de la structure interne des énoncés du naija : la microsyntaxe et la macrosyntaxe.

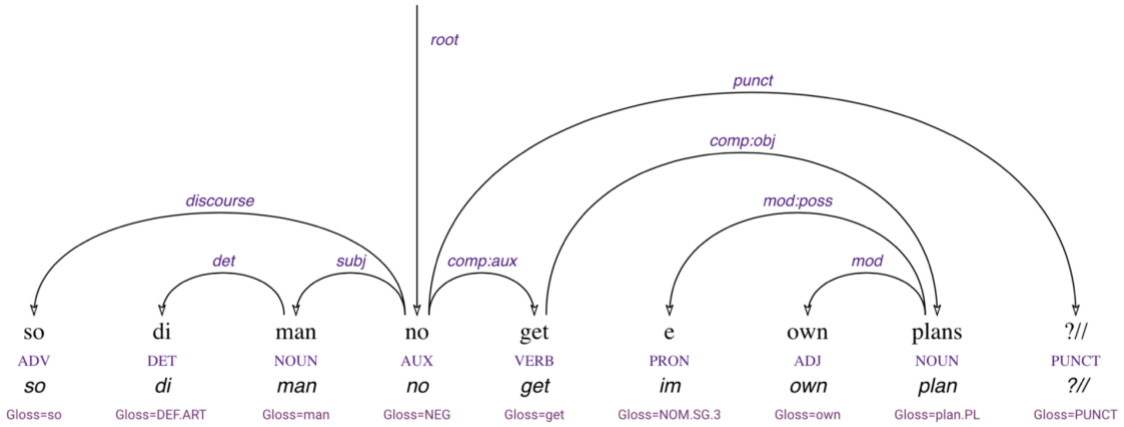
2.3.1 La microsyntaxe

Nous utilisons le terme **microsyntaxe** pour désigner les relations hiérarchiques entre les mots qui constituent un énoncé, et les règles qui les régissent. La microsyntaxe organise les mots d'un énoncé en arbres hiérarchiques qui modélisent leurs relations grammaticales avec les autres. Cette approche de l'étude de la grammaire a été fortement influencé par (Chomsky 1957), ouvrage fondateur dans l'histoire de la syntaxe et plus largement de la linguistique. Diverses approches de modélisation des relations microsyntaxiques se sont développées au fil des ans. Dans ce projet, nous employons une approche basée sur la grammaire de dépendance, une théorie basée largement sur (Tesnière 1959) qui conceptualise les relations entre les mots comme étant des relations de gouvernance et de dépendance entre des couples de mots. Les arguments verbaux tels que les sujets et les objets sont considérés

comme les dépendants des verbes. Les verbes eux-mêmes sont considérés comme les gouverneurs de ces arguments. Les dépendants typiques des noms comprennent leurs déterminants ainsi que les adjectifs qui les modifient. Ce projet utilise une norme d'annotation syntaxique connue sous le nom de *Surface-Syntactic Universal Dependencies* (SUD), décrite en détail dans Gerdes *et al*. (2018).

Dans ce système d'annotation, chaque énoncé est représenté sous la forme d'un arbre de dépendance, comme illustré dans la figure 1 ci-dessous.

Figure 1 - Un énoncé en naija sous forme d'arbre de dépendance



Chaque flèche représente une relation entre un gouverneur, situé à l'origine de la flèche, et son dépendant, situé au bout de la flèche. Notez qu'un mot peut avoir plusieurs dépendants, mais seulement un gouverneur. Lorsque le dépendant d'un gouverneur donné a également d'autres dépendants, on dit que ce gouverneur est la **tête** d'un syntagme composé de lui-même, de son dépendant, et de tout ce qui est inférieur dans la hiérarchie syntaxique. Ainsi, le verbe *get* dans cet exemple est considéré comme la tête de *get e own plans*, et *no* serait la tête de l'énoncé entier. Un mot qui est la tête d'un énoncé entier est également dit être la racine de l'énoncé.

Chaque relation de dépendance est également marquée par une étiquette décrivant la nature de la relation syntaxique. Par exemple, les objets directs sont reliés aux verbes par la relation `comp:obj`, tandis que les modificateurs sont reliés à leurs gouverneurs par la relation `mod`. Plusieurs de ces étiquettes sont particulièrement utiles pour l'annotation du discours oral. Par exemple, l'étiquette `discourse` est utilisée pour identifier les marqueurs de discours contenant un seul mot. Cette dernière étiquette est visible dans l'exemple ci-dessus sur la flèche reliant *no* au marqueur de discours *so*.

Le contenu de ces arbres de dépendance est stocké dans le format de fichier `.conllu`, qui fournit un moyen normalisé de représenter les arbres de dépendance. Un fichier `.conllu` est

un fichier texte brut dans lequel chaque énoncé est représenté par un bloc de lignes, chacune correspondant à un token. Chaque ligne est divisée en dix colonnes, délimitées par une tabulation, chacune d'entre elles contenant entre autres diverses informations morphologiques et syntaxiques sur le token correspondant. Nous ne décrirons pas la fonction de chaque colonne, mais les suivantes sont les plus fréquemment utilisées et les plus importantes pour la compréhension de ce projet.

- Colonne 1 : un nombre indiquant la position du token dans l'énoncé
- Colonne 2 : le mot-forme (le token tel qu'il est prononcé, avec inflection)
- Colonne 3 : le lemme (forme canonique du token, sans inflection)
- Colonne 4 : la partie de discours du token (nom, verbe…)
- Colonne 7 : la position du gouverneur du token dans l'énoncé (voir Colonne 1)
- Colonne 8 : le type de relation syntaxique avec le gouverneur (`comp:aux`, `mod`…)
- Colonne 10 : informations diverses, telles que des gloses et les alignements temporels.

Dans le cadre du projet NaijaSynCor, la colonne 10 est centrale pour l'étude de la prosodie. Dans cette colonne, chaque token est assigné à une paire de traits `AlignBegin=` et `AlignEnd=`, chacun d'entre eux étant suivi d'une valeur numérique. Ces traits fournissent des informations temporelles détaillant les positions dans le fichier sonore correspondant où le mot commence et se termine. Ces informations sont destinées à permettre aux chercheurs de générer automatiquement un alignement temporel entre certaines structures syntaxiques et leurs positions correspondantes dans le fichier sonore. Ce point est un élément central de ce travail, et sera revisité dans les sections suivantes.

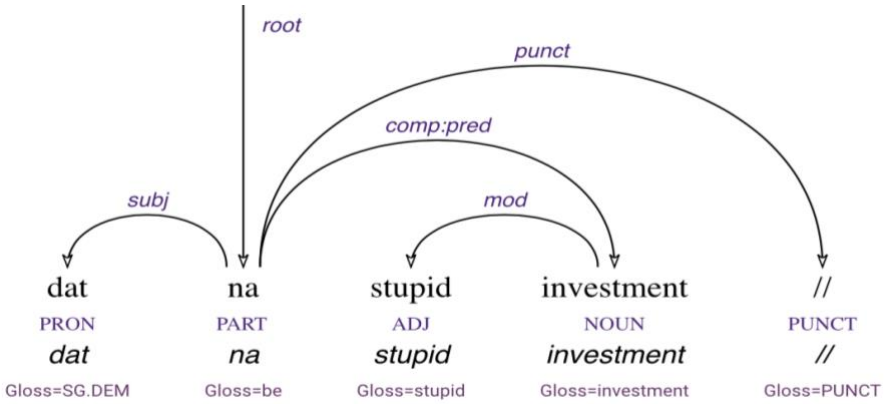
En plus des lignes décrivant les caractéristiques de chaque token, chaque énoncé dans un fichier `.conllu` est également précédé d'une série de lignes commençant par le symbole #. Ces lignes contiennent différents types de métadonnées. Les métadonnées utilisées dans le projet NaijaSynCor comprennent notamment un identifiant pour chaque énoncé, ainsi qu'une traduction anglaise.

Dans ce projet, chaque fichier `.conllu` correspond à un seul fichier sonore. Chaque énoncé présent dans le fichier sonore est représenté par un arbre correspondant dans le fichier `.conllu`. Les blocs correspondant à chaque énoncé sont séparés dans le fichier `.conllu` par une seule ligne vide. Les deux figures suivantes montrent un énoncé en naija tel qu'il est représenté dans un fichier `.conllu` (figure 2) suivi par la représentation graphique de son arbre syntaxique (figure 3).

Figure 2 – Arbre de dépendance en format .conllu

```
# sent_id = WAZP_03_Education_MG__13
# text = dat na stupid investment //
# text_en = That's a stupid investment.
# text_ortho = Dat na stupid investment.
1	dat	dat	PRON	_	Number=Sing|PronType=Dem	2	subj	_	AlignBegin=36666|AlignEnd=36892|Gloss=SG.DEM
2	na	na	PART	_	PartType=Cop	0	root	_	AlignBegin=36892|AlignEnd=37176|Gloss=be
3	stupid	stupid	ADJ	_	_	4	mod	_	AlignBegin=37481|AlignEnd=38287|Gloss=stupid
4	investment	investment	NOUN	_	_	2	comp:pred	_	AlignBegin=38287|AlignEnd=39109|Gloss=investment
5	//	//	PUNCT	_	_	2	punct	_	AlignBegin=39109|AlignEnd=39139|Gloss=PUNCT
```

Figure 3 - Représentation visuelle d'un arbre de dépendance



Bien que le format `.conllu` soit extrêmement utile pour encoder succinctement diverses informations macrosyntaxiques, l'interprétation d'un arbre syntaxique dans un tel format textuel requiert un effort considérable pour un être humain. L'ajout ou la suppression de tokens ou la modification des relations syntaxiques qui les relient dans un fichier `.conllu` est encore plus difficile, car l'utilisateur doit s'assurer que chaque token et la position de son gouverneur sont correctement numérotés. Afin de réduire les erreurs humaines et de faciliter les corrections manuelles, nous avons utilisé l'outil Arborator-Grew[3], qui permet aux utilisateurs de télécharger des collections de fichiers `.conllu` et de représenter leur contenu par une interface graphique intuitive. Ainsi, tous les arbres représentés ci-dessus ont été générés par Arborator-Grew. En plus de permettre aux utilisateurs de représenter graphiquement les arbres, Arborator-Grew fournit également un ensemble d'outils conviviaux pour les modifier. Notamment, des relations syntaxiques peuvent être établies entre les tokens simplement en glissant une flèche d'un gouverneur à son dépendant. Les utilisateurs peuvent également ajouter ou supprimer des tokens, et la numérotation des mots dans le fichier `.conllu` sous-

[3] https://arboratorgrew.elizia.net/#/

jacent sera modifiée en conséquence. Les utilisateurs peuvent ensuite exporter les fichiers `.conllu` modifiés. Une description complète de cet outil peut être trouvée dans Guibon *et al.* (2020).

Arborator-Grew intègre également un outil informatique appelé ***G**raph **Rew**riting* (GREW[4]), qui permet aux utilisateurs d'écrire et d'exécuter des règles généralisées pour la modification des fichiers `.conllu`. Ceci est particulièrement utile pour la correction automatique de certaines erreurs récurrentes qu'il serait laborieux de corriger manuellement. Par exemple, on peut écrire une règle imposant que tous les tokens représentant un certain lemme portent une certaine partie du discours. On pourrait aussi théoriquement écrire une règle dictant que toutes les relations de dépendance entre un auxiliaire et un verbe doivent porter l'étiquette `comp:aux` si ce n'est pas déjà le cas. Bien entendu, GREW offre la possibilité d'écrire des exceptions aux règles en dictant les conditions dans lesquelles une règle ne sera pas appliquée. Les spécificités de cet outil sont décrites en détail dans Guillaume (2021).

2.3.2 La macrosyntaxe

En plus des annotations microsyntaxiques, ce projet comprend également une couche d'annotation macrosyntaxique. Contrairement à la microsyntaxe, qui décrit les relations de dépendance syntaxique entre les mots, la macrosyntaxe divise les énoncés en unités basées sur leur autonomie illocutoire, c'est-à-dire sur le fait qu'un segment donné peut ou non former un énoncé acceptable en isolation. Notre approche de l'annotation macrosyntaxique est basée sur celle utilisée dans le projet ANR Rhapsodie, qui a créé un treebank similaire de français oral. Cette approche est décrite en détail dans Pietrandrea et Kahane (2019), un chapitre paru dans Lacheret-Dujour *et al.* (2019).

Pour les besoins de ce projet de recherche, nous pouvons conceptualiser la macrosyntaxe comme un moyen de segmenter un énoncé en unités majeures basées sur certaines propriétés sémantiques et distributionnelles. Les trois unités principales sont les suivantes :

- **Le noyau** – Le noyau porte la force illocutoire d'un énoncé, encodant l'intention première du locuteur de produire cet énoncé. Par exemple, on peut dire que dans l'énoncé *Lucie, peux-tu me passer le sel*, le segment *peux-tu me passer le sel* porte la force illocutoire. Tout énoncé ininterrompu doit contenir au moins un noyau, bien que des unités macrosyntaxiques supplémentaires puissent également être présentes. Contrairement aux autres unités, les noyaux sont totalement autonomes. Si l'on

[4] https://grew.fr

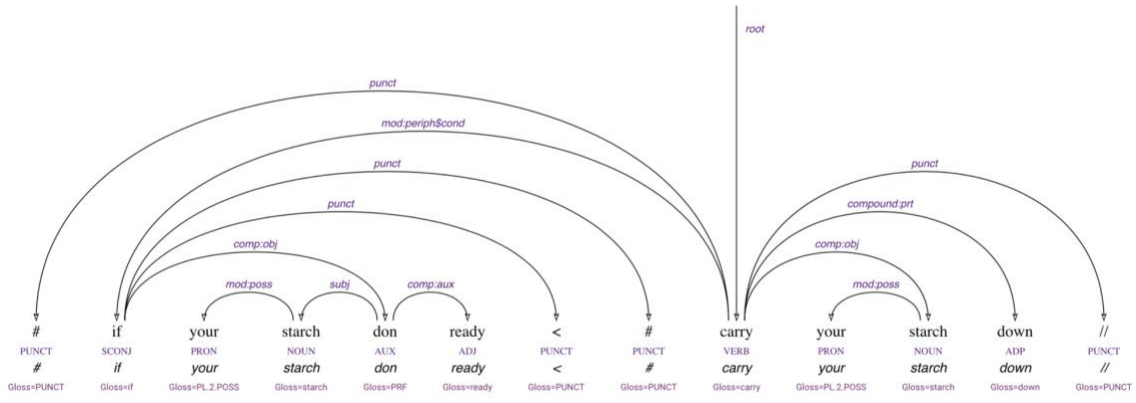
supprime tous les éléments d'un énoncé à l'exception du noyau, on obtient tout de même un énoncé acceptable ayant à peu près le même sens.

- **Le prénoyau** – Le prénoyau est une unité macrosyntaxique facultative située à gauche d'un noyau. Les prénoyaux peuvent jouer un certain nombre de rôles y compris celui du marquage des topics. Ces unités pourraient être supprimées de l'énoncé sans que cela n'ait d'impact sur son sens fondamental. Parfois, un noyau peut être précédé d'une longue série de prénoyaux.

- **Postnoyau** – Comme un prénoyau, un postnoyau est une unité macrosyntaxique qui ne répond pas aux critères définis pour un noyau, et qui pourrait être éliminée sans affecter l'acceptabilité de l'énoncé. Comme son nom l'indique, un postnoyau se trouve à droite d'un noyau.

Les annotations macrosyntaxiques qui identifient ces unités sont représentées directement dans les fichiers `.conllu` par l'ajout de signes de ponctuation utilisant des symboles spéciaux. Les prénoyaux sont délimités par le symbole <, et les postnoyaux par le symbole >. Le symbole macrosyntaxique >+ est également utilisé pour délimiter les postnoyaux faisant partie d'une construction clivée de type *na nineteen eighty four* >+ *wey de born me* 'c'est en 1984 qu'ils m'ont donné naissance'. Notez dans l'arbre syntaxique représenté dans la figure 4 que le symbole macrosyntaxique < est traité comme un token exactement comme les mots de l'énoncé. Les mots apparaissant à gauche du symbole constituent le prénoyau, tandis que ceux apparaissant à droite font partie du noyau.

Figure 4 - Un énoncé en naija contenant un noyau et prénoyau



Pour donner un exemple clair des propriétés distributionnelles des noyaux, prénoyaux et postnoyaux, considérons l'énoncé suivant, qui contient les trois types d'unités :

and dat starch < # e dey get am from # inside cassava > # cassava wey de take dey do gari //
 # et cet amidon < # ils l'obtiennet de l'intérieur # du manioc > # le manioc qu'ils utilisent pour produire le gari //

Notez que si nous supprimions à la fois le prénoyau et le postnoyau, il nous resterait un énoncé syntaxiquement er sémantiquement acceptable, *e dey get am from # inside cassava* 'ils l'extraient du manioc'. Cependant, aucun de ces composants optionnels ne peut former un énoncé acceptable en isolation. Il existe plusieurs autres symboles macrosyntaxiques dans ce corpus qui sont utilisés pour identifier d'autres structures. Cependant, ces structures dépassent le cadre de ce projet et ne seront pas abordées dans ce mémoire.

Ce projet de recherche se concentre sur le prénoyau, que nous appelons aussi la **position initiale**. Nous avons limité le champ de notre étude aux prénoyaux car ces unités macrosyntaxiques semblent jouer un rôle important dans la structure informationnelle du naija et d'autres langues. En naija, les prénoyaux ont tendance à remplir d'importantes fonctions pragmatiques, comme désigner le topic qui sera abordé dans la suite de l'énoncé, ou situer la portée de l'énoncé dans l'espace ou le temps. Considérons les deux exemples suivants.

1) *di number one problem for dis country now < na { poverty |c and hunger }*
   'le problème numéro un dans ce pays < c'est la pauvreté et la faim'

2) *when di time near < you go see*
   'quand le moment sera venu < tu verras'

Dans l'énoncé 1, le prénoyau introduit le topic discuté dans le noyau. Dans l'énoncé 2, le prénoyau sert à situer le contenu du noyau dans un certain cadre temporel. Ces fonctions seront abordées plus en détail dans la section 2.5 consacrée à la structure informationnelle.

2.4 La prosodie

Avec le niveau segmental, la prosodie, appelée également niveau suprasegmental, constitue la charpente sonore du langage. Si le niveau segmental concerne les consonnes et les voyelles, leurs catégories, et les règles qui régissent leur utilisation dans les langues, la prosodie s'intéresse principalement aux autres aspects des vocalisations humaines. Parmi les caractéristiques prosodiques fréquemment citées figurent l'intensité, i.e. le volume sonore réel et perçu de la parole; la durée, i.e. la longueur temporelle des sons de la parole, et la hauteur de la voix (Lacheret-Dujour et Beaugendre 1999:12). Cette dernière est une fonction directe de la fréquence fondamentale (F0) des sons émis par le locuteur, ou de la vitesse à laquelle ses cordes vocales vibrent.

Il existe des différences significatives de F0 d'une personne à l'autre en fonction de facteurs tels que l'âge et le sexe. Les hommes adultes, par exemple, ont tendance à avoir une F0 d'environ 120 Hz, tandis que les femmes ont tendance à avoir une F0 d'environ 200 Hz (Pépiot 2014). Cependant, même au cours d'un seul acte de parole, la F0 d'une personne a tendance

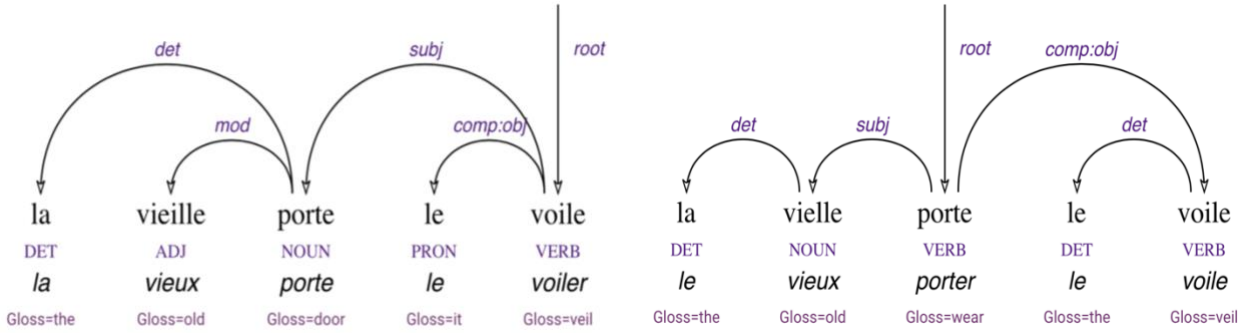
à subir des modulations importantes. Ces hausses et baisses de la hauteur constituent un aspect fondamental de la prosodie et jouent souvent un rôle communicationnel. En français, par exemple, une hausse ou une baisse de la F0 fait la différence entre la question *tu viens avec moi ?* et l'injonction *tu viens avec moi*. Si les différences intonatives entre les questions et les déclarations sont peut-être les exemples les plus évidents de l'importance de la prosodie, les modulations de la F0 peuvent jouer un large éventail d'autres fonctions. Les locuteurs peuvent notamment exprimer des attitudes, des états émotionnels, la politesse, et l'ironie par des modulations de leur F0. La prosodie peut même jouer un rôle dans l'interprétation des structures syntaxiques. Prenons l'exemple de la phrase ambiguë *la vieille porte le voile*[5]*,* qui peut avoir deux interprétations différentes soulignées par ces paraphrases :

1) La vieille dame porte le voile.

2) La porte qui est vieille cache l'objet.

Ces deux sens correspondent en fait à deux analyses syntaxiques distinctes. Dans le premier, *la vieille* constitue un syntagme nominal qui est le sujet syntaxique du groupe verbal *porte le voile*. Dans la deuxième interprétation, *la vieille porte* constitue un syntagme nominal qui est le sujet du verbe *voile*. Ces deux interprétations sont illustrées dans la figure 5 ci-dessous.

Figure 5 - Analyses syntaxiques d'un énoncé ambigu



Ces deux structures syntaxiques sont également reflétées par des modulations différentes de la F0. Dans la première, la F0 monte sur le sujet *la vieille* alors que le prédicat *porte le voile* est caractérisé par une F0 descendante. Dans le deuxième, la F0 monte tout au long du segment *la vieille porte* et descend sur le prédicat *le voile.* Cet exemple montre une correspondance particulièrement forte entre les structures syntaxiques et la structure prosodique. Dans les deux interprétations, il y a une montée de la F0 sur le sujet indépendamment de sa structure syntaxique interne. Le prédicat verbal est également caractérisé par des contours F0 descendants dans les deux exemples. **L'interface syntaxe-**

[5] Exemple donné par le professeur Cédric Gendrot lors de son cours sur la synthèse vocale à Paris 3 en 2019.

prosodie est un terme souvent utilisé pour décrire cette relation. Si cet exemple montre une relation particulièrement forte entre syntaxe et prosodie, la correspondance entre les deux n'est pas toujours aussi évidente. Des études précédentes suggèrent que si la structure prosodique peut jouer un rôle de désambiguïsation dans l'interprétation des structures syntaxiques, il n'y a pas toujours un alignement parfait entre les deux dans le discours oral (Lacheret-Dujour et Beaugendre 1999:12).

Des facteurs prosodiques peuvent également faire la différence entre de la parole qui est perçue comme naturelle et celle qui est perçue comme plate, maladroite, voire robotique. Pour cette raison, une meilleure compréhension de la prosodie peut apporter des contributions importantes au secteur du traitement automatique des langues, particulièrement en ce qui concerne la synthèse vocale et la reconnaissance automatique de la parole.

2.4.1 La prosodie du naija

La typologie prosodique exacte du naija reste un sujet de débat. Certains chercheurs ont affirmé que le naija est une langue à tons, parce que la hauteur est utilisée pour contraster entre un ensemble limité de lexèmes tels que [fádà] 'père' et [fādá] 'prêtre catholique'. D'autres ont soutenu que le naija est une langue à accent de hauteur, une langue dans laquelle certaines syllabes sont perçues comme étant plus proéminentes que d'autres, et que cette proéminence perçue est marquée par une différence de F0 (Oyelere 2021:18). D'autres encore affirment que le système prosodique du naija ne correspond à aucune catégorie typique (Oyelere 2021:19).

Des études récentes sur les caractéristiques prosodiques du naija présentent un certain nombre d'observations intéressantes. Les mots lexicaux semblent se distinguer par une fréquence fondamentale élevée (Simard *et al.* 2019). Dans un exemple particulièrement intéressant, les différentes utilisations de *dey*, une copule et un marqueur de l'aspect imperfectif, se distinguent par une différence de hauteur, l'usage imperfectif portant un ton bas (Agbo et Plag 2020).

Les proéminences syllabiques perçues jouent également un rôle dans la structure informationnelle de cette langue, contribuant au marquage des focus (Simard *et al.* 2019). Cependant, à notre connaissance, aucun travail n'a été effectué sur le marquage prosodique des topics ou d'autres composants pragmatiques dans les prénoyaux du naija.

2.4.2 La prosodie et le TAL

2.4.2.1 Histoire

Les chercheurs ont depuis longtemps compris l'utilité des données prosodiques dans le traitement automatique de la parole, notamment dans les domaines de la génération automatique de la parole à partir d'un texte (synthèse vocale) et de la reconnaissance automatique de la parole. Depuis les années 1970, les chercheurs ont développé diverses méthodes pour émuler automatiquement la structure prosodique du français dans le but de développer une synthèse vocale plus réaliste. Divers modèles et techniques ont été développés au fil des ans, y compris ceux prenant en compte les informations syntaxiques et lexicales (pour un aperçu complet de ces premiers systèmes, voir Lacheret-Dujour et Beaugendre 1999:203-230). Parallèlement, la prosodie de la parole enregistrée a également été utilisée pour effectuer diverses tâches comme la détection des limites entre les phrases et les syntagmes, et la segmentation des mots en syllabes (Nasri et Caelen-Haumont 1991, Pérennou et Caelen 1982). Price et Baer (1990) ont notamment utilisé des données relatives à la durée pour désambiguïser certaines constructions syntaxiquement ambiguës. En dehors de la synthèse et du traitement de la parole, les données prosodiques ont été utilisées avec succès dans certaines tâches de classification automatique. Ranganath *et al.* (2009) ont notamment démontré que la prise en compte de certaines caractéristiques prosodiques de base comme l'intensité, la variation de la F0 et le débit de la parole est utile pour classifier la parole par intentions sociales humaines comme le flirt.

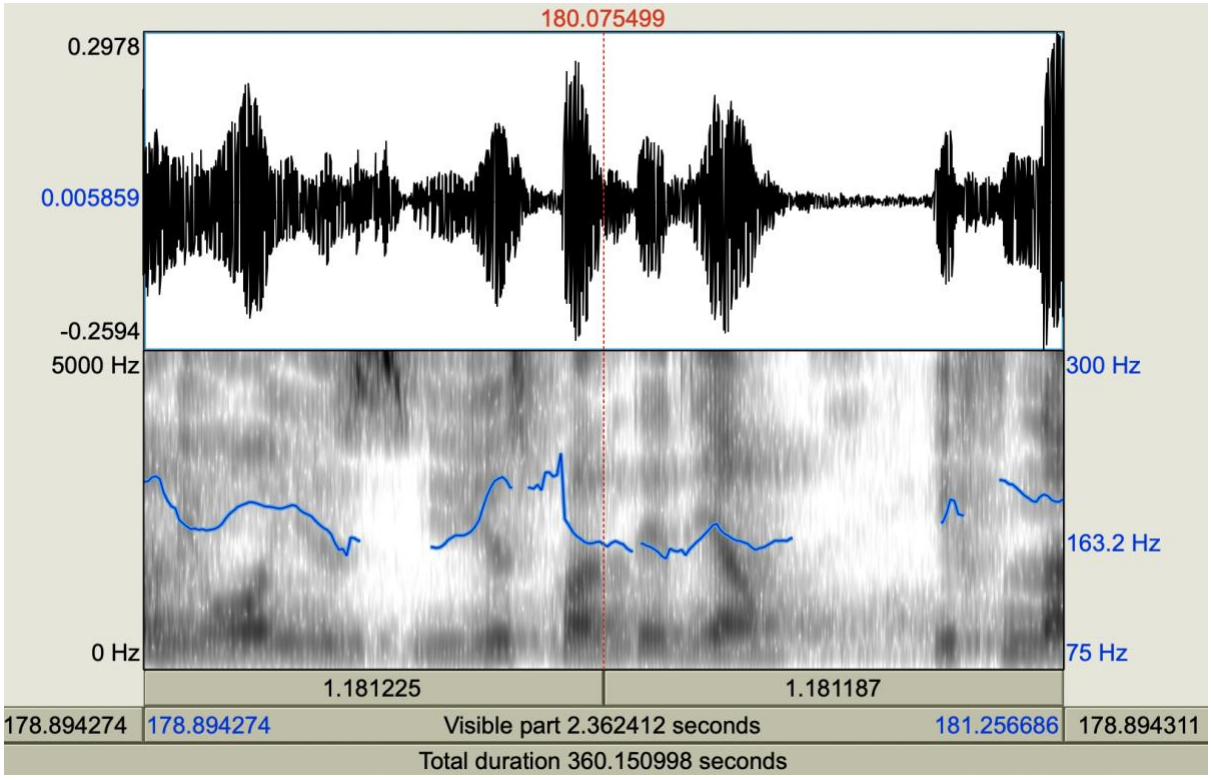
En raison des récents progrès en termes de la puissance de calcul et de la disponibilité accrue de données linguistiques, les chercheurs font de plus en plus appel aux réseaux neuronaux pour modéliser la prosodie (Bernandy et Themistocleous 2017). La prosodie de la parole humaine étant caractérisée par d'innombrables micro-variations souvent imperceptibles, un objectif commun aux tentatives de modélisation de la prosodie est d'identifier et de modéliser uniquement les variations qui sont perceptibles et significatives pour un auditeur. Ceci est au cœur d'un outil appelé SLAM+ qui a été exploité dans le développement de ce projet, et qui sera présenté dans la section suivante.

2.4.2.2 Outils pour le traitement automatique de la prosodie

Un certain nombre d'outils informatiques ont été développés pour faciliter l'étude et le traitement de la prosodie. L'un des programmes les plus utilisés pour étudier la prosodie et la phonétique de façon plus générale est **Praat**, dont les applications sont décrites par Boersma et van Heuven (2001). Créé en 1992, Praat fournit une vaste boîte à outils pour l'analyse, l'annotation et la modification des sons de la parole humaine. Praat permet notamment aux

utilisateurs de télécharger des fichiers sonores au format `.wav`, et de générer une représentation visuelle de son contenu sonore grâce à son interface graphique. La figure 6 illustre cette interface. En haut, nous voyons l'oscillogramme (*wave form*) d'un segment de parole, qui fournit une visualisation de l'onde sonore. En dessous, nous voyons le spectrogramme, qui permet de visualiser les fréquences les plus intenses dans le signal vocal (les stries foncés, appelés formants, représentent les fréquences les plus intenses). Sur le spectrogramme, la ligne bleue représente la F0, dont la valeur en Hz est calculée par Praat à partir du fichier sonore.

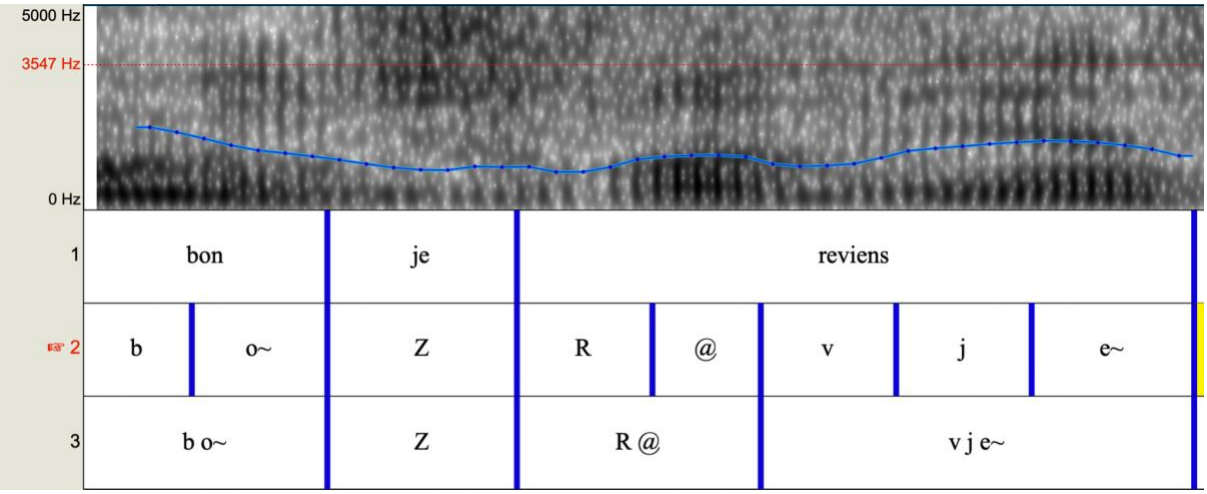
Figure 6 - Un segment de parole visualisé sur Praat



Praat permet aux utilisateurs d'extraire automatiquement ces modulations de la F0 à partir d'un fichier sonore et de les enregistrer dans le format de fichier **.PitchTier**. Notez que cette extraction de la F0 est souvent imparfaite, contenant parfois des fréquences provenant du bruit de fond, ou des sons vocaux interprétés par erreur comme de la fréquence fondamentale.

La deuxième fonctionnalité de Praat qui est pertinente pour ce projet est la segmentation et l'annotation des fichiers sonores. Praat permet de créer plusieurs niveaux d'annotation appelés *tiers*, qui permettent aux utilisateurs de segmenter un fichier et d'attribuer une annotation textuelle à chaque segment. La figure 7 montre le spectrogramme du segment de parole *bon je reviens*, accompagné de trois tiers en dessous. De haut en bas, ceux-ci segmentent le fichier sonore au niveau du mot, du phonème et de la syllabe.

Figure 7 - Segment de parole segmenté annoté sur Praat



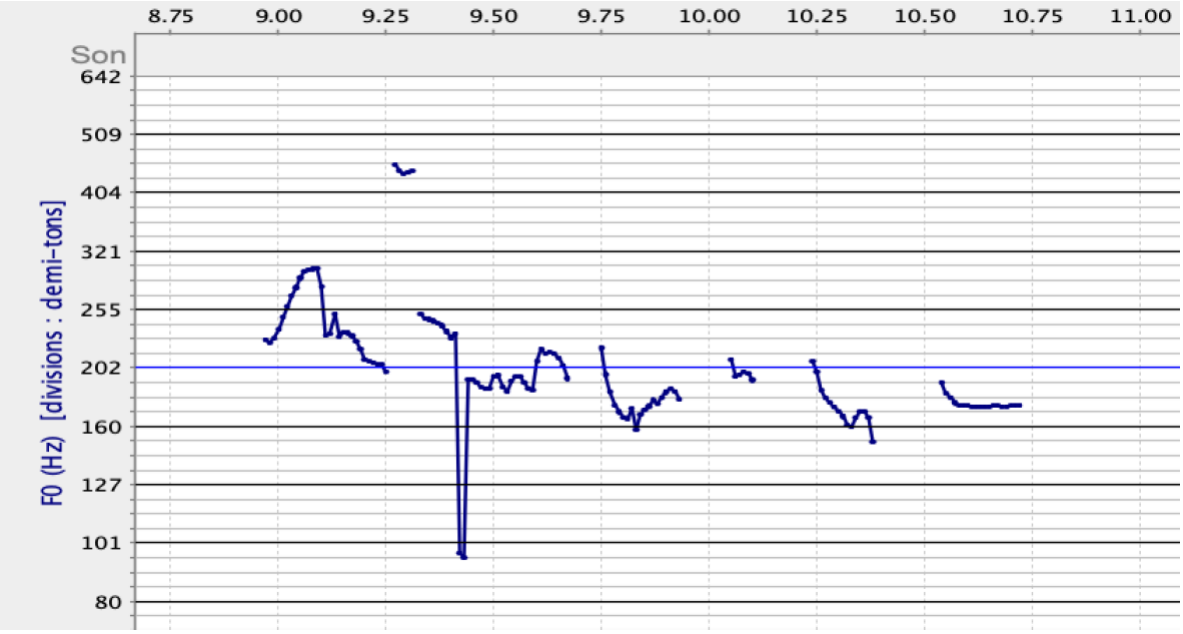
Le contenu textuel et les positions des limites entre chaque segment peuvent être modifiés librement par l'utilisateur. Dans cet exemple, le tier segmentant la parole au niveau du mot contient une représentation orthographique de chaque mot, tandis que les tiers phonémique et syllabique contiennent des représentations phonétiques. Ces annotations peuvent être enregistrées pour une utilisation ultérieure sous la forme d'un fichier texte au format `.TextGrid`. Les fichiers `.TextGrid` contiennent, pour chaque tier, les valeurs temporelles correspondant au début et à la fin de chaque segment, ainsi que leur contenu textuel.

Un outil supplémentaire d'analyse phonétique, qui a également joué un rôle important dans la réalisation de ce projet, est ANALOR, dont les fonctionnalités sont décrites par Avanzi *et al.* (2008). L'interface graphique d'ANALOR présente une ressemblance superficielle avec celle de Praat. Cependant, ce programme est adapté à l'étude de la prosodie. ANALOR permet notamment la détection automatique des proéminences accentuelles et des unités prosodiques majeures, appelées périodes intonatives, dont les limites sont définies par des pauses de plus de 300 ms qui sont directement précédées et suivies de changements significatifs de la F0.

ANALOR est également utile pour le nettoyage des fichiers `.PitchTier`. L'interface de ce logiciel permet aux utilisateurs d'ouvrir les fichiers `.PitchTier` avec les fichiers sonores `.wav` correspondants, et d'inspecter les sections où la F0 détectée semble erronée. Considérons la figure 8 ci-dessous, où la F0 subit une chute soudaine et rapide et revient immédiatement à sa hauteur précédente. Dans ce cas, un utilisateur pourrait écouter le fichier sonore pour déterminer si cette chute de la F0 est réellement reflétée dans la voix du locuteur. Si ce n'est pas le cas, ANALOR fournit les outils nécessaires pour augmenter la hauteur du segment fautif afin qu'il corresponde au reste du contour F0. Les utilisateurs peuvent même

supprimer des segments du contour F0 s'ils décident qu'ils sont causés par des facteurs tels que le bruit ambiant et ne correspondent pas à la véritable fréquence fondamentale du locuteur. La F0 corrigée peut alors être sauvegardée comme un nouveau fichier `.PitchTier`.

Figure 8 - Contour douteux de la F0 sur ANALOR



La capacité d'extraire automatiquement des contours stylisés de F0 à partir de la parole est également au cœur de ce projet. Un **contour stylisé** est une modélisation des contours réels de F0 que nous avons vus plus haut qui vise à prendre en compte uniquement les variations qui sont perceptibles par l'auditeur. À notre connaissance, il n'existe qu'un seul outil disponible publiquement capable d'extraire automatiquement des contours de hauteur stylisés de la parole sur des segments de nature et d'empan variables : SLAM+ (Liu *et al*. 2019).

SLAM+ est un logiciel qui calcule les contours mélodiques en prenant comme entrées un fichier `.TextGrid` et un fichier `.PitchTier` correspondant. Pour un segment donné du fichier `.TextGrid`, SLAM+ analyse le contour F0 correspondant et génère un code textuel décrivant son contour mélodique. Ces codes sont constitués d'un alphabet de cinq tons élémentaires décrivant la hauteur relative du contour à un moment donné. Ceux-ci sont représentés par les symboles H (extrêmement élevé), h (élevé), m (moyen), l (bas) et L (extrêmement bas). Notez que ces valeurs sont toutes relatives à la F0 moyenne du locuteur sur un intervalle donné. Le ton m représente une F0 comparable à la F0 moyenne du segment utilisé pour calculer la moyenne du locuteur.

Pour les contours simples, tels qu'un ton plat ou un simple ton descendant ou montant, le code contiendra deux lettres décrivant la hauteur de la F0 au début et à la fin d'un segment. Le code hh correspondrait donc à un contour plat avec une F0 élevée, tandis que lh

correspondrait à un simple contour montant. Cependant, les codes SLAM+ peuvent également contenir des informations correspondant aux modulations de la F0 à l'intérieur d'un contour. Si le contour F0 contient une saillance interne significative (un pic ou un creux significatif de la F0), une troisième étiquette sera générée pour décrire sa hauteur. Cette troisième valeur sera accompagnée d'un descripteur numérique décrivant sa position dans le contour. h1 correspond à une saillance interne élevée près du début du contour, l2 correspond à une saillance interne faible au milieu du contour, et m3 à une saillance interne moyenne près de la fin du contour. Les contours ayant une saillance interne sont appelés contours complexes.

L'étiquette complète d'un contour complexe est ainsi composée de trois lettres et d'un chiffre. Les deux premières lettres correspondent à la F0 du début et de la fin du contour, et la dernière lettre et le nombre décrivent la hauteur et la position de la saillance interne. Par exemple, mmh2 désigne un contour complexe qui commence et se termine par une F0 moyenne et qui a une saillance interne élevée au milieu.

La figure 9 montre comment SLAM+ génère des contours stylisés à partir de l'énoncé *if e get more than PhD < I go get more than PhD* (‘s’il obtient plus qu'un doctorat, je vais obtenir plus qu'un doctorat’). Dans cet exemple, un tier `.TextGrid` a été utilisé pour segmenter l'énoncé en deux éléments : un prénoyau et un noyau. En utilisant les données prosodiques contenues dans le fichier `.PitchTier` et les données temporelles utilisées dans le fichier `.TextGrid`, une paire de contours stylisés a été produite pour chacun.

Figure 9 - Contours prosodiques générés à partir d'un prénoyau et noyau

Dans l'image du dessus, nous voyons les modulations de la F0 dans chacun des deux segments. En dessous, nous voyons en rouge et en vert une paire de contours stylisés produits pour chaque segment en utilisant ces informations prosodiques. Notez que dans chacun des contours stylisés, la hauteur augmente jusqu'à un pic avant de redescendre. Ces pics correspondent aux saillances internes, qui correspondent elles-mêmes aux pics les plus significatifs observés dans les modulations de F0 dans l'image du haut. La structure de chacun de ces contours est également encodée dans les codes textuels en rouge et vert visibles dans l'image du bas.

Les codes rouge et vert générés pour chaque segment correspondent à deux façons différentes de calculer le contour stylisé. Les codes rouges représentent les contours locaux, qui ne prennent en compte pour le calcul que l'information prosodique contenue dans le segment cible (le prénoyau ou noyau dans ce cas). Les codes verts représentent les contours globaux, dans lesquels la valeur des cinq tons élémentaires est influencée par la F0 moyenne observée dans un segment de parole plus large (l'ensemble de l'énoncé dans le cadre de ce projet). Ce qui est considéré comme une hauteur élevée dans le contour local (h) peut être

considéré comme extrêmement élevé (H) dans le contour global si la F0 moyenne du locuteur est plus basse dans ce segment plus large.

SLAM+ permet aux utilisateurs de définir quels tiers des fichiers `.TextGrid` doivent être utilisés dans ses calculs. En définissant le **targetTier**, les utilisateurs peuvent définir quel tier contient la segmentation utilisée pour générer les contours prosodiques. Pour générer les contours des unités macrosyntaxiques majeures telles que les noyaux et les prénoyaux, il faudrait sélectionner le tier qui segmente les énoncés en fonction de ces unités. Pour générer les contours des composants internes des prénoyaux, il faudrait choisir comme targetTier le tier contenant l'alignement correspondant. En définissant le **speakerTier**, les utilisateurs peuvent définir le tier dont la F0 moyenne influence le calcul des contours globaux. Ce tier doit contenir une segmentation plus large que celle utilisée pour le targetTier. Dans ce projet, nous avons systématiquement utilisé comme speakerTier le tier qui segmente les fichiers en énoncés, ce qui signifie que le contour global produit pour un segment cible sera influencé par les informations prosodiques contenues dans l'énoncé plus large dans lequel il apparaît.

A partir de ces informations, SLAM+ génère un nouveau fichier `.TextGrid` contenant deux tiers supplémentaires contenant respectivement les contours locaux et globaux calculés pour chaque segment du targetTier. Il génère également un troisième tier contenant une copie du tier défini comme **tagTier**. Le tier sélectionné pour ce rôle doit contenir le même nombre de segments que le targetTier, et le contenu de ses segments devrait d'une certaine manière être descriptif des segments équivalents de ce dernier. Pour les besoins de ce projet, il suffit de comprendre que pour générer des contours à partir des segments d'un tier donné, il doit exister un deuxième tier contenant la même segmentation.

2.5 La structure informationnelle

La structure informationnelle désigne la manière dont un locuteur encode des informations en fonction des connaissances de son interlocuteur. Plus précisément, la structure informationnelle d'un énoncé reflète les hypothèses d'un locuteur sur les connaissances de son interlocuteur à un moment donné (Simard 2010:172). Notamment, la structure informationnelle étudie la manière dont les locuteurs font la distinction entre les informations qu'ils considèrent comme étant connues de leur interlocuteur, et les informations qui sont introduites pour la première fois (Arnold *et al.* 2015). Le concept de structure informationnelle est basé sur la supposition que les langues permettent aux locuteurs d'encoder ces distinctions de façon explicite, notamment par le biais de diverses constructions syntaxiques ou de schémas d'accentuation. En français, la déclaration *Lucie est la femme de Gilles* peut porter au moins trois schémas d'accentuation selon le contexte. Les mots en gras

représentent des mots accentués dans ces exemples. ***Lucie*** *est la femme de Gilles* indique que Lucie est un nouvel élément pour l'auditeur, qui est déjà au courant du fait qu'il existe une personne nommée Gilles qui a une femme. En revanche, *Lucie est la femme de* ***Gilles*** suggère que l'interlocuteur sait que Lucie a un mari, mais ne connaissait pas son identité. Enfin, *Lucie est* ***la femme*** *de Gilles* suggère que l'interlocuteur connait les deux personnes, mais que la nature de leur relation est une nouvelle information.

Deux concepts essentiels dans l'étude de la structure informationnelle sont le **topic** et le **focus**. Le topic désigne une information connue dont un groupe d'interlocuteurs parle déjà. Quant au concept de focus, il fait référence à un nouvel élément d'information fourni sur le topic. Dans l'énoncé ***Lucie*** *est la femme de Gilles*, *la femme de Gilles* serait le topic, et *Lucie* le focus.

Les topics peuvent différer dans leur degré d'accessibilité cognitive pour l'auditeur. Un topic qui est mentionné pour la toute première fois, ou qui est réintroduit dans la discussion après un long détour, peut être considéré comme moins accessible cognitivement qu'un topic qui est déjà au centre de la discussion. Dans la paire d'énoncés "*Il y avait un invité nommé Robert à la fête hier. Il était plutôt sympathique.*", *un invité nommé Robert* est moins accessible que *il* alors que les deux font référence à la même personne.

Dans le cadre de ce projet, nous avons divisé les topics en différentes catégories selon leur fonction et leur niveau d'accessibilité cognitive. La première catégorie, à laquelle nous nous référons avec le terme anglais ***aboutness topics***, couvre tous les topics qui sont le centre d'intérêt principal d'un énoncé. Nous les divisons en plusieurs sous-catégories :

- **Nouveaux topics** : les topics qui sont introduits dans une discussion pour la toute première fois. Nous les étiquetons avec le label TOP_A NEW.
- **Topics continus** : les topics qui sont déjà au centre de la discussion. Dans ce projet, nous les définissons comme des topics qui ont été mentionnés au moins une fois dans les cinq derniers énoncés d'un locuteur. Nous les étiquetons avec le label TOP_A CONTINUING.
- **Topics repris** (*resumed topics*) : Des topics qui sont réintroduits dans la conversation. Dans ce projet, nous définissons ces topics comme étant ceux qui ont été mentionnés pour la dernière fois six énoncés ou plus avant l'énoncé en question. Nous les étiquetons avec le label TOP_A RESUMED.

- **Topics récapitulatifs** (*summative topics*): les topics qui regroupent plusieurs topics précédents en un seul groupe, par exemple *toutes ces personnes dont j'ai parlé*. Nous les étiquetons avec le label TOP_A SUM.

Nous identifions également une catégorie de **topics contrastifs**, qui sont utilisés pour signaler un contraste entre deux référents. Par exemple, les deux personnes mentionnées dans l'énoncé *Pauline aime la pizza, mais Marie préfère les pâtes* sont des topiques contrastifs. Les topics contrastifs sont annotés avec l'étiquette TOP_CONTRAST.

La dernière catégorie de topics est consacrée aux cadres de discours, qui jouent généralement un rôle pour définir la portée d'un événement. Nous définissons quatre sous-catégories des cadres de discours:

- **Cadres temporels** : éléments qui situent un événement dans le temps (quand j'étais jeune, hier, en 1997...). Nous les annotons avec l'étiquette TOP-FRAME TEMPORAL.
- **Cadres spatiaux** : éléments qui situent un événement dans l'espace (chez moi, en France, dans la cuisine). Nous les annotons avec l'étiquette TOP-FRAME SPATIAL.
- **Cadres conditionnels** : éléments qui spécifient une condition requise pour qu'un certain événement se produise (si j'étais riche, si tu as toujours faim…). Nous les annotons avec l'étiquette TOP-FRAME CONDIT.
- **Autres** : les cadres qui n'entrent dans aucune des catégories ci-dessus. Cette catégorie contient généralement des adverbes ou d'autres éléments qui jouent un rôle énonciatif (marque l'attitude du sujet par rapport au propos qu'il décrit) ou descriptif (heureusement, lentement, etc.). Nous les annotons avec l'étiquette TOP-FRAME OTHER.

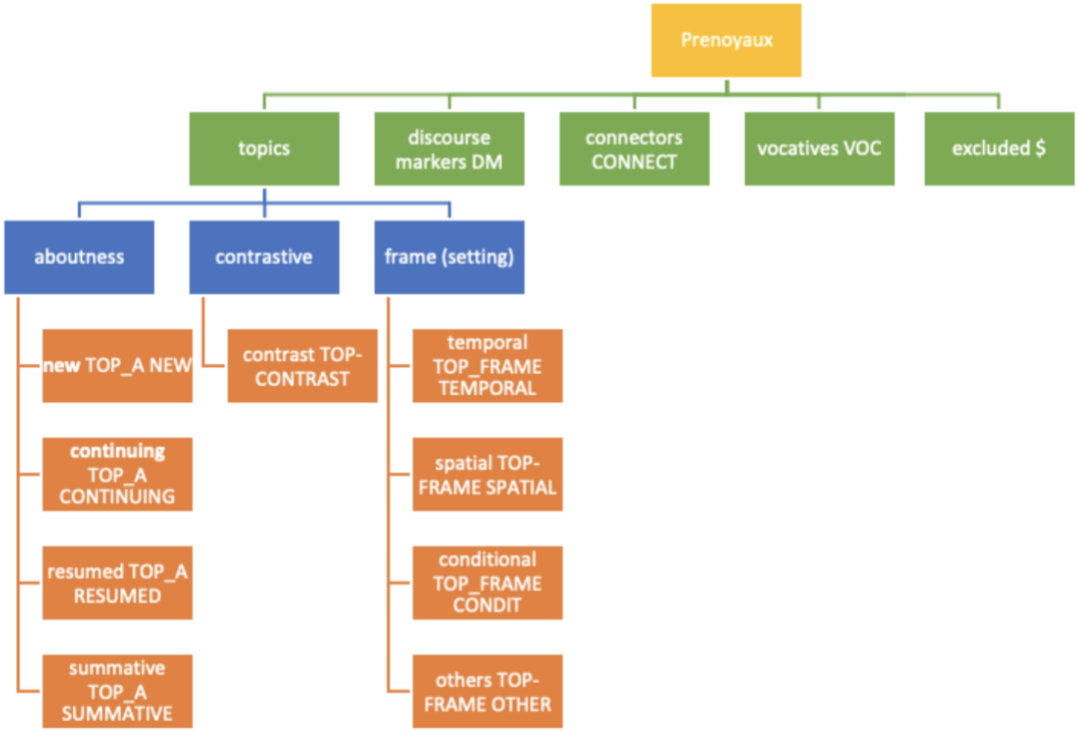
En dehors des topics, nous avons défini cinq autres catégories pour décrire la structure informationnelle du naija. Il s'agit des catégories suivantes :

- **Marqueurs de discours** : éléments qui jouent un rôle expressif ou modal (*tu sais, mais bon*). Annotés avec l'étiquette DM.
- **Connecteurs** : éléments qui relient un énoncé à un énoncé précédent (*et, mais*). Annotés avec l'étiquette CONNECT.
- **Vocatifs** : Éléments du discours où le locuteur s'adresse directement à un interlocuteur, typiquement un prénom (**Jean**, ferme la fenêtre ; **Mon ami**, comment vas-tu ?). Annotés avec l'étiquette VOC.

Pratiquement tous les prénoyaux étudiés dans le cadre de ce projet peuvent entrer dans au moins une de ces catégories. Cependant, il est important de noter que de nombreux prénoyaux peuvent être subdivisés en plusieurs unités internes qui portent chacune une seule fonction pragmatique. Par exemple, l'énoncé *and me < I no worry* (‘et moi, je ne m’inquiète pas’) contient à la fois un connecteur (*and*) et un topic (*me*). Dans le cadre de ce projet, chaque composant pragmatique d’un prénoyau est annoté individuellement selon sa fonction. Le prénoyau lui-même porte également une annotation contenant une liste d'étiquettes correspondant aux étiquettes de chaque composant. Ainsi, le *and me* porterait l’étiquette ‘CONNECT, TOP_A NEW’.

Enfin, nous avons également introduit une catégorie pour les prénoyaux que nous avons décidé d'exclure pour diverses raisons, comme la difficulté de déterminer sa fonction pragmatique. Typiquement, nous avons exclu les prénoyaux qui semblaient être très disfluents, avec un grand nombre de pauses, de reformulations ou d'hésitations. Nous avons également exclu les prénoyaux de plus de dix syllabes, qui nous semblaient trop longs pour générer des données prosodiques significatives. Ces prénoyaux ont été annotés avec l'étiquette $. Notez que cette étiquette a été appliquée uniquement aux prénoyaux entiers, plutôt qu'aux constituants des prénoyaux. La figure 10 résume toutes les catégories pragmatiques décrites ci-dessus avec leurs étiquettes correspondantes.

Figure 10 - Catégories pragmatiques des prénoyaux



2.6 La statistique

La statistique est un domaine de la mathématique qui couvre un large éventail de méthodes visant à faciliter l'analyse et l'interprétation des données. Les statistiques sont particulièrement utiles dans le domaine de la linguistique de corpus (voir Brezina 2018, Wallis 2021, Desagulier 2017), permettant aux chercheurs d'établir des relations entre différentes variables linguistiques. Par exemple, un chercheur peut utiliser des analyses statistiques sur un grand ensemble de données linguistiques pour déterminer les corrélations entre l'âge ou le sexe d'un locuteur et son utilisation de certaines constructions grammaticales.

L'objectif de notre analyse statistique était de voir s'il existait des corrélations entre les variations des contours prosodiques observées sur les prénoyaux, leurs types pragmatiques et leurs fonctions dans l'actualisation de la structure informationnelle. À cette fin, nous nous sommes appuyés sur deux méthodes statistiques : l'Analyse en composants multiples (ACM) et le test exact de Fisher.

2.6.1 Analyse en composants multiples

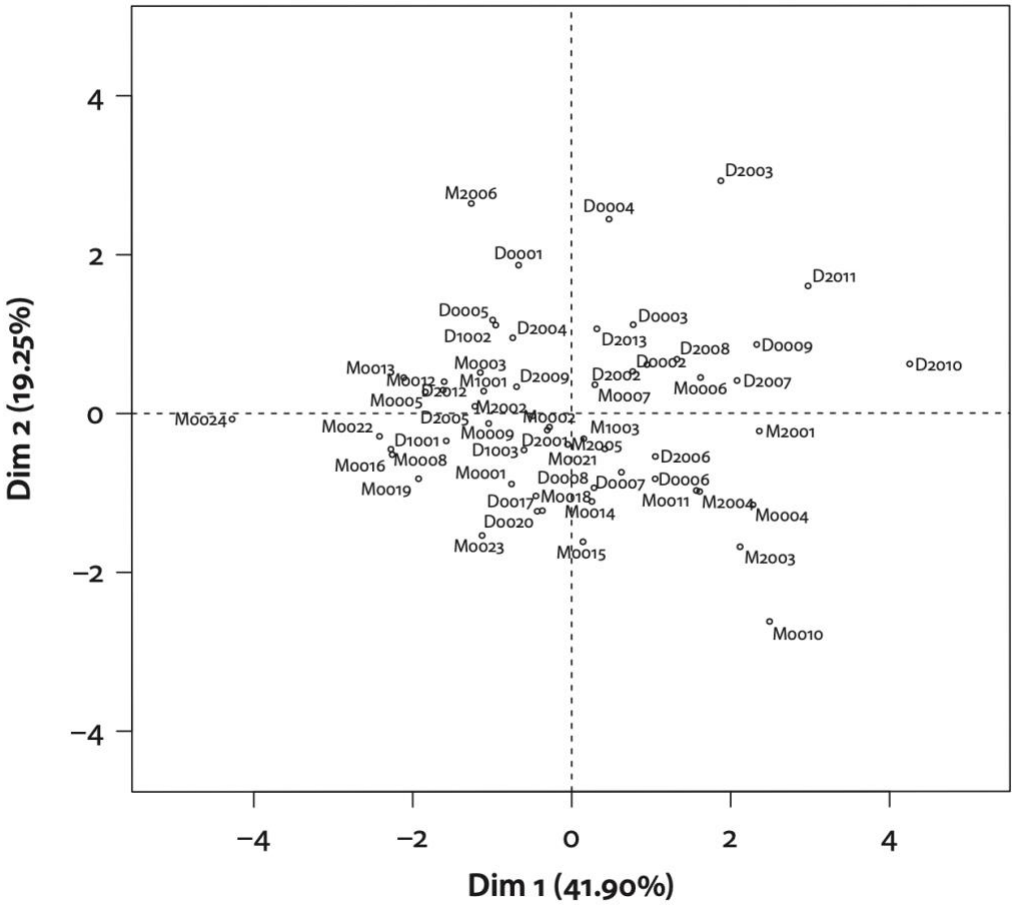
Notre méthodologie a été basée en partie sur le projet ANR Rhapsodie, dirigé par Anne Lacheret-Dujour, qui avait pour objectif d'explorer la structure prosodique du français oral (voir

Lacheret-Dujour *et al.* 2018). Cette étude s'est appuyée sur une technique statistique connue sous le nom d'Analyse en composantes principales (ACP), qui est utile pour comparer les profils de différentes variables et visualiser les corrélations entre elles (Lacheret-Dujour *et al.* 2018:319). L'ACP fonctionne en examinant la totalité de la variation d'un ensemble de données qualitatives et quantitatives, et en réduisant cette variation à un nombre limité de dimensions qui expliquent la variabilité du corpus. À partir de là, il est possible de projeter les variables sur un graphique en deux dimensions où chaque axe représente une de ces dimensions. Les graphiques les plus utiles sont ceux qui représentent les deux plus grandes dimensions, car ils expliquent la plus grande partie de la variabilité du corpus.

Cependant, cette technique statistique est généralement utilisée pour des ensembles de données impliquant une ou plusieurs variables continues. Comme toutes les variables utilisées dans ce mémoire de recherche étaient catégoriques, nous avons utilisé une technique similaire mais plus adaptée, connue sous le nom d'Analyse en correspondances multiples (ACM). Les graphiques générés par ces deux techniques sont similaires en apparence et peuvent être interprétés en utilisant les mêmes principes de base. Notez que les exemples suivants proviennent de graphiques ACP générés dans le cadre du projet Rhapsodie, mais que ceux-ci constituent néanmoins une illustration utile des concepts de base des graphiques ACM utilisés dans ce mémoire de recherche.

Le graphique ACP représenté dans la figure 11 (Lacheret-Dujour *et al.* 2018:322) sert d'exemple utile de la manière dont les graphiques ACP peuvent être interprétés. Ce graphique a été généré à partir d'un ensemble de données comprenant une série d'échantillons de discours. Chacun portait diverses données phonétiques quantitatives, telles que le nombre de pauses ou le nombre de euh d'hésitation par seconde dans chaque échantillon, ainsi que des données qualitatives telles que la structure de l'échantillon de discours (dialogue ou monologue), son contexte social (public ou privé) et son genre (narration, description, etc.). A partir de ces informations, les chercheurs ont produit un graphique ACP qui montre comment chaque échantillon de parole est associé à chacune de ces variables. Notez que les différentes variables auraient pu être affichées sur le même graphique, mais cela aurait produit un graphique avec trop de données pour être lu facilement.

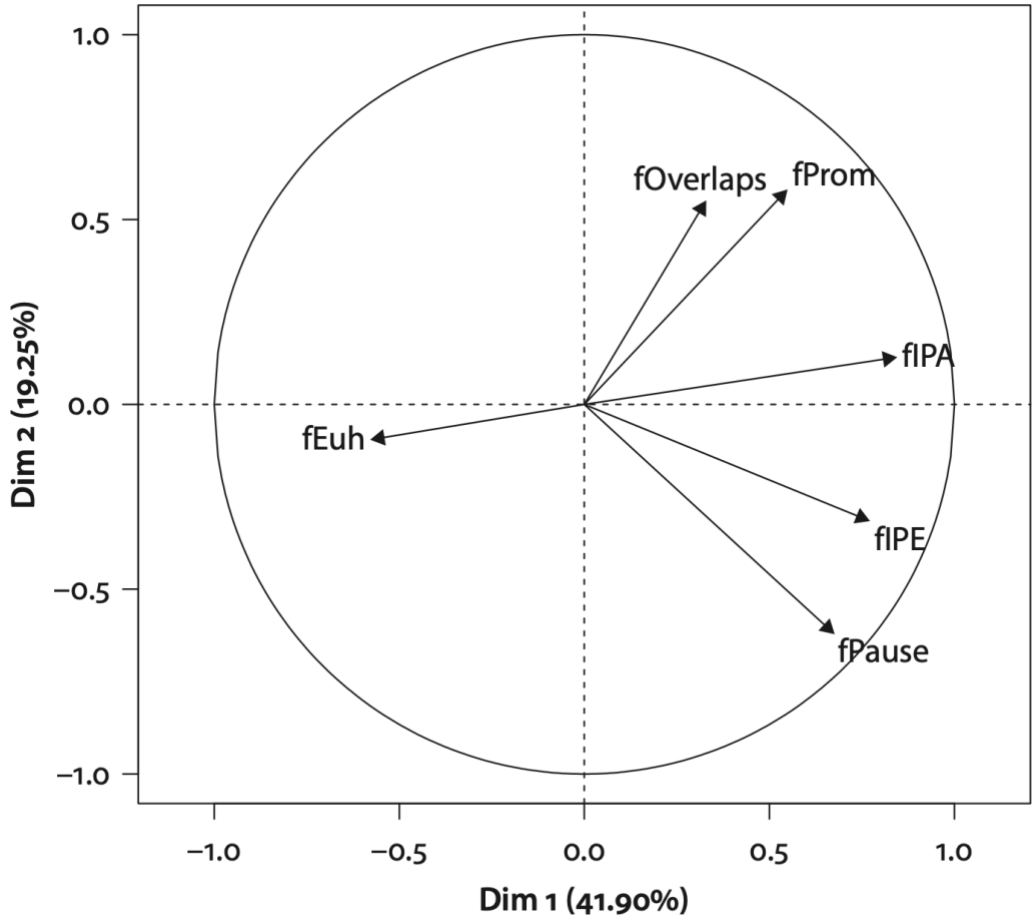
Figure 11 - Graphique ACP généré dans le cadre du projet Rhapsodie (échantillons de parole)



Dans ce graphique, chaque point correspond à un seul échantillon de parole. Les échantillons de parole situés à proximité les uns des autres présentent de grandes similitudes en termes des variables utilisées pour construire le graphique. En revanche, les échantillons situés très loin les uns des autres sont beaucoup moins similaires. Notez que l'axe horizontal représente 41,9% de la variation totale de l'ensemble de données, tandis que l'axe vertical représente 19,25%. Cela signifie que la majorité (61,15%) de la variabilité de l'ensemble de données est représentée sur ce seul graphique. Les graphiques ACP sont par nature très opaques, ce qui signifie qu'il n'est pas toujours facile de comprendre exactement quelles informations sont représentées dans ces dimensions. Si nous pouvons constater que deux échantillons de cet ensemble de données sont très similaires, il n'est pas nécessairement évident de savoir en quoi ils sont similaires.

Cependant, il est également possible de projeter les diverses variables prosodiques quantitatives sur le même ensemble de dimensions, comme le montre la figure 12 ci-dessous.

Figure 12 - Graphique ACP généré dans le cadre du projet Rhapsodie (variables quantitatives)

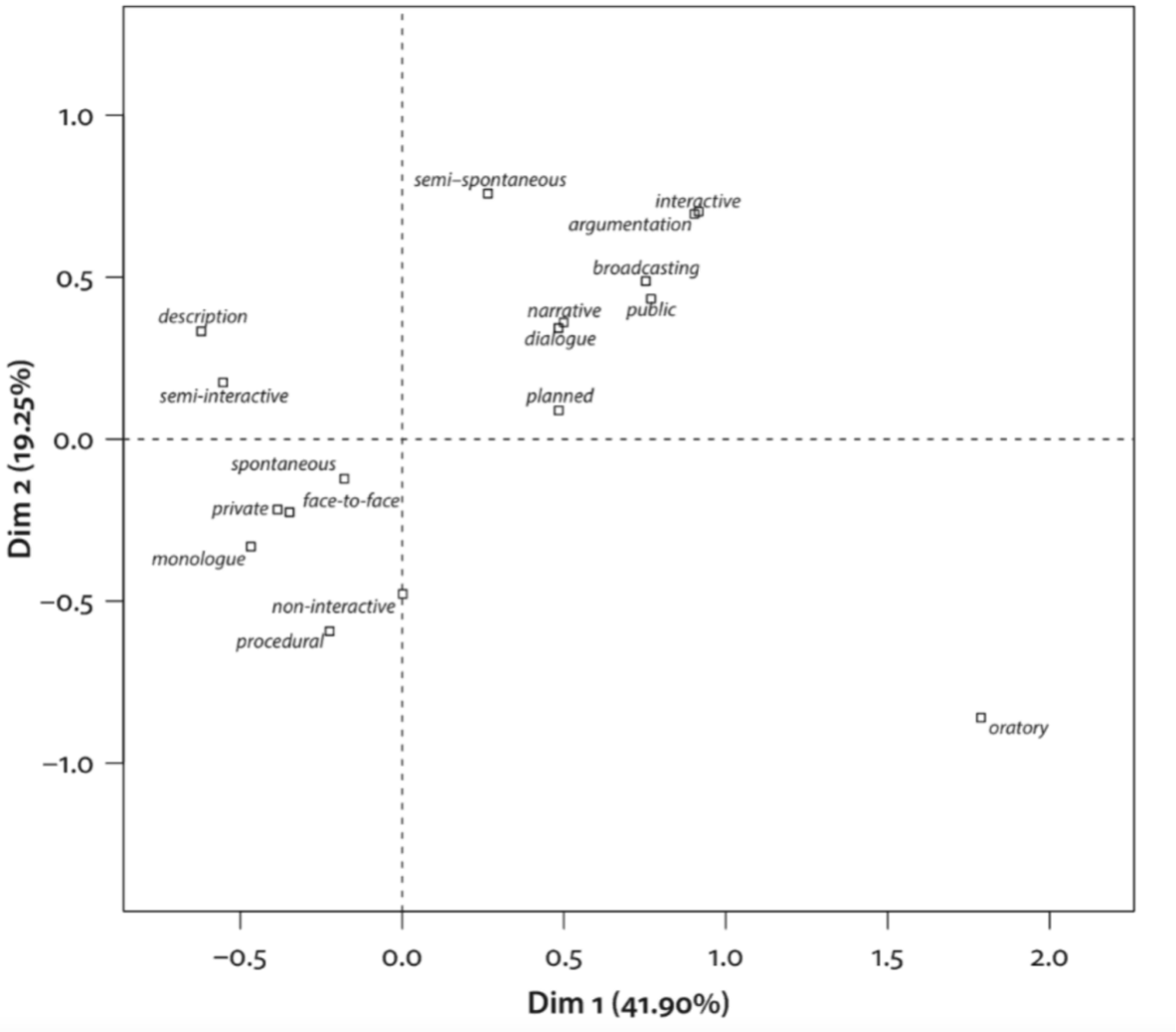


Dans ce graphique, les variables prosodiques quantitatives sont représentées par des flèches plutôt que par des points, mais les principes restent les mêmes : les flèches qui se rapprochent les unes des autres sont fortement associées et sont susceptibles d'apparaître dans les mêmes échantillons. Cette représentation est utile pour interpréter la distribution des échantillons vue dans la figure 11. Par exemple, on peut voir que les échantillons situés dans le coin supérieur droit du graphique de la figure 11 sont associés aux variables fOverlaps et fProm situées dans la partie correspondante de la figure 12, ce qui signifie que ces échantillons ont tendance à contenir plus de chevauchements entre les locuteurs et de syllabes proéminents.

Il est également possible de produire un graphique affichant uniquement les variables qualitatives, ce qui peut également permettre de faire des observations intéressantes. La figure 13 ci-dessous représente le même graphique, mais avec les différentes variables qualitatives affichées à la place des échantillons de parole. Ces variables illustratives aident à montrer les similarités partagées entre les échantillons de parole de différents types. Nous constatons notamment que les monologues, les discours privés et les discours spontanés

présentent tous des similitudes, tandis que les discours argumentatifs, les discours publics et les discours planifiés partagent également des caractéristiques en commun. Cela peut confirmer les intuitions du lecteur selon lesquelles il existe un grand nombre de chevauchements et de similitudes entre les différents membres de ces deux groupes.

Figure 13 - Graphique ACP généré dans le cadre du projet Rhapsodie (variables qualitatives)



Les graphiques ACP les plus intéressants sont ceux dans lesquels les dimensions représentent un haut degré de variabilité, et dans lesquels les différents points sont distribués d'une manière facilement interprétable, comme les clusters que nous avons vus dans l'exemple précédent.

Dans ce projet, nous avons utilisé un ensemble de données plus simple composé de deux variables qualitatives : les étiquettes pragmatiques et les étiquettes prosodiques. Dans le cadre de notre recherche, nous avons décidé de produire deux principaux types de graphiques ACM : ceux qui tracent les positions des étiquettes pragmatiques en fonction de l'étiquette prosodique, et ceux qui tracent les positions des étiquettes prosodiques en fonction

des étiquettes pragmatiques. Notre ensemble de données a ensuite été élargi à un plus grand nombre de variables prosodiques, en tenant compte de diverses caractéristiques prosodiques de chaque unité de parole de notre corpus. Des graphiques ACM supplémentaires ont été produits à partir de cet ensemble de données élargi.

2.6.2 Le test exact de Fisher

Les graphiques ACM constituent un moyen utile et intuitif de faire des observations générales sur la relation entre les variables d'un ensemble de données. Cependant, ils présentent également plusieurs inconvénients. Tout d'abord, ils sont mal adaptés à la tâche de quantifier les relations entre une seule paire de valeurs dans l'ensemble de données. Deuxièmement, comme l'ACM peut impliquer un grand nombre de dimensions significatives, l'examen d'un graphique représentant même les deux plus grandes dimensions peut cacher une grande quantité d'informations concernant la variation au sein du corpus. De ce fait, on peut conclure à tort que les valeurs correspondant à deux points proches sont plus fortement corrélées qu'elles ne le sont réellement, et vice-versa.

Le test exact de Fisher permet de mesurer la probabilité que la distribution des valeurs dans un tableau de contingence 2x2 soit le fruit du hasard ou non. Il existe une variété d'outils permettant d'appliquer ce test à un ensemble de données en utilisant des langages de programmation comme Python ou R. La réalisation de ce test génère une *p-value* comprise entre 0 et 1, représentant la probabilité que la distribution des valeurs soit le fruit du hasard. Une faible p-value indique une faible probabilité que la distribution soit le produit du hasard, indiquant ainsi une forte probabilité que les variables étudiées soient liées. Une p-value de 0,05 représente un degré de confiance de 95% que les variables ne sont pas indépendantes. Il s'agit traditionnellement du seuil à partir duquel les variables peuvent être présumées liées. Il convient néanmoins de noter que ce seuil est arbitraire, ce qui signifie que des valeurs p situées légèrement au-dessus peuvent encore être prises en considération.

Considérons le jeu de données suivant[6], qui classe 29 patients atteints de maladies en fonction de deux variables binaires : leur état de santé actuel (malade ou guéri) et le traitement qu'ils utilisent pour combattre la maladie (traitement A ou traitement B).

Tableau 1 – Tableau de contingence

	Traitement A	**Traitement B**
Malade	3	9
Guéri	13	4

[6] Exemple adapté de McDonald (2014).

L'application du test exact de Fisher permet de déterminer si ces deux variables sont corrélées ou non, c'est-à-dire si le choix du traitement a un impact sur le résultat de santé. L'application du test génère une p-value de 0,00772, indiquant un degré élevé de probabilité que les variables soient corrélées. On peut donc raisonnablement supposer que le choix du traitement peut déterminer si un patient sera guéri ou non.

Pour une description plus détaillée du test exact de Fisher et de ses principes, voir McDonald (2014).

4 Méthodologie

4.1 Étapes principales

Dans cette section, nous décrirons brièvement chaque étape majeure du projet pour conduire notre étude sur les corrélations entre les caractéristiques prosodiques et les fonctions pragmatiques des prénoyaux.

La première étape consiste à utiliser les 80 monologues du corpus Gold pour extraire les positions temporelles des unités macrosyntaxiques pertinentes, à savoir les prénoyaux et leurs composants pragmatiques internes. En termes plus techniques, nous avons utilisé les informations temporelles de chaque fichier `.conllu` pour produire une série de fichiers `.TextGrid` qui segmentent le fichier sonore correspondant en fonction des positions de diverses unités macrosyntaxiques. Il a fallu ensuite confirmer l'exactitude de ces alignements en comparant chaque segment de chaque fichier `.TextGrid` au segment audio correspondant à l'aide du logiciel Praat.

Une fois l'exactitude de l'alignement temporel sur les fichiers `.TextGrid` vérifiée, ils ont été utilisés pour produire des contours stylisés de chaque unité macrosyntaxique pertinente. Cette étape du processus a reposé sur le logiciel SLAM+, qui requiert deux formats de fichiers en entrée : un fichier `.TextGrid` contenant la segmentation temporelle, et le fichier `.PitchTier` correspondant, contenant les informations prosodiques relatives aux modulations de la fréquence fondamentale. Les fichiers `.PitchTier` peuvent être extraits directement d'un fichier `.wav` à l'aide de Praat. À une exception près, décrite dans la section 4.2.3, les fichiers `.PitchTier` nécessaires avaient déjà été produits et nettoyés par Biola Oyelere, un membre de l'équipe NaijaSynCor. Le fichier `.PitchTier` restant a dû être généré et nettoyé pendant la rédaction de ce mémoire.

La génération des contours stylisés nous a obligé à faire tourner SLAM+ deux fois sur nos données. La première fois, nous avons défini les paramètres de SLAM+ afin d'extraire un contour de chaque segment de la tier IC-text, qui segmente chaque énoncé en prénoyaux, noyaux et postnoyaux. La deuxième fois, nous avons extrait les contours prosodiques pour chaque segment du tier InPN-text, qui donne une segmentation pour chaque composant pragmatique interne de chaque prénoyau.

Nous avons ensuite exporté les informations prosodiques, macrosyntaxiques et temporelles pertinentes de chaque fichier `.TextGrid` sous forme de fichier `.tsv`. Pour faciliter nos analyses, chaque paire de fichiers `.tsv` correspondant à un même fichier a été fusionnée,

rassemblant à la fois les informations relatives aux prénoyaux et aux composants des prénoyaux.

À partir de là, nous avons entrepris une annotation pragmatique approfondie de chaque prénoyau et de ses composants pragmatiques internes. Pour des raisons de temps, nous avons dû limiter notre étude aux 23 fichiers pour lesquels une première analyse pragmatique avait déjà été produite par la chercheuse Candide Simard. Au cours de cette phase du projet, les 23 fichiers `.tsv` pertinents ont été fusionnés, et les annotations pragmatiques préalables ont été ajoutées. Ces annotations ont ensuite été vérifiées afin de corriger diverses erreurs décrites dans la section 4.2.4.

Le résultat de cet exercice a été la création d'un seul fichier `.tsv` contenant tous les prénoyaux de ces 23 fichiers, ainsi que leurs composants pragmatiques internes. Chacune de ces unités a été annotée en fonction de son type (prénoyau simple, prénoyau complexe et composant interne), de son type pragmatique et de ses contours prosodiques stylisés. Enfin, ce fichier a été utilisé pour produire une série de graphiques ACM pour chacun des trois niveaux d'analyse, qui ont ensuite été interprétés manuellement. Nous avons ensuite utilisé ce fichier `.tsv` pour ajouter un ensemble de descripteurs prosodiques à chaque segment, basés principalement sur le contenu de leurs contours stylisés associés. Ceci a été utilisé pour générer une nouvelle série de graphiques ACM. Le test exact de Fisher a également été appliqué à cet ensemble de données. Cette phase de notre exploration du corpus est traité dans la section 4.2.6.

4.2 Description des tâches principales

4.2.1 Extraction des unités macrosyntaxiques

Afin de réaliser ce projet, nous avons commencé par appliquer un script développé dans le cadre du projet NaijaSynCor qui permet d'extraire des 80 monologues du corpus Gold les informations temporelles correspondantes à certaines catégories d'unités macrosyntaxiques. Ces informations temporelles ont ensuite été utilisées pour générer un fichier `.TextGrid` pour chaque monologue qui segmente le fichier selon divers types d'unités macrosyntaxiques. Les fichiers `.TextGrid` générés en sortie contiennent 14 tiers, en fonction des positions de cinq types d'unités macrosyntaxiques.

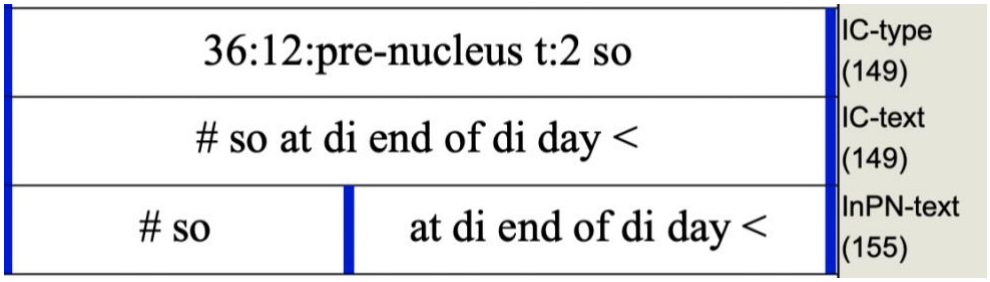
Unités illocutoires (UI) : la position de chaque UI, ou énoncé, est définie par une paire de tiers. Le premier tier, **IU-type**, fournit un identifiant numérique pour chaque énoncé et décrit son type (déclaratif, interrogatif, exclamatif, ou inachevé). Le deuxième tier, **IU-text** contient

le contenu textuel de l'unité illocutoire. L'unité illocutoire représente l'unité fondamentale de segmentation et d'annotation du corpus.

Parataxes et unités dissociés : Si une unité illocutoire contient une unité paratactique ou une unité dissociée, la position de chaque élément est délimitée par une paire de tiers. **In-IU-type** contient un identifiant pour chaque unité concernée, et In-IU-text contient son contenu textuel.

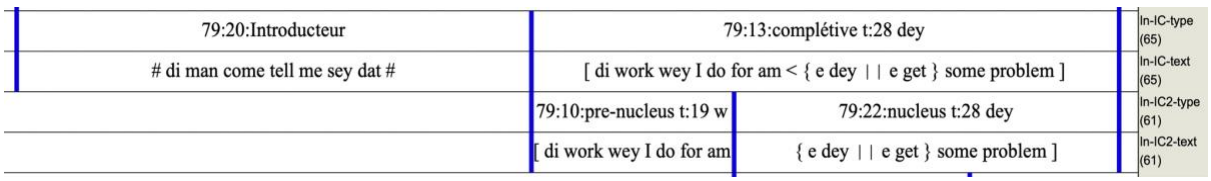
Unités macrosyntaxiques majeures : La position de chaque noyau, prénoyau, et postnoyau est délimitée par les tiers **IC-type** et **IC-text**. IC-type contient un identifiant unique pour chaque occurrence d'un élément de ces trois catégories, et le distingue selon son type. IC-text contient le contenu textuel de chaque unité concernée. Un troisième tier, **InPN-text** fournit une deuxième segmentation des prénoyaux visant à séparer les marqueurs de discours et les connecteurs comme *and*. Ce tier contient uniquement le contenu textuel de chaque segmentation. La figure 13 montre comment le prénoyau *# so at the end of di day <* est segmenté par ces trois tiers.

Figure 13 - Prénoyau représenté par trois tiers



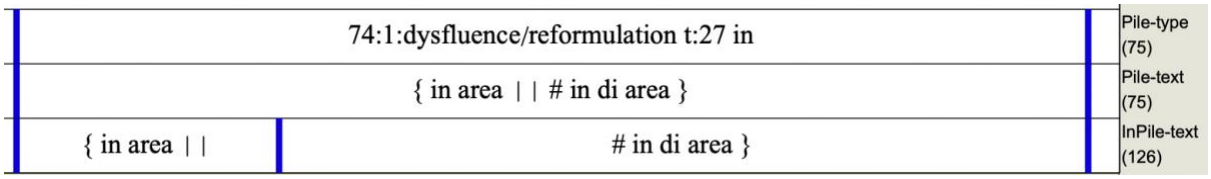
Unités enchâssées : Un quatrième groupe de tiers a été produit pour délimiter les occurrences de discours rapporté, de complétives, et de parenthèses. **In-IC-type** contient un identifiant unique pour chacun de ces types d'unités, et **In-IC-text** son contenu textuel. Dans le cas du discours rapporté et des complétives, les segments introduisant ces unités sont également délimités par ces tiers. Une deuxième paire de tiers, **In-IC2-type** et **In-IC2-text**, délimite les unités macrosyntaxiques majeures se trouvant à l'intérieur d'une unité enchassée. La figure 14 montre comment l'unité illocutoire *# di man come tell me sey dat # [di work wey I do for am < { e dey || e get } some problem]* est segmenté selon ces quatre tiers.

Figure 14 - Quatre tiers représentant une complétive et son introducteur



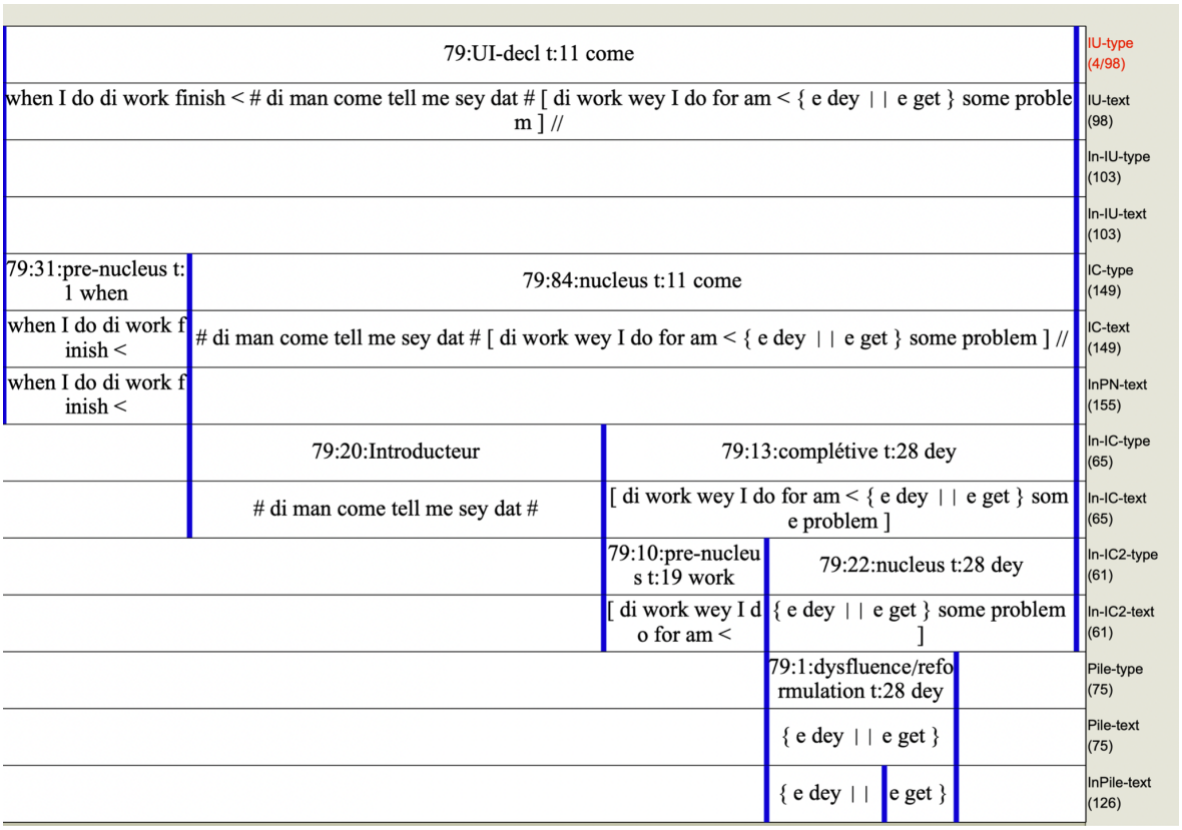
Piles : les reformulations, coordinations, appositions, et réduplications sont délimitées par les tiers **Pile-type**, **Pile-text**, et **InPile-text**. Les tiers Pile-type et Pile-texte délimitent chaque pile, et contiennent respectivement son identifiant et son contenu textuel. Le tier InPile-text sépare chaque élément de la pile et contient son contenu textuel. La figure 15 montre comment la pile *{ in area || # in di area }* est segmenté par ces trois tiers.

Figure 15 - Une pile représentée par quatre tiers



La figure 16 montre un énoncé segmenté selon les quatorze tiers.

Figure 16 - Un énoncé segmenté selon 14 tiers



Le script utilisé a également permis de générer une liste d'erreurs macrosyntaxiques possibles. Parmi les problèmes les plus fréquents, il y avait notamment des unités enchâssées pour lesquelles il manquait un crochet ouvrant ou fermant, des piles auxquelles il manquait des accolades délimitantes ou des signes macrosyntaxiques internes, et des énoncés auxquels il manquait des signes de fin d'unité illocutoire. Pour la plupart, ces problèmes ont

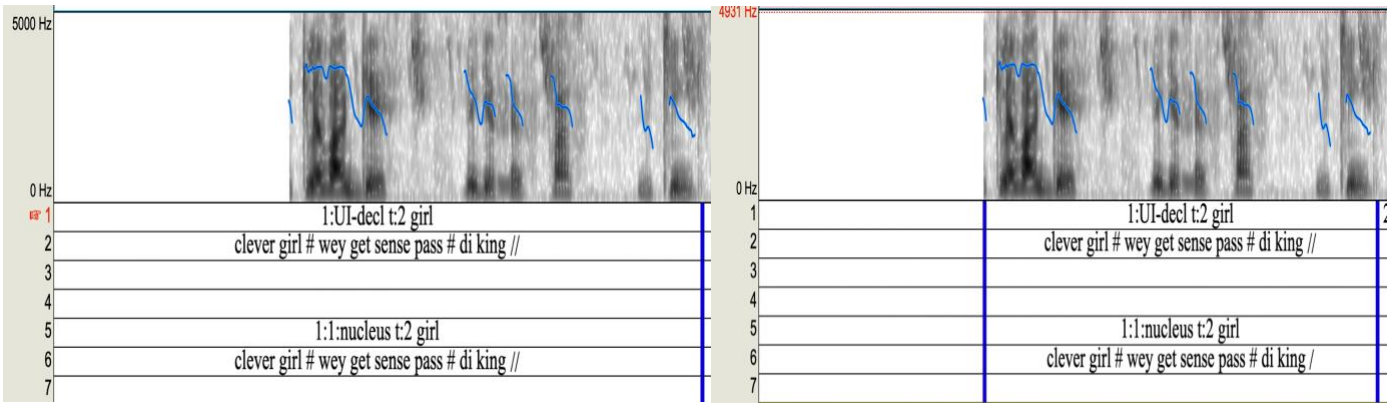
pu être réglés en insérant les signes macrosyntaxiques sur Arborator en amont de l'analyse prosodique.

Parfois, pour certains tokens ou unités illocutoires entières, il manquait également les informations temporelles permettant de générer ces alignements. Généralement, les informations manquantes avaient été vraisemblablement supprimées par accident lors des nombreuses interventions sur le corpus NaijaSynCor. Dans ces cas, les informations temporelles étaient facilement trouvables dans les anciennes versions du corpus accessibles sur GitHub. Plus rarement, les informations manquantes n'étaient pas récupérables. Il a ainsi été nécessaire d'estimer la longueur temporelle de chaque token en trouvant la différence entre la dernière valeur `AlignEnd` présente, et la prochaine valeur `AlignBegin`, et en divisant cette différence par le nombre de tokens (sauf signes macro) se trouvant entre ces deux valeurs. Des valeurs `AlignBegin` et `AlignEnd` respectant cette longueur estimée ont ensuite été ajoutées. Ces corrections ont non seulement permis d'obtenir des alignements beaucoup plus fiables, mais ont également donné l'occasion de corriger diverses erreurs qui avaient été ignorées dans le treebank.

4.2.2 Correction manuelle des textgrids

Après avoir corrigé les différentes erreurs restant dans le corpus et régénéré des alignements à partir du corpus corrigé, nous avons commencé un processus de vérification manuelle des alignements. Il s'agissait principalement d'ouvrir chaque TextGrid sur Praat et d'écouter le plus soigneusement possible chaque segment de chaque tier afin de vérifier que le fichier était correctement segmenté. La plupart des alignements ont été générés correctement, mais nous avons souvent légèrement modifié certains intervalles pour avoir un découpage encore plus précis. Ce type de vérification minutieuse ne pouvait pas être automatisé, ce qui en a fait la partie la plus longue de la préparation des données. Cependant, cette étape a permis de garantir que les patrons prosodiques soient calculés sur un ensemble de données hautement contrôlé. Une modification qui a été apportée systématiquement à chaque fichier a été de corriger le début de la première unité illocutoire. Dans chaque fichier `.conllu`, le début de la première unité illocutoire a été placé de manière à s'aligner avec le tout début de chaque fichier audio. Cependant, dans les fichiers audio utilisés, plusieurs secondes s'écoulaient avant que les locuteurs ne commencent à parler. Nous avons donc corrigé les premiers alignements de chaque fichier. Cette correction a également été effectuée à la fin de chaque fichier, bien que les problèmes sérieux d'alignement aient été rares.

Figure 17 - Début d'un fichier .TextGrid avant et après correction



L'inspection des alignements a également permis d'identifier d'autres erreurs dans les signes macro qui n'avaient pas été corrigés lors de la première série de corrections. Dans ces cas, le TextGrid a été corrigé manuellement, et les changements nécessaires ont été intégrés au fichier `.conllu` correspondant sur Arborator.

Un problème fréquemment rencontré lors de ces vérifications concernait le mot *because* lorsqu'il se trouvait dans un prénoyau. Dans le tier InPN-text, ce mot a été fréquemment séparé du reste du prénoyau, indiquant qu'il fonctionnait comme un marqueur de discours. Pendant l'inspection des données, nous avons remarqué que ce mot peut être utilisé de deux façons lorsqu'il apparaît dans un prénoyau, illustrées par les exemples suivants :

1. **because I still like eh broadcasting < I dey read news //**
   parce que j'aime toujours la radiodiffusion < je lis les actualités //

2. **because me < I dey read news //**
   parce que moi < je lis les actualités //

Dans le premier exemple, *because* fonctionne comme une conjonction de subordination, étant le gouverneur de la subordonnée *j'aime toujours la diffusion*. Dans le deuxième, il fonctionne comme un marqueur de discours, n'étant pas le gouverneur syntaxique de *me*. Dans ces cas, nous avons décidé de joindre *because* au segment suivant s'il était suivi d'une proposition, et de le garder séparé s'il était suivi d'un pronom ou d'un syntagme nominal. Cette décision était cohérente avec les annotations syntaxiques du treebank, où les différentes utilisations de *because* étaient annotées différemment.

Un dernier problème rencontré au cours de cette étape concernait un bug dans lequel le début d'un intervalle avait une valeur temporelle plus élevée que la fin. Dans l'interface graphique de Praat, ceci a rendu les fichiers impossibles à corriger manuellement.

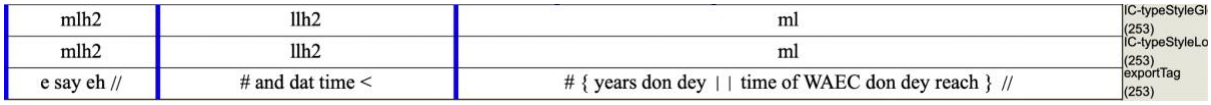
Une inspection approfondie des données n'a révélé aucune cause évidente : les traits `AlignBegin` et `AlignEnd` des unités illocutoires concernées semblaient correctes, et il n'y avait pas de points communs évidents entre les énoncés auxquels on avait attribué des alignements fautifs. Ce problème étant relativement rare, nous avons décidé d'éviter une refonte complète du code qui a généré les fichiers `.TextGrid`. La solution la plus efficace a été de créer une fonction qui itère sur chaque intervalle concerné. Si la fonction trouvait un intervalle dont le temps de début était antérieur à son temps de fin, le temps de début était remplacé par celui de la fin du segment précédent. Le temps de fin, quant à lui, a été remplacé par le temps de début du segment suivant. Cette approche était facile à intégrer et corrigeait généralement le problème sans nécessiter de modifications manuelles importantes des alignements. Dans les rares cas où le fichier `.TextGrid` résultant était mal aligné, le changement permettait néanmoins de modifier facilement le fichier à l'aide de l'interface graphique de Praat.

Plus tard, nous avons découvert un problème similaire dans quelques rares fichiers dans lesquels certains intervalles se chevauchaient. En d'autres termes, le début d'un segment apparaissait avant la fin du segment précédent. Ceci a également rendu les fichiers inopérables et impossibles à corriger via Praat. Étant donné la rareté relative de ce problème, la solution la plus efficace a été de modifier les intervalles en question directement dans le fichier `.TextGrid` à l'aide d'un éditeur de texte.

4.2.3 Génération des contours via SLAM+

Après la correction des fichiers `.TextGrid`, nous avons utilisé l'outil SLAM+ pour générer les patrons prosodiques des intervalles présents dans le tier IC-type, correspondant à l'identifiant de chaque prénoyau, noyau et postnoyau. Pour ce faire, le tier IC-type a été choisi comme targetTier, le tier IC-text comme tagTier, et IU-type comme speakerTier.[7] Ceci a permis de générer un fichier `.TextGrid` avec deux nouveaux tiers correspondant aux contours globaux et locaux de chaque segment.

Figure 18 - Nouveaux tiers générés par SLAM+

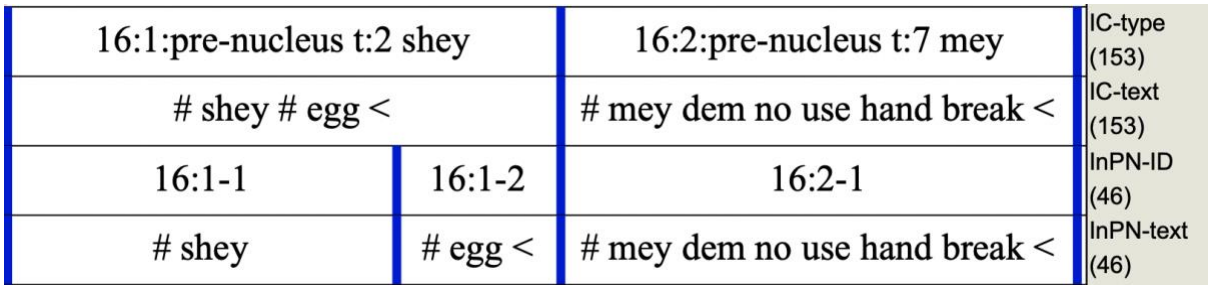


Nous avons appliqué le même processus au tier InIC-text, correspondant aux éléments pragmatiques internes de chaque prénoyau. Comme SLAM+ nécessite deux tiers avec une

[7] Ces termes ont été décrits précédemment dans la description de SLAM+ dans la section 2.4.2.2.

segmentation comparable pour fonctionner, nous avons dû élaborer un script pour modifier légèrement le fichier `.TextGrid` utilisé en entrée en ajoutant un nouveau tier InPN-ID contenant un identifiant unique pour chaque constituant d'un prénoyau. Cet identifiant est un code simple contenant le code numérique d'un prénoyau donné, suivi par un tiret et un chiffre pour chaque composant interne. Ainsi, pour un prénoyau 16:1 avec deux composants pragmatiques, le nouveau tier contient un segment 16:1-1 et un segment 16:1-2 délimitant les deux composants.

Figure 19 - Fichier .PitchTier modifié contenant un nouveau tier InPN-ID



Une fois cette modification terminée, nous avons appliqué SLAM+ deux fois au corpus, en utilisant à chaque fois un ensemble de variables distinct. Lors de la première itération, nous avons assigné aux variables targetTier et tagTier les tiers IC-type et IC-text. Cela a permis de générer un contour mélodique pour chaque unité macrosyntaxique majeure, dont les prénoyaux. Lors de la deuxième itération, nous avons assigné aux variables targetTier et tagTier les tiers InPN-ID et InPN-text, ce qui a permis de générer un contour mélodique pour chaque composant de chaque prénoyau. Le tier IU-type a été utilisé comme speakerTier pendant les deux itérations.

Cette étape a abouti à la création de deux nouveaux fichiers `.TextGrid` pour chaque fichier utilisé en entrée, soit un fichier pour chaque TargetTier. Les figures 20 et 21 montrent les contours prosodiques locaux et globaux générés pour le prénoyau *so now* <. Dans la figure 20, le contour a été généré à partir du tier 'IC-type', donc la totalité du prénoyau. Dans la figure 21, les contours ont été générés à partir du tier 'InPN-ID'. Il y a donc un tier pour chacun des deux composants.

Figure 20 - Contours générés à partir du tier IC-type

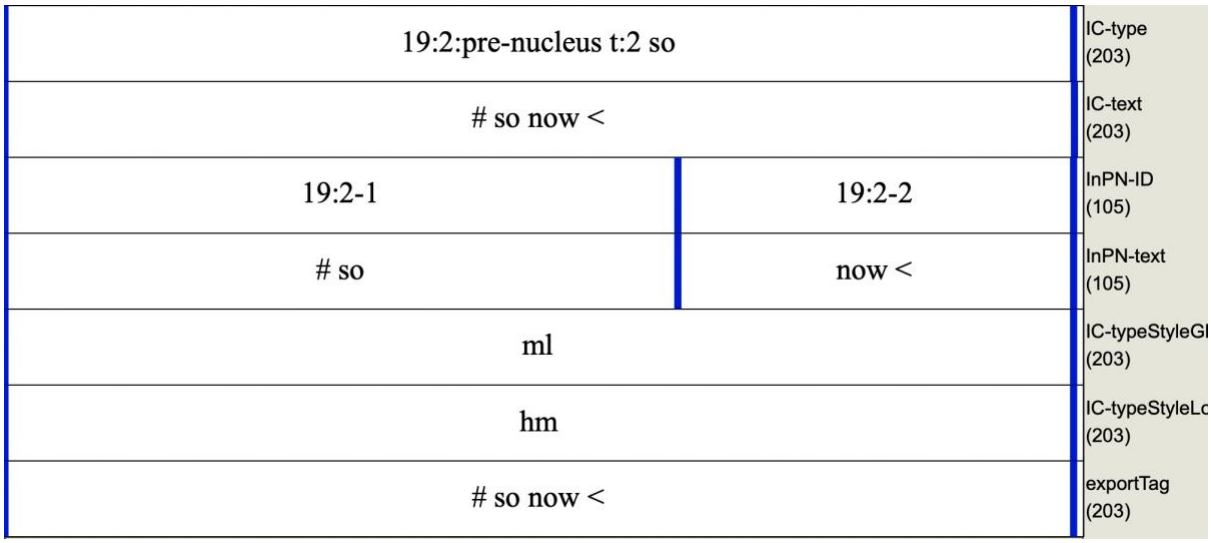
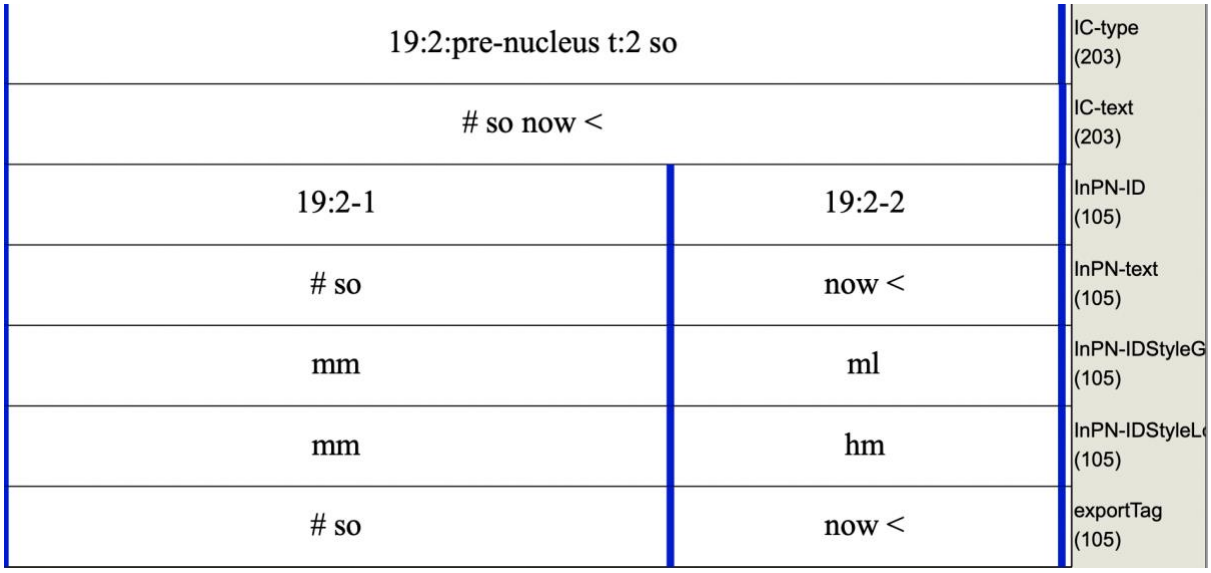
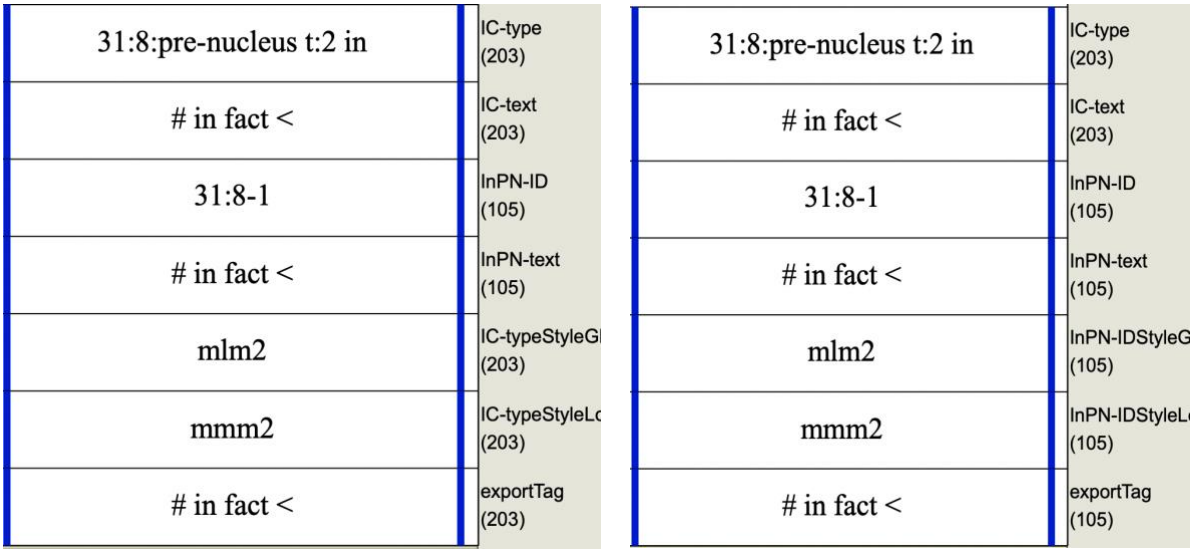


Figure 21 - Contours générés à partir du tier InPN-ID



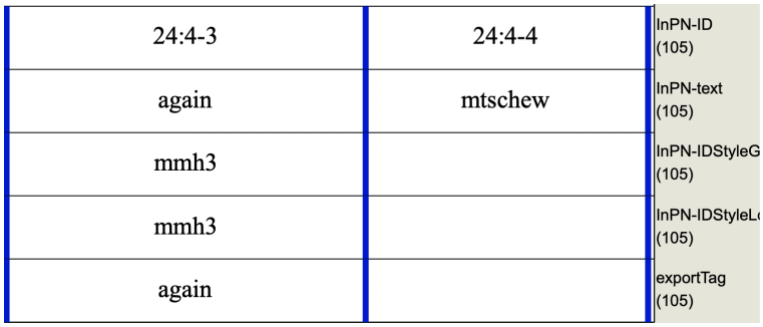
Notez que de nombreux prénoyaux dans ce corpus ne contiennent qu'un seul composant. Dans ces cas, chacun des deux fichiers contient le même contour mélodique, puisque la segmentation du tier InPN-ID est identique à la segmentation du tier IC-type.

Figure 22 - Contours identiques produits à partir d'un prénoyau simple



Notez également que le SLAM+ n'a parfois pas réussi à générer un contour pour quelques rares segments. Typiquement, ces segments correspondent à de très courts laps de temps, ou à des sections de l'enregistrement de très mauvaise qualité audio. Dans ces conditions, le SLAM+ ne disposait pas d'informations prosodiques suffisantes pour générer un contour mélodique. Très souvent, ces segments correspondent à l'interjection *mtchew*, un clic sourd utilisé comme marqueur de discours. Dans ces cas, les tiers prosodiques générés par SLAM+ étaient simplement vides, comme montre la figure 23.

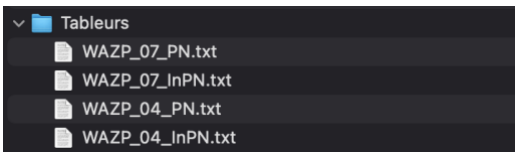
Figure 23 - Tiers prosodiques vides produits à partir de l'interjection mtchew



Après avoir créé les fichiers `.TextGrid` contenant les contours mélodiques, l'étape suivante a consisté à convertir les données dans un format plus utilisable. Pour cela, chaque fichier `.TextGrid` a été ouvert avec le logiciel d'analyse prosodique ANALOR, qui offre la possibilité d'exporter les données dans un fichier `.tsv` qui peut être manipulé avec un tableur comme Excel. Pour chaque fichier `.TextGrid`, nous avons exporté un fichier `.tsv` contenant les informations contenues dans cinq tiers : les trois nouveaux tiers générés par SLAM+, et les deux tiers utilisés comme TargetTier et TagTier.

Chaque fichier exporté a été nommé avec le nom du fichier correspondant, suivi d'un ou deux suffixes en fonction du tier utilisé pour calculer les contours mélodiques. Les fichiers contenant un contour pour chaque composant de chaque prénoyau ont été distingués par le suffixe _InPN, et les fichiers contenant les contours pour chaque unité macrosyntaxique majeure ont reçu le suffixe _PN. Les deux fichiers associés au fichier ABJ_GWA_03 sont donc ABJ_GWA_03_PN et ABJ_GWA_03_InPN. La totalité des 160 fichiers générés ont été rangés dans un même répertoire.

Figure 24 - Tableurs rangés dans un même répertoire



Enfin, un script Python a été élaboré pour combiner chaque paire de fichiers en un seul fichier `.tsv`. Le script a converti les deux fichiers en une structure de *dataframes*, a harmonisé les noms des colonnes de ces *dataframes*, puis les a concaténés. S'ensuit le tri des données selon les informations temporelles pour finalement générer un nouveau fichier `.tsv` contenant les informations combinées. En plus d'avoir toutes les informations prosodiques liées à un fichier donné dans un seul fichier, cette approche permet également de visualiser chaque segment, son contenu textuel, et ses contours prosodiques en un format convivial et facile à comprendre. Dans la figure 25 ci-dessous, on voit notamment que le prénoyau 17:7 est suivi directement par ses composants.

Figure 25 - Visualisation du fichier .tsv combiné

	1	2	3	4	5	6	7	8	9
30	38.907	40.491	12:14:nucleus t:6 like	I no like sciences //	mlh1	mlh1	I no like sciences //		
31	40.491	42.918	13:15:nucleus t:4 write	# so I write my WAEC //	hmh2	hmh2	# so I write my WAEC //		
32	42.918	44.561	14:16:nucleus t:3 pass	I no pass all di subjects //	ml	ml	I no pass all di subjects //		
33	44.561	46.698	15:17:nucleus t:3 con	# she con say [okay make I just do pre-degree first] //	mlh1	mlh1	# she con say [okay make I just do pr		
34	46.698	50.024	16:18:nucleus t:4 go	# den I go rewrite again till I must study di medicine //	mmh3	mmh3	# den I go rewrite again till I must stu		
35	50.024	52.32	17:7:pre-nucleus t:2 but	# but as I enter pre-degree do library science <	mlH1	mlH1	# but as I enter pre-degree do library		
36	50.024	50.692	17:7-1	# but	mm	mm	# but		
37	50.692	52.32	17:7-2	as I enter pre-degree do library science <	mlH1	mlH1	as I enter pre-degree do library scien		
38	52.32	55.061	17:19:nucleus t:12 tell	I tell mysef sey [noting go make me go back to sciences] //	llH2	llH2	I tell mysef sey [noting go make me		
39	55.061	57.506	18:20:nucleus t:3 na	# so na wetin make me study library science for school //	hl	hl	# so na wetin make me study library		
40	57.506	59.27	19:21:nucleus t:9 do	# and { I || I } don't do bad //+	mlH2	mLH2	# and { I || I } don't do bad //+		
41	59.27	59.687	19:22:nucleus t:13 for	# for school //	ll	mm	# for school //		

Cette structure de données a permis de faire une dernière vérification manuelle du découpage de chaque prénoyau en composants internes, et de détecter quelques marqueurs de discours qui n'étaient pas correctement isolés. Dans ces cas, nous avons corrigé manuellement les fichiers `.TextGrid` utilisés en entrée sur Praat. Si le problème était causé par une erreur d'annotation syntaxique, nous avons également corrigé l'erreur sur Arborator.

Ce processus de vérification a également permis d'identifier un problème avec le fichier `.PitchTier` associé au fichier BEN_08. En inspectant le fichier `.tsv` contenant les

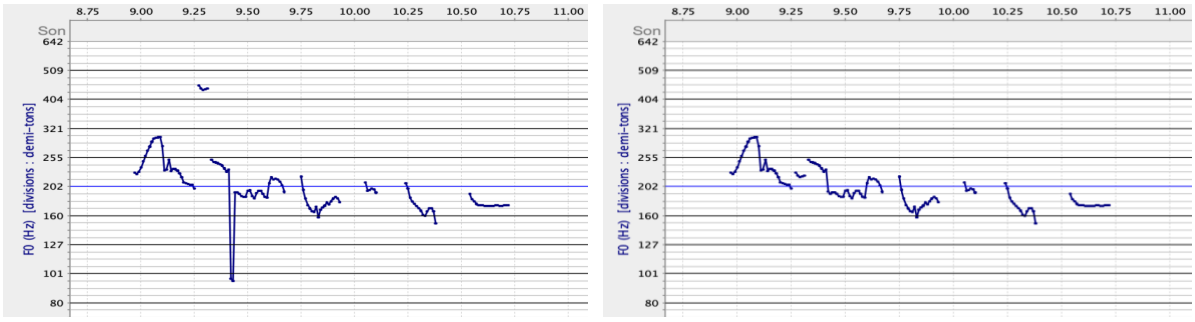
contours prosodiques associés à ce dernier, nous avons remarqué qu'aucun contour prosodique n'avait été généré pour les segments commençant après 297 secondes.

Figure 26 - Contours prosodiques manquants dans le fichier BEN_08

295.031	296.324	77:50:pre-nucleus t:16 if	# if na fresh fish <	hl	ml	# if na fresh fish <
295.031	296.324	77:50-1	# if na fresh fish <	hl	ml	# if na fresh fish <
296.324	297.704	77:51:pre-nucleus t:23 if	# { if na || if na ponmo } <	ml	mm	# { if na || if na ponmo } <
296.324	297.704	77:51-1	# { if na || if na ponmo } <	ml	mm	# { if na || if na ponmo } <
297.704	302.542	77:84:nucleus t:33 go	# e go even good because # before ponmo go even done # for fire < # di soup go don # thick //			
302.542	303.571	78:52:pre-nucleus t:2 but	# but if na meat <			
302.572	302.921	78:52-1	# but			
302.921	303.571	78:52-2	if na meat <			
303.571	304.117	78:85:nucleus t:9 good	e still good //			
304.117	305.447	79:53:pre-nucleus t:2 if	# if na fresh fish <			
304.117	305.447	79:53-1	# if na fresh fish <			
305.447	309.678	79:86:nucleus t:9 go	# you go know sey [{ di ban~ || di soup } go don thick small before you put your fresh fish] //			

En inspectant le fichier `BEN_08.PitchTier`, nous avons remarqué que le fichier ne contenait pas d'informations prosodiques au-delà des 297 secondes, ce qui laisse penser qu'il s'agissait d'un fichier mal nommé. Pour régler ce problème rapidement, nous avons généré un nouveau fichier `.PitchTier` à partir du fichier audio `BEN_08.wav` à l'aide de Praat. Les informations sur les fréquences fondamentales contenues dans ce fichier ayant été extraites automatiquement, il a été nécessaire de corriger manuellement le fichier à l'aide du logiciel ANALOR. Il s'agissait essentiellement de supprimer les segments de fréquence fondamentale qui ne semblaient pas correspondre à la véritable fréquence fondamentale du locuteur, comme ceux causés par le bruit de fond. Les contours de F0 présentant des changements brusques et extrêmes de hauteur ont également été considérés comme suspects. Dans ces cas, nous avons écouté attentivement les segments audio correspondants pour voir si ces modulations correspondaient à des irrégularités audibles de la fréquence fondamentale du locuteur. Dans de nombreux cas, les sauts soudains dans la fréquence fondamentale détectée semblaient être causés par la fricative [s], qui produit parfois un sifflement aigu qui avait été apparemment confondu avec la fréquence fondamentale du locuteur. Dans les cas où le contour F0 extrait ne semblait pas correspondre à la hauteur du locuteur entendue dans le fichier audio, les segments incriminés ont été supprimés ou ajustés.

Figure 27 - Un segment de BEN_08.PitchTier avant et après correction manuel



Une fois le nouveau fichier `.PitchTier` généré, nous avons appliqué SLAM+ de nouveau sur les données pour générer les contours manquants.

4.2.4 Annotations pragmatiques

Une fois les fichiers `.tsv` finaux générés, les informations ont été fusionnées dans une feuille de calcul contenant les annotations pragmatiques pour chaque prénoyau et ses composants.

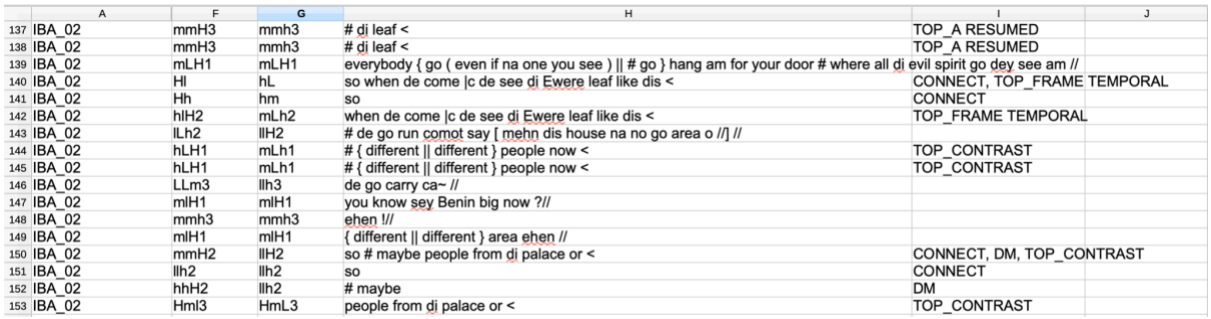
Avant que ce mémoire de recherche ne soit commencé, un ensemble d'annotations pragmatiques avait déjà été produit par Candide Simard et Anne Lacheret-Dujour pour les prénoyaux dans 23 monologues. Cependant, ce premier travail présentait plusieurs problèmes. Tout d'abord, ce travail a été effectué sur une version antérieure du corpus NaijaSynCor, qui a subi des changements importants au cours de l'année dernière. De ce fait, les annotations macro-syntaxiques ont parfois été modifiées. Certains segments précédemment annotés comme des prénoyaux n'étaient plus annotés comme tels, et vice-versa. Les annotations pragmatiques précédentes ont donc dû être modifiées.

Deuxièmement, le script utilisé pour segmenter les prénoyaux en composants pragmatiques internes a parfois produit une segmentation différente de celle qui avait été utilisée lors de cette première tentative de produire des annotations pragmatiques. Après avoir comparé les deux segmentations, nous avons décidé de conserver celle qui avait été produite dans le cadre de ce mémoire de recherche, et de modifier les annotations précédentes en fonction des différences de segmentation.

Enfin, les annotations pragmatiques originelles avaient été produites par des chercheuses qui ne connaissaient pas la langue et qui s'étaient donc largement appuyées sur la ressemblance du naija avec l'anglais pour interpréter les énoncés. De ce fait, de nombreuses étiquettes pragmatiques étaient erronées en raison d'une mauvaise interprétation sémantique de la part des annotatrices.

Afin d'assurer la qualité des données, nous avons donc décidé de vérifier chaque annotation pragmatique et de corriger les étiquettes si nécessaire. En cas de doute, nous avons collaboré avec Candide Simard pour mieux comprendre son raisonnement dans le choix de certaines annotations pragmatiques.

Figure 28 - Fichier contenant des annotations pragmatiques et prosodiques pour chaque segment



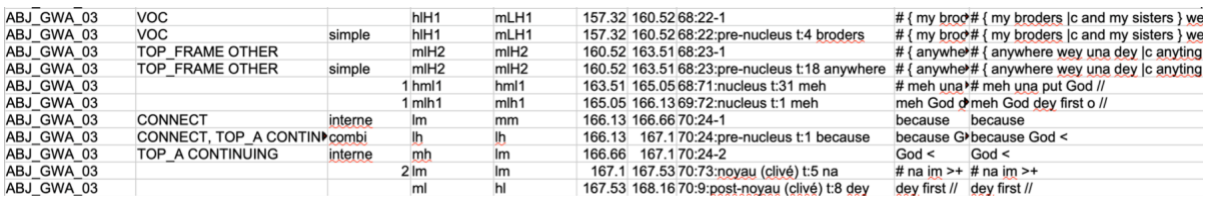
	A	F	G	H	I	J
137	IBA_02	mmH3	mmh3	# di leaf <	TOP_A RESUMED	
138	IBA_02	mmH3	mmh3	# di leaf <	TOP_A RESUMED	
139	IBA_02	mLH1	mLH1	everybody { go (even if na one you see) || # go } hang am for your door # where all di evil spirit go dey see am //		
140	IBA_02	Hl	hL	so when de come |c de see di Ewere leaf like dis <	CONNECT, TOP_FRAME TEMPORAL	
141	IBA_02	Hh	hm	so	CONNECT	
142	IBA_02	hlH2	mLh2	when de come |c de see di Ewere leaf like dis <	TOP_FRAME TEMPORAL	
143	IBA_02	lLh2	llH2	# de go run comot say [mehn dis house na no go area o //] //		
144	IBA_02	hLH1	mLh1	# { different || different } people now <	TOP_CONTRAST	
145	IBA_02	hLH1	mLh1	# { different || different } people now <	TOP_CONTRAST	
146	IBA_02	LLm3	llh3	de go carry ca~ //		
147	IBA_02	mlH1	mlH1	you know sey Benin big now ?//		
148	IBA_02	mmh3	mmh3	ehen !//		
149	IBA_02	mlH1	mlH1	{ different || different } area ehen //		
150	IBA_02	mmH2	llH2	so # maybe people from di palace or <	CONNECT, DM, TOP_CONTRAST	
151	IBA_02	llh2	llh2	so	CONNECT	
152	IBA_02	hhH2	llh2	# maybe	DM	
153	IBA_02	Hml3	HmL3	people from di palace or <	TOP_CONTRAST	

4.2.5 Analyse statistique

Une fois les annotations pragmatiques terminées et les courbes extraites, nous sommes passé à notre phase d'analyse statistique. Nous avons commencé par générer un nouveau tableau contenant toutes les données collectées jusqu'à présent, en ajoutant une colonne supplémentaire indiquant la catégorie syntaxique de chaque unité macrosyntaxique. Cette colonne divise les noyaux en deux sous-catégories : les noyaux faisant partie d'une construction clivée (annotés avec l'étiquette '1'), et les noyaux ne faisant pas partie d'une construction clivée (annotés avec l'étiquette '2'). Nous avons également divisé l'ensemble des prénoyaux en deux groupes principaux : les prénoyaux simples et les prénoyaux complexes. Les prénoyaux simples sont ceux qui ne contiennent qu'une seule fonction pragmatique, c'est-à-dire les prénoyaux qui ne contiennent aucune segmentation interne. Les prénoyaux complexes sont ceux qui contiennent deux ou plusieurs unités pragmatiques internes. Chaque unité pragmatique se trouvant à l'intérieur d'un prénoyau complexe a été catégorisée dans la nouvelle colonne comme une unité 'interne'. Notez que notre méthode de production des données a pour résultat deux lignes pour chaque prénoyau simple, chacune contenant des données identiques. En d'autres termes, les prénoyaux simples sont apparus dans notre ensemble de données comme des prénoyaux avec exactement un composant pragmatique interne. Pour éviter que des données redondantes n'entrent dans nos calculs, les prénoyaux ont été annotés dans la nouvelle colonne avec l'étiquette 'simple', et leurs composants internes n'ont pas été annotés.

Notre corpus final contenait 555 prénoyaux 'simples', 183 prénoyaux 'complexes', et 368 composants pragmatiques 'internes'.

Figure 29 - Extrait de jeu de données classifiant unités syntaxiques par catégorie



ABJ_GWA_03	VOC		hlH1	mLH1	157.32	160.52	68:22-1	# { my broc # { my broders |c and my sisters } we
ABJ_GWA_03	VOC	simple	hlH1	mLH1	157.32	160.52	68:22:pre-nucleus t:4 broders	# { my broc # { my broders |c and my sisters } we
ABJ_GWA_03	TOP_FRAME OTHER		mlH2	mlH2	160.52	163.51	68:23-1	# { anywhe # { anywhere wey una dey |c anyting
ABJ_GWA_03	TOP_FRAME OTHER	simple	mlH2	mlH2	160.52	163.51	68:23:pre-nucleus t:18 anywhere	# { anywhe # { anywhere wey una dey |c anyting
ABJ_GWA_03		1	hml1	hml1	163.51	165.05	68:71:nucleus t:31 meh	# meh una # meh una put God //
ABJ_GWA_03		1	mlh1	mlh1	165.05	166.13	69:72:nucleus t:1 meh	meh God d meh God dey first o //
ABJ_GWA_03	CONNECT	interne	lm	mm	166.13	166.66	70:24-1	because because
ABJ_GWA_03	CONNECT, TOP_A CONTIN	combi	lh	lh	166.13	167.1	70:24:pre-nucleus t:1 because	because G because God <
ABJ_GWA_03	TOP_A CONTINUING	interne	mh	lm	166.66	167.1	70:24-2	God < God <
ABJ_GWA_03		2	lm	lm	167.1	167.53	70:73:noyau (clivé) t:5 na	# na im >+ # na im >+
ABJ_GWA_03			ml	hl	167.53	168.16	70:9:post-noyau (clivé) t:8 dey	dey first // dey first //

Au cours de cette phase du projet, nous avons remarqué qu'un grand nombre d'étiquettes prosodiques n'apparaissait qu'un très petit nombre de fois. Pour chaque niveau d'analyse (prénoyaux simples, prénoyaux complexes, ou composants internes), nous avons décidé de générer deux grandes catégories de graphiques ACM : ceux qui prennent en compte tous les contours présents, et ceux qui ne prennent en compte que les contours les plus fréquents. Cette dernière approche a été réalisée en classant les contours par nombre d'occurrences, et en incluant ceux dont le nombre combiné d'occurrences dépassait 50% du nombre total d'occurrences dans le corpus pour le niveau d'analyse donné.

Pour chaque niveau d'analyse, et pour chaque champ d'application des données utilisées (tous les contours pertinents ou seulement les plus communs), une gamme de graphiques ACM a été produite en tenant compte de trois variables additionnelles : si les contours ou les étiquettes pragmatiques ont été utilisés comme variable primaire, si les calculs ont été fait à partir des contours globaux ou locaux, et quels types pragmatiques ont été inclus dans les calculs. Cette dernière variable ne concerne que les niveaux d'analyse des prénoyaux simples et des composants pragmatiques internes. Pour ces derniers, nous avons produit un ensemble de graphiques ACM excluant toutes les catégories pragmatiques sauf les *aboutness topics*, et les topics contrastifs (i.e., gardant un maximum de 5 étiquettes).

Le tableau 2 présente 32 fichiers produits selon ces critères, utilisant comme niveau d'analyse le composant pragmatique interne. Les noms de fichiers en **gras** représentent les fichiers dans lesquels les variables primaires et secondaires sont visibles, tandis que les fichiers en *italique* représentent ceux dans lesquels les variables secondaires sont cachées. Les noms de fichiers en noir sont les fichiers dans lesquels toutes les étiquettes pragmatiques ont été incluses dans les calculs, tandis que les noms de fichiers en bleu représentent les fichiers dans lesquels seul le sous-ensemble de 5 étiquettes pragmatiques (*aboutness topics* et topics contrastifs) a été utilisé dans les calculs. Un ensemble semblable de 32 fichiers a également été produit en utilisant les prénoyaux simples comme niveau d'analyse. Pour les prénoyaux complexes, nous avons décidé de ne pas produire de fichiers utilisant un sous-ensemble d'étiquettes pragmatiques. Cela est dû au fait que les annotations pragmatiques produites à ce niveau d'analyse sont des étiquettes composites de type "*DM, TOP_A NEW*". Il serait donc impossible de générer des graphiques équivalents contenant les *aboutness topics* et les topics contrastifs isolés. Pour visualiser les types de fichiers produits pour les prénoyaux complexes, il faut simplement imaginer le tableau 2, mais sans les fichiers en bleu. 80 graphiques ACM ont été produits au total.

Tableau 2 – 32 graphiques ACM générés au niveau des composants pragmatiques internes

Var. primaire	*Var. secondaire*	*Nom du fichier*	
		100% des contours	Contours les plus communs
Type pragmatique	Contour local	**ptloc** *pt_generedepuisloc* **Listeptloc** *Listept_generedepuisloc*	**50ptloc** *50pt_genereloc* **50ptloc** *Liste50pt_generedepuisloc*
	Contour global	**ptglo** *pt_generedepuisglo* **Listeptglo** *Listept_generedepuisglo*	**50ptglo** *50pt_generedepuisglo* **Liste50ptglo** *Liste50pt_generedepuisglo*
Contour global	Type pragmatique	**glopt** *glo* **Listeglopt** *Listeglo*	**50glopt** *50glo* **Liste50glopt** *Liste50glo*
Contour local	Type pragmatique	**Locpt** *loc* **Listelocpt** *Listeloc*	**50locpt** *50loc* **Liste50locpt** *Liste50loc*

4.2.6 Tests statistiques supplémentaires

Après avoir appliqué ces méthodes et analysé les résultats décrits dans la section suivante, nous avons décidé d'approfondir nos analyses en utilisant deux tests supplémentaires afin d'identifier des correspondances entre les variables prosodiques et pragmatiques. Au lieu de chercher des correspondances entre les types pragmatiques et les contours SLAM+ eux-mêmes, nous avons décidé de décrire chaque contour à l'aide de six traits correspondant à certaines caractéristiques de base, comme la position de sa saillance interne. Deux traits additionnels ont été ajoutés pour décrire chaque segment étudié en termes de longueur temporelle et syllabique. Chaque trait que nous avons défini possède un petit nombre de valeurs possibles, dont l'une est attribuée à chaque unité pragmatique en fonction de la nature du contour global qui lui est associé. Cette tâche a été réalisée automatiquement à l'aide d'un script Python que nous avons développé pour analyser chaque topic et calculer les étiquettes à partir de son contour global, ses alignements temporels, ou son contenu textuel.

Huit traits ont été développés au total, pour un total combiné de 30 valeurs possibles. Elles sont énumérées ci-dessous.

- **Amplitude_endpoints**, représentant la différence de hauteur entre le début et la fin du segment. Il s'agit d'une mesure de l'amplitude, ou de variation de la F0 à l'intérieur du segment. Les valeurs possibles sont **AmpEndpoints_Low**, **AmpEndpoints_Mid** et **AmpEndpoints_High**.
- **Amplitude_overall**, une deuxième mesure de l'amplitude. Celle-ci représente la différence entre la hauteur la plus élevée du contour et la hauteur la plus basse, tenant en compte la valeur de la saillance interne. Les valeurs possibles sont **AmpOverall_Low**, **AmpOverall _Mid** et **AmpOverall _High**.
- **Position_saliency**, décrivant la position de la saillance interne d'un contour. Les valeurs possibles sont **SaliencyBegin**, **SaliencyMiddle**, **SaliencyEnd**. Une quatrième valeur, **SaliencyNull**, existe pour les contours manquant une saillance interne.
- **Curve**, représentant la courbe du contour. Les valeurs sont **CurveConvex** (saillance plus élevée que le début et la fin), **CurveConcav** (saillance plus basse que le début et la fin), **CurveIntermed** (saillance entre le début et la fin), et **CurveNull** (aucune saillance).
- **Slope,** représentant la direction du contour. Les valeurs sont **SlopeRising** (hauteur finale plus élevée que la valeur initiale), **SlopeFalling** (hauteur de fin plus basse que celle du début), et **SlopeFlat** (hauteur de fin égale à celle du début).
- **Average_height**, représentant la moyenne des hauteurs encodées dans le contour. Les valeurs possibles sont **AvgHeight_Low**, **AvgHeight_Mid**, et **AvgHeight_High**.
- **Duration**, représentant la durée relative de chaque segment. Les valeurs possibles sont **Duration_Low**, **Duration_Mid**, et **Duration_High**. Il s'agit de l'un des deux traits prosodiques générés à partir de données situées en dehors des contours SLAM+. Bien que les valeurs soient relatives à la longueur moyenne de tous les segments majeurs de notre corpus (y compris les noyaux et les prénoyaux), chaque étiquette possible est représentée dans les prénoyaux.
- **Syllables**, représentant les nombres de syllabes dans le segment étudié. Il y a sept valeurs possibles allant de **1_syl** pour les unités monosyllabiques, à **7+_syl** pour les unités de sept syllabes ou plus. Ces valeurs ont été calculées à l'aide d'une fonction qui génère une représentation phonétique de chaque mot du segment en utilisant le *Carnegie Mellon Pronouncing Dictionary* (`cmudict`) et compte le nombre de voyelles. Cela a permis un comptage précis des voyelles dans les mots d'origine anglaise, y compris ceux dont l'orthographie diverge fortement de la prononciation. Pour les mots non reconnus par le dictionnaire, un nombre estimé de syllabes a été obtenu en comptant les groupes de voyelles dans chaque mot.

Pour chaque segment de notre corpus, ces traits ont été calculés et assemblés dans un nouveau jeu de données. La figure 30 ci-dessous montre les étiquettes prosodiques associées à un ensemble de topics.

Figure 30 – Jeux de données contenant les traits prosodiques

Pragmatic_type	StyleGlo	Syllables	Amplitude_endpoints	Amplitude_overall	Position saliency	Curve	Average_height	Slope	Duration
TOP_CONTRAST	mlh2	2_syl	AmpEndpoints_Low	AmpOverall_Mid	SaliencyMiddle	CurveConvex	AvgHeight_Mid	SlopeFalling	Duration_Mid
TOP_A RESUMED	mmh1	4_syl	AmpEndpoints_Low	AmpOverall_Low	SaliencyBegin	CurveConvex	AvgHeight_Mid	SlopeFlat	Duration_Mid
TOP_A CONTINUING	hlH1	7+_syl	AmpEndpoints_Mid	AmpOverall_High	SaliencyBegin	CurveConvex	AvgHeight_Mid	SlopeFalling	Duration_Mid
TOP_A RESUMED	lmh2	2_syl	AmpEndpoints_Low	AmpOverall_Mid	SaliencyMiddle	CurveConvex	AvgHeight_Mid	SlopeRising	Duration_Mid
TOP_CONTRAST	mmh1	4_syl	AmpEndpoints_Low	AmpOverall_Low	SaliencyBegin	CurveConvex	AvgHeight_Mid	SlopeFlat	Duration_Mid
TOP_CONTRAST	Lm	1_syl	AmpEndpoints_Mid	AmpOverall_Mid	SaliencyNull	CurveNull	AvgHeight_Low	SlopeRising	Duration_Low
TOP_CONTRAST	mlH1	5_syl	AmpEndpoints_Low	AmpOverall_High	SaliencyBegin	CurveConvex	AvgHeight_Mid	SlopeFalling	Duration_High
TOP_A RESUMED	mmh2	3_syl	AmpEndpoints_Low	AmpOverall_Low	SaliencyMiddle	CurveConvex	AvgHeight_Mid	SlopeFlat	Duration_Mid
TOP_A NEW	mlH1	5_syl	AmpEndpoints_Low	AmpOverall_High	SaliencyBegin	CurveConvex	AvgHeight_Mid	SlopeFalling	Duration_Mid

Nous avons ensuite utilisé ce jeu de données pour générer une paire de graphiques ACM représentant certaines étiquettes pragmatiques et leurs caractéristiques prosodiques associées : une pour les prénoyaux simples et une pour les composantes internes des prénoyaux complexes. Nous avons limité la portée de cette étude à cinq étiquettes pragmatiques représentant les quatre aboutness topics en plus des topics contrastifs. Pour les raisons décrites dans la section suivante, les graphiques ACM que nous avions précédemment générés en utilisant uniquement ces topics ont donné les résultats de loin les plus intéressants. Nous avons donc décidé de concentrer le reste de notre recherche sur ce sous-ensemble de notre corpus.

Ces graphiques comportaient un total de neuf variables primaires : les huit nouveaux traits prosodiques ainsi que les étiquettes pragmatiques. Les contours originaux de SLAM+ n'ont pas été pris en compte. L'interprétation des graphiques ACM étant largement basée sur les impressions, nous avons également décidé d'appliquer le test exact de Fisher à chaque paire de valeurs prosodiques et pragmatiques.

Pour ce faire, un script Python a été développé pour parcourir ce sous-ensemble du corpus et, pour chaque paire de valeurs, calculer un tableau de contingence basé sur le modèle suivant.

Tableau 3 – Modèle de tableau de contingence

	Unités assignées l'etiquette Y	Unités sans l'étiquette Y
Unités assignées l'étiquette X	*a*	*b*
Unités sans l'étiquette X	*c*	*d*

La bibliothèque `scipy` a ensuite été utilisée pour appliquer le test exact de Fisher sur le tableau de contingence. Pour chaque niveau d'analyse (topics en prénoyaux simples vs complexes), les p-values résultantes ont été triées par ordre croissant afin d'obtenir facilement les étiquettes les plus significativement corrélées. Une paire de tableaux a également été

produite pour afficher les 150 p-values de chaque paire possible entre une étiquette pragmatique et une étiquette prosodique. Ce dernier tableau pouvait être utilisé pour identifier les corrélations prosodiques les plus fortes de chaque catégorie de topic, tandis que la liste plus complète était destinée à identifier toutes les variables prosodiques qui ont tendance à être fortement corrélées.

5 Résultats et analyses

5.1 Introduction et rappel des hypothèses

Dans cette section, nous présenterons les principaux résultats des analyses qualitatives et quantitatives de notre corpus. Nous rappelons que ces résultats seront analysés dans le but de valider ou d'invalider nos trois hypothèses :

1. Qu'il existe des différences prosodiques statistiquement observables entre les différents types pragmatiques.
2. Qu'il est possible d'identifier les contours prosodiques qui sont le plus fortement associés à certains types pragmatiques.
3. Que les locuteurs sont plus susceptibles de marquer prosodiquement les topics qu'ils supposent moins accessibles dans la représentation mentale de leurs interlocuteurs.

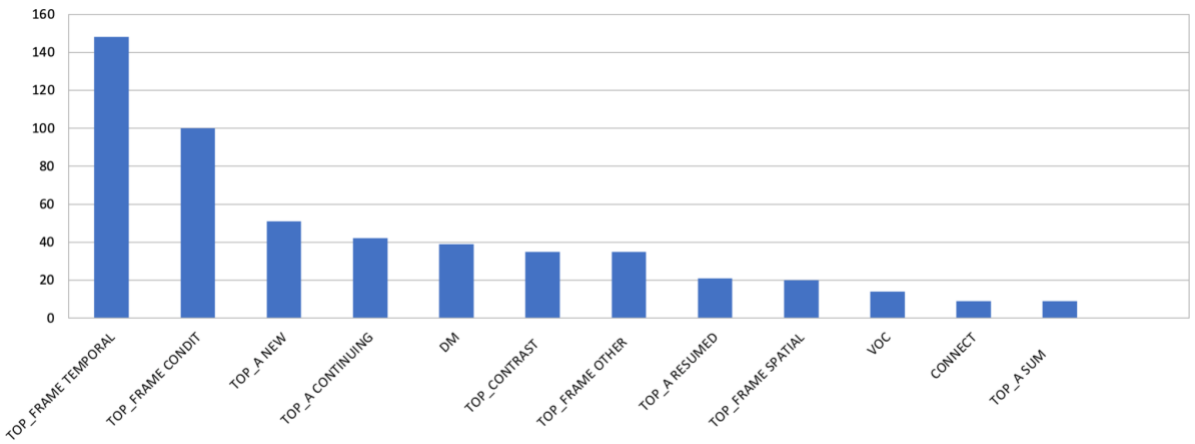
Avant de nous plonger dans nos résultats statistiques et leurs implications par rapport à nos hypothèses, nous allons commencer cette section par une brève comparaison quantitative des différents types pragmatiques présents dans nos trois niveaux d'analyse : les prénoyaux simples, les prénoyaux complexes et les constituants des prénoyaux complexes. Cela permettra avant tout d'avoir un panorama général du contenu pragmatique de chacun des trois types d'unités. Dans un deuxième temps, cela nous permettra de mettre en évidence et de commenter certaines différences fonctionnelles découvertes entre ces trois niveaux d'analyse.

5.2 Contenu pragmatique des niveaux d'analyse

Dans cette section, nous allons donner un aperçu de la fréquence des différents types pragmatiques dans les trois niveaux d'analyse utilisés dans cette étude : les prénoyaux simples, les prénoyaux complexes et les constituants pragmatiques internes des prénoyaux complexes.

5.2.1 Contenu pragmatique des prénoyaux simples

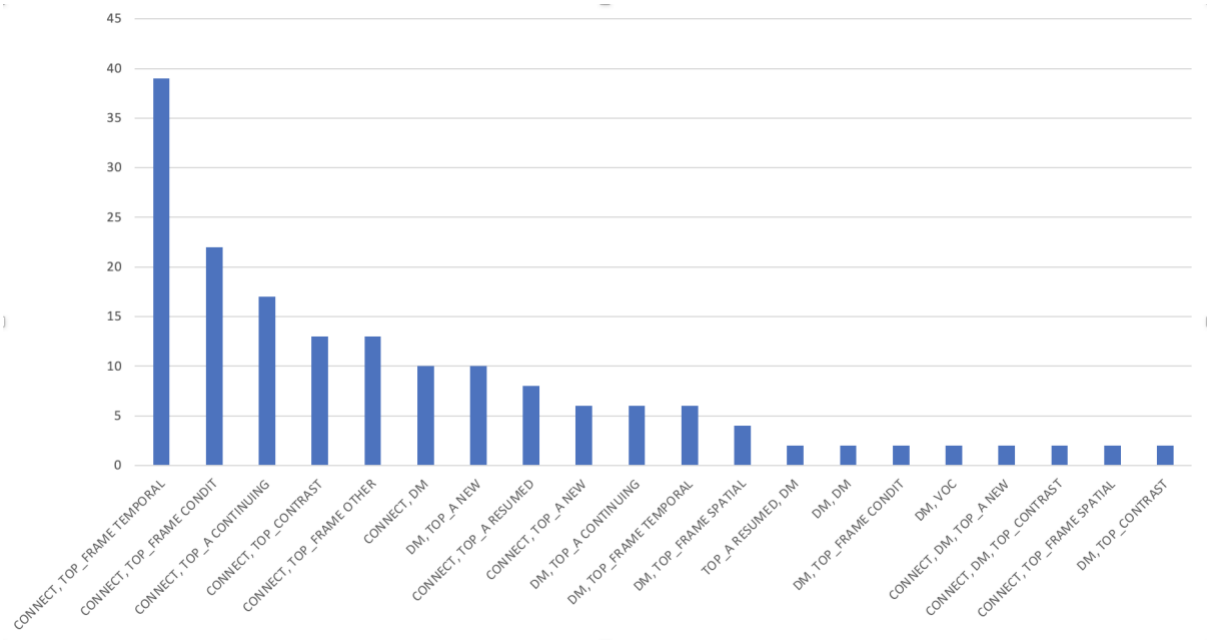
Figure 31 - Classement des étiquettes pragmatiques des prénoyaux simples



Dans les prénoyaux simples, les marqueurs pragmatiques les plus fréquents sont de loin *TOP_FRAME TEMPORAL* (148) et *TOP_FRAME CONDIT* (100), suivis de *TOP_A NEW* (51) et *TOP_A CONTINUING* (42). Les moins fréquents sont *VOC* (14), *CONNECT* (9), et *TOP_A SUM* (9).

5.2.2 Contenu pragmatique des prénoyaux complexes

Figure 32 - Classement des étiquettes pragmatiques des prénoyaux complexes

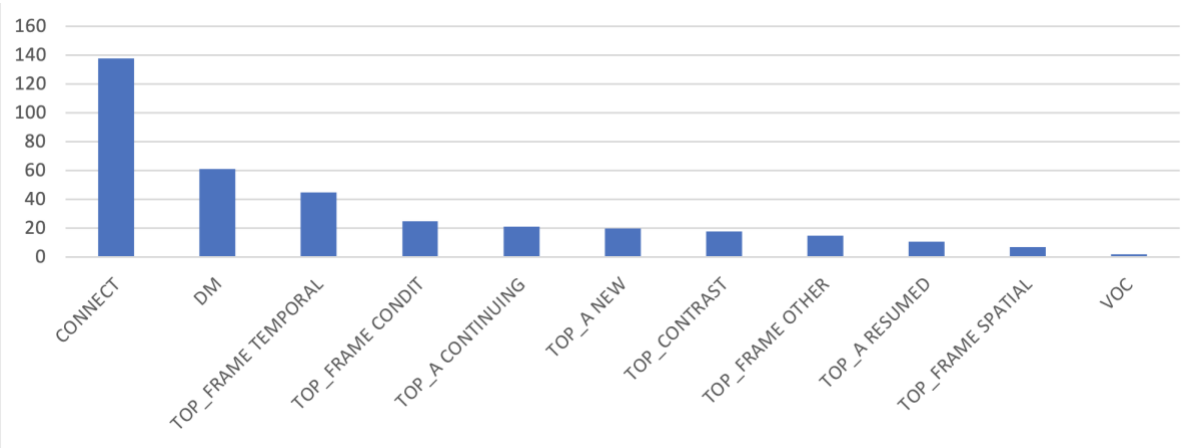


Dans les prénoyaux complexes, les configurations pragmatiques les plus courantes consistent en un simple *CONNECT* suivi d'un *TOP_FRAME TEMPORAL* (39), d'un *TOP_FRAME CONDIT* (22), d'un *TOP_A CONTINUING* (17), ou d'un *TOP_CONTRAST* (13). La majorité des configurations présentes dans le corpus n'apparaissent qu'une ou deux fois. Cela inclut

des configurations comme un *DM* suivi d'un *DM*, ou d'un *DM* suivi d'un *TOP_CONTRAST*. Notez que pour des raisons de lisibilité, les configurations n'apparaissant qu'une seule fois ont été exclues de la figure 32.

5.2.3 Contenu pragmatique des composants des prénoyaux complexes

Figure 33 - Classement des étiquettes pragmatiques des composants des prénoyaux complexes



Dans les composants pragmatiques des prénoyaux complexes, les fonctions les plus courantes sont *CONNECT* (138), *DM* (161) et *TOP_FRAME TEMPORAL* (45). Les moins fréquentes sont *TOP_A RESUMED* (11), *TOP_FRAME SPATIAL* (7) et *VOC* (2).

5.2.4 Analyse de la fréquence des catégories pragmatiques

En comparant les fréquences des différents types pragmatiques présents dans ces trois niveaux d'analyse, on peut noter plusieurs tendances intéressantes. Dans les prénoyaux simples, les deux fonctions pragmatiques les plus courantes sont *TOP_FRAME TEMPORAL* et *TOP_FRAME CONDIT*. De même, dans les prénoyaux complexes, les deux fonctions pragmatiques les plus courantes sont un *CONNECT* suivi d'un *TOP_FRAME TEMPORAL* ou d'un *TOP_FRAME CONDIT*.

Cependant, on peut également constater des divergences. La troisième fonction pragmatique la plus courante dans les prénoyaux simples est *TOP_A NEW*. Alors que l'on pourrait s'attendre à ce que la troisième fonction la plus courante des prénoyaux complexes soit un *CONNECT* suivi d'un *TOP_A NEW*, ce dernier n'apparaît que dans la neuvième fonction pragmatique la plus courante des prénoyaux complexes. Un regard sur les composants internes des prénoyaux complexes montre que *TOP_A NEW* est la sixième étiquette la plus courante, ou la quatrième si l'on exclut les *CONNECT* et *DM*. En revanche, *TOP_CONTRAST* et *TOP_FRAME OTHER* semblent occuper une place beaucoup plus importante dans les prénoyaux complexes que dans les prénoyaux simples. Nous pouvons également observer

que les *VOC*, très rares dans l'ensemble de notre corpus, sont concentrés dans les prénoyaux simples. En d'autres termes, les vocatifs apparaissent souvent seuls, sans être accompagnés d'un connecteur ou marqueur de discours.

Cette analyse nous permet de faire quelques observations. La première est qu'il semble exister des différences réelles et concrètes entre les fonctions pragmatiques qui sont typiques des prénoyaux complexes, et celles qui sont typiques des prénoyaux simples. Il serait donc incorrect de conceptualiser le prénoyau complexe prototypique comme étant un prénoyau simple qui contient un connecteur ou un marqueur de discours supplémentaire. En même temps, il serait peu judicieux de considérer les prénoyaux complexes et simples comme des catégories mentales distinctes auxquelles les locuteurs attribuent des fonctions différentes. Une interprétation beaucoup plus simple serait de dire que les différents types pragmatiques ont des probabilités différentes d'être accompagnés d'un type pragmatique supplémentaire, tel qu'un connecteur ou un marqueur de discours. Les vocatifs et les nouveaux topics ont une forte probabilité d'apparaître seuls, alors qu'il n'est pas rare d'observer des topics contrastifs ou des cadres de discours après un connecteur.

Cette observation soulève plusieurs questions qui dépassent largement le cadre de cette étude. Les connecteurs apparaissant relativement fréquemment à côté des topics contrastifs ou des cadres de discours, il est raisonnable de se demander si les connecteurs servent également à marquer ces fonctions pragmatiques d'une certaine manière. Selon cette interprétation, un connecteur pourrait servir non seulement à relier un énoncé à un énoncé précédent, mais aussi à souligner indirectement le caractère contrastif d'un topic, par exemple. Une deuxième interprétation de ces données est qu'il n'y a pas de lien linguistique direct entre les connecteurs et les topics qu'ils précèdent, mais que certaines fonctions pragmatiques sont simplement plus ou moins susceptibles d'apparaître dans des contextes où il y a également un connecteur.

Une observation intéressante à l'appui de cette interprétation est que la grande majorité des nouveaux topics apparaissent sans connecteur. Pendant la phase d'annotation pragmatique, nous avons remarqué que les nouveaux topics étaient relativement communs dans les premiers énoncés de chaque monologue, c'est-à-dire les contextes dans lesquels les connecteurs ne sont pas typiquement utilisés. Une analyse statistique plus détaillée de ces données serait probablement nécessaire pour tirer des conclusions plus significatives. Une étude future sur la structure pragmatique des prénoyaux du naija pourrait envisager de quantifier les corrélations entre les types pragmatiques et les contextes dans lesquels ils apparaissent.

On peut enfin noter que les cadres de discours sont globalement plus représentés que les topics dans tous les niveaux d'analyse. Ceci suggère que la position initiale n'est pas nécessairement la position préférée pour désigner les topics en naija. Il convient également de noter que les cadres temporels sont beaucoup plus courants que les cadres spatiaux. Nous pensons que cela est en partie dû aux types d'enregistrements utilisés dans cette étude, qui sont en grande partie composés de genres narratifs avec des locuteurs qui racontent leur vie. Puisque, dans ce genre de discours, les locuteurs passent la plupart de leur temps à décrire des événements de leur passé, il est logique que les cadres temporels soient fréquemment utilisés.

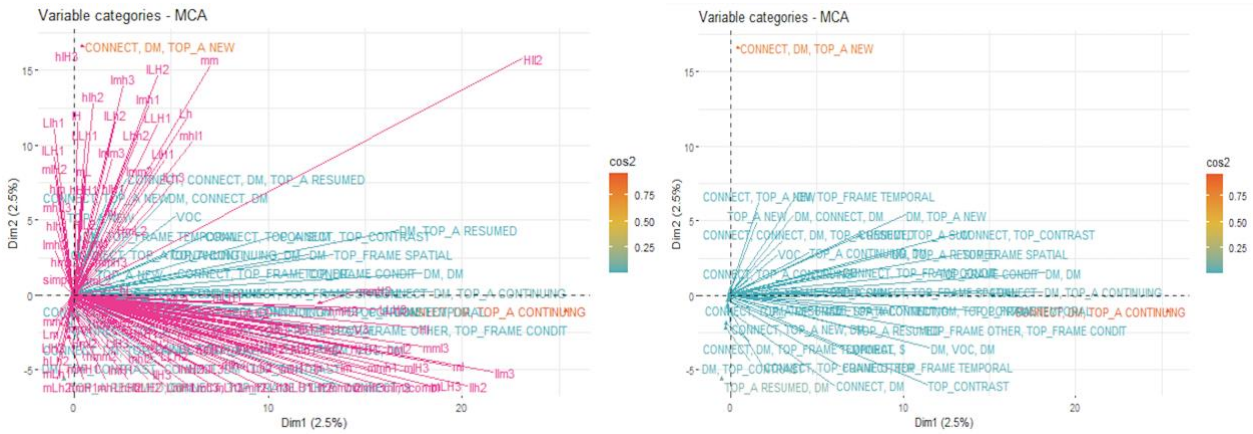
5.3 Analyse des graphiques ACM

Dans cette section, nous irons au-delà d'une simple analyse quantitative du contenu pragmatique des prénoyaux. Ainsi, nous examinerons les graphiques ACM produits dans le cadre de cette étude dans le but de valider nos trois hypothèses. Cette section commence par un aperçu des défis liés à l'interprétation de nos résultats, suivi de trois sous-sections supplémentaires commentant chaque hypothèse.

5.3.1 Difficultés liées à l'interprétation des graphiques ACM

Un total de 80 graphiques ACM a été produit dans le cadre de ce projet en faisant varier différents paramètres décrits dans la section méthodologique. Dans la majorité des cas, les graphiques ACM générés contenaient trop de points de données pour être interprétables.

Figure 34 - Graphiques ACM CSptloc et CSpt_generedepuisloc



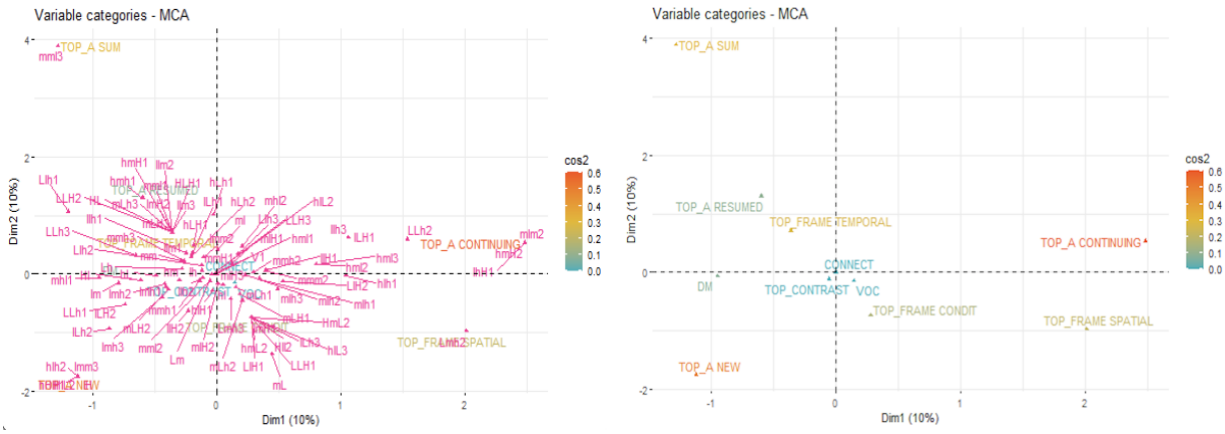
La paire d'images dans la figure 34 est représentative de la plupart des graphiques ACM générés pendant notre étude. À gauche, nous voyons une projection de tous les types pragmatiques présents dans les prénoyaux complexes, avec leurs contours prosodiques

locaux affichés comme variable secondaire. L'image de droite représente la même distribution, avec les contours prosodiques cachés. Avec ou sans les contours prosodiques cachés, ces graphiques sont en l'état ininterprétables. Presque tous les types pragmatiques et les contours prosodiques semblent se regrouper autour du même point sur le graphique, ce qui suggère qu'il n'y a pas de relation significative entre les contours et les types pragmatiques. En outre, les deux dimensions ne représentent que 2,5 % de la variance des données, ce qui est négligeable.

Seuls deux labels pragmatiques semblent se distinguer des autres : *CONNECT, DM, TOP_A NEW* et *CONNECT, DM, TOP_A CONTINUING*. Leurs positions s'alignent étroitement avec respectivement les contours hlH2 et mmH2.

Un examen de notre ensemble de données révèle que le type pragmatique *CONNECT, DM, TOP_A NEW* n'apparaît que deux fois, tandis que *CONNECT, DM, TOP_A CONTINUING* apparaît exactement une fois. Étant donné leur faible nombre d'occurrences, il est naturel qu'ils soient étroitement associés à certains contours prosodiques. Le seul prénoyau complexe portant l'étiquette pragmatique *CONNECT, DM, TOP_A CONTINUING* porte également le contour prosodique local mmH2, tandis que l'une des deux CONNECT, DM, TOP_A NEW porte le contour hlH2. Ces deux contours apparaissent moins de cinq fois dans les prénoyaux complexes. La corrélation est donc forte car ces deux étiquettes pragmatiques et prosodiques sont peu représentées.

Figure 35 - Graphiques ACM SIMPLESptloc et SIMPLESpt_generedepuisloc

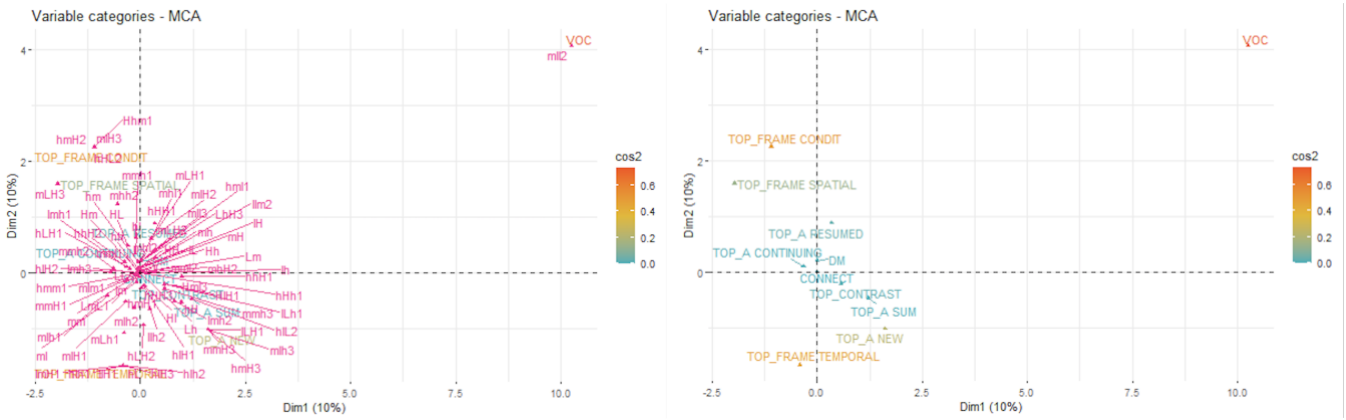


Si nous examinons les graphiques ACM générés au niveau des prénoyaux simples en utilisant les mêmes paramètres, les résultats deviennent plus lisibles. Ceux-ci peuvent être vus dans la figure 35 ci-dessus. Quand les contours prosodiques sont rendus visibles (figure de gauche), le graphique est dense mais il est toujours possible de visualiser certains groupes de contours prosodiques qui sont associés à certains types pragmatiques. On peut également noter que chaque dimension compte pour 10% de la variation représentée dans nos données.

Bien que ce chiffre soit encore relativement faible, il convient de noter qu'il est nettement supérieur à celui de l'ACM générée pour les contours complexes. Nous pouvons également constater une répartition claire entre les différentes étiquettes pragmatiques, ce qui donne un graphique globalement plus interprétable. Les améliorations constatées dans ces graphiques ACM sont probablement liées au fait qu'il y a moins d'étiquettes pragmatiques possibles dans les prénoyaux simples, ce qui donne un ensemble de données plus simple et plus équilibré. En effet, tous les types pragmatiques des prénoyaux simples apparaissent au moins neuf fois. Dans les prénoyaux complexes, une partie importante des types pragmatiques n'apparaît qu'une ou deux fois, ce qui donne un ensemble de données plus complexe et moins interprétable.

Si l'on examine les composants internes des prénoyaux complexes, on obtient des graphiques ACM similaires à ceux des prénoyaux simples. Les exemples dans la figure 36 sont également basés sur les contours locaux, bien que nous ayons également remarqué que l'utilisation des contours globaux n'avait pas d'effet significatif sur la qualité ou l'interprétabilité des données.

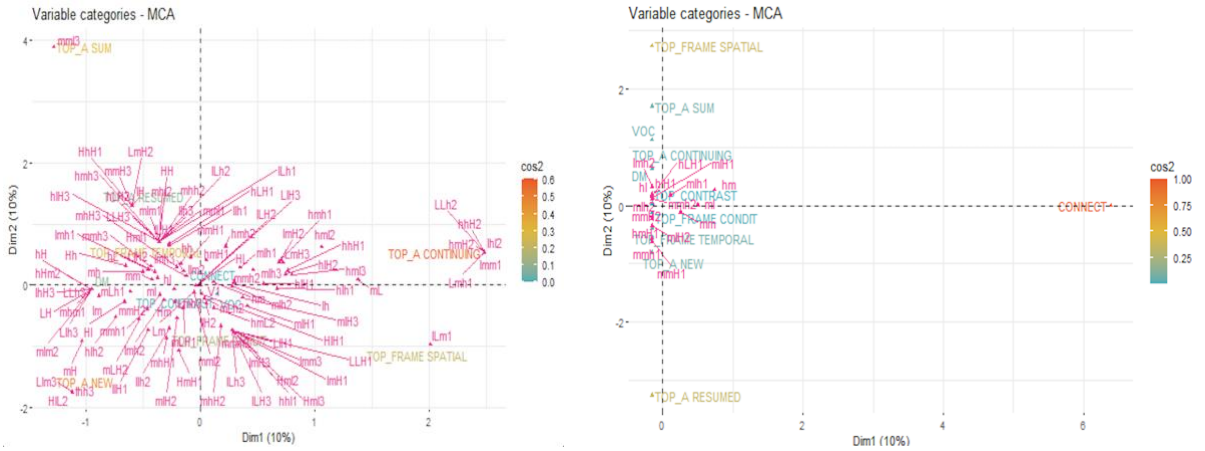
Figure 36 - Graphiques ACM internesptloc et internespt_generedepuisloc



Comme dans le précédent exemple, nous voyons des dimensions qui représentent 10% de la variation totale, et une dispersion relativement claire des variables. Cependant, nous constatons également que les vocatifs sont extrêmement éloignés du reste des données. Rappelons que les vocatifs sont extrêmement rares dans les prénoyaux complexes, et par extension dans leurs composants. En fait, seuls deux vocatifs sont présents à ce niveau d'analyse. Une fois encore, l'inclusion d'une étiquette pragmatique fortement sous-représentée semble fausser les résultats. Les vocatifs sont considérés comme très distincts des autres simplement parce que l'un des deux vocatifs a un contour prosodique de mll2, contour qui n'apparaît qu'une seule fois dans le corpus. Cela illustre surtout l'importance d'un ensemble de données relativement équilibré.

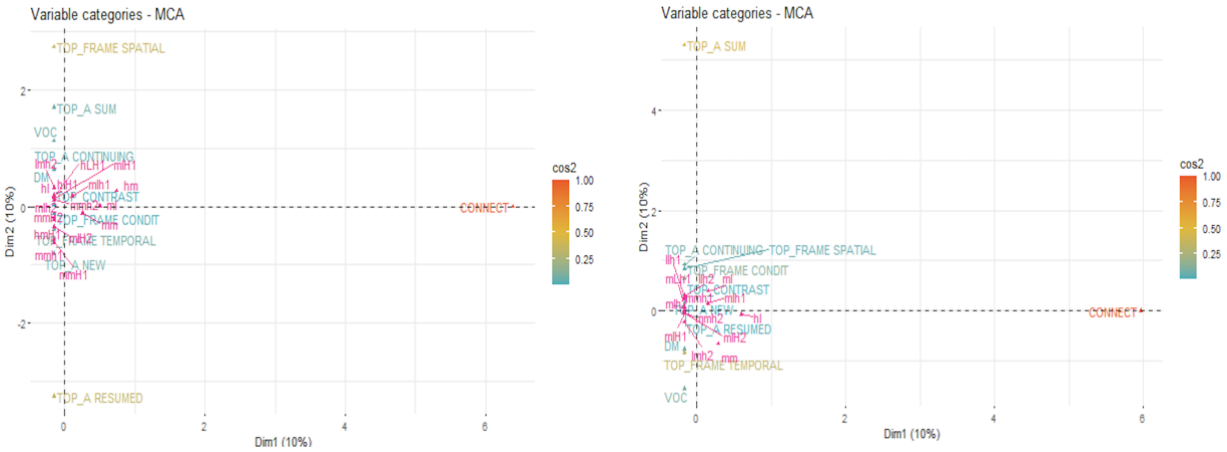
Lors de la conception du projet, nous avons anticipé ce problème. Pour cette raison, nous avons également généré une série de graphiques ACM qui ne tient compte que des contours les plus fréquents. Cependant, comme le montre la figure 37, cela ne donne pas nécessairement des graphiques ACM plus interprétables.

Figure 37 - Graphiques ACM SIMPLESptglo et SIMPLES50ptglo



L'image de gauche représente un graphique ACM généré à partir de prénoyaux simples en utilisant tous les contours globaux comme variable secondaire. L'image de droite représente les mêmes paramètres, mais avec seulement les contours les plus fréquents. Bien qu'il y ait moins d'étiquettes obscurcissant l'image, le graphique résultant est sans doute encore plus difficile à interpréter, puisque certaines étiquettes ne semblent pas être associées à des contours prosodiques. En effet, le fait que *TOP_FRAME SPATIAL*, *CONNECT* et *TOP_A RESUMED* soient tous situés aux extrémités des deux axes soulève des questions quant au type d'information que chaque axe représente. *TOP_A SUM* et *VOC* sont également très rares, mais sont situés près de l'origine. Étant donné la nature opaque des ACM, les raisons de ce phénomène ne sont pas entièrement claires. Il semble probable que lorsque nous avons exclu les contours les moins communs, la plupart ou la totalité des étiquettes prosodiques associées aux contours *TOP_FRAME SPATIAL*, *CONNECT* et *TOP_A RESUMED* ont été éliminées. S'il semble difficile d'expliquer ce résultat, on peut supposer qu'il reflète une erreur méthodologique importante : en excluant les contours les moins représentatifs du corpus (couverture inférieure à 50%), nous avons du coup exclu les contours qui représentaient ces étiquettes et le fait que ces étiquettes se retrouvent aux extrémités des axes n'est que la conséquence de cette erreur méthodologique. En utilisant la même logique, nous pourrions supposer que *TOP_A SUM* et *VOC* sont plus proches de l'origine parce que les contours prosodiques associés sont communs ailleurs. Un phénomène similaire apparaît si nous observons le graphique équivalent produit à partir des contours locaux les plus courants.

Figure 38 - Graphiques ACM SIMPLES50ptglo et SIMPLES50ptloc

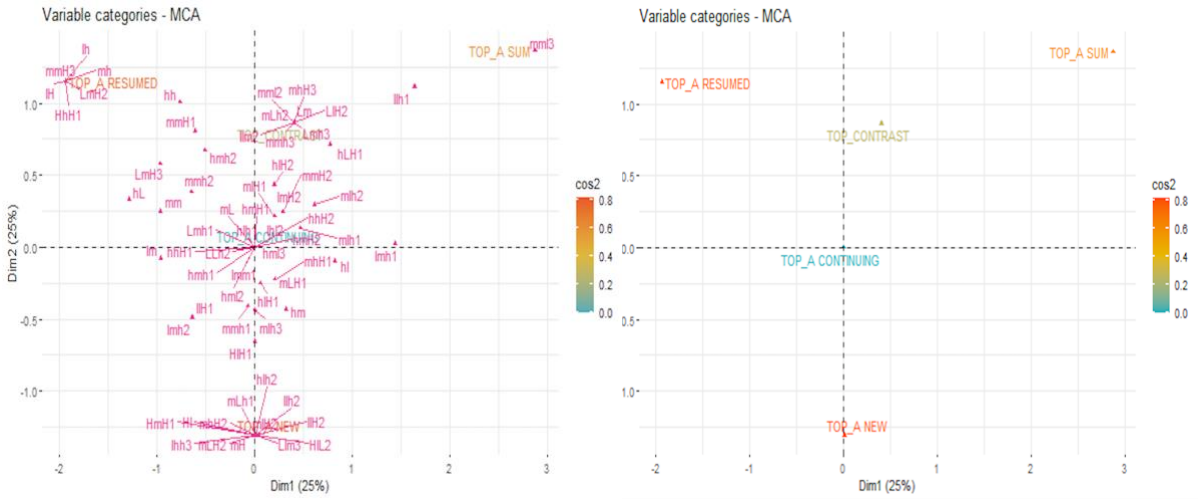


Que l'on utilise ou non les contours locaux ou globaux pour générer les graphiques, les étiquettes pragmatiques qui apparaissent aux extrémités sont toujours parmi les moins courantes. Cependant, la nature de ces rares étiquettes apparaissant aux extrémités des axes peut varier. Notez que dans les exemples dans la figure 38 ci-dessus, l'étiquette TOP_FRAME SPATIAL apparaît à l'extrémité supérieure de l'axe vertical dans le graphique généré à partir des contours globaux (gauche) mais est proche de l'origine dans le graphique généré à partir des contours locaux (droit). L'explication la plus raisonnable est que dans certains cas, une méthode de génération d'un contour (local vs global) assignera à un prénoyau donné un contour prosodique plus fréquent que l'autre méthode. Pour les types pragmatiques qui sont très rares, cela fait la différence entre avoir des contours prosodiques associés qui sont suffisamment rares pour être exclus et avoir des contours qui sont suffisamment courants pour être inclus.

Si ces explications sont justes, ce problème représente un défaut méthodologique qui devrait être corrigé dans les études futures. Cependant, un ensemble de données beaucoup plus important pourrait également permettre de rectifier ce problème. Ces exemples montrent qu'il y a au moins un inconvénient important à exclure les contours prosodiques les moins courants.

Les graphiques ACM les plus intéressants sont ceux qui ont été développés avec seulement un sous-ensemble des étiquettes pragmatiques, les topics, comme variable primaire. Cela a eu tendance à produire des graphiques avec une distribution relativement égale des étiquettes pragmatiques. Ces graphiques étaient également assez facilement interprétables, avec ou sans les contours prosodiques moins courants.

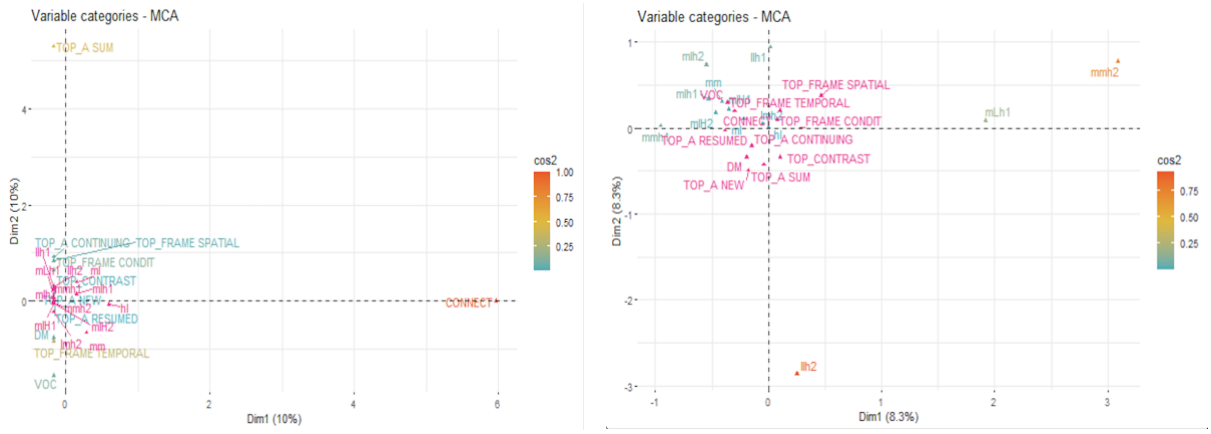
Figure 39 - Graphiques ACM SimpleListeptglo et SimpleListept_generedepuisglo



Les exemples dans la figure 39 montrent la distribution des différentes catégories de topics contenues dans les prénoyaux simples, avec la totalité des contours prosodiques globaux utilisés comme variable secondaire (cachés dans l'image de droite). Dans ces images, nous pouvons voir une distribution relativement égale des étiquettes pragmatiques, et toutes sont associées à un groupe de contours. Certains contours apparaissent entre deux étiquettes pragmatiques, ce qui suggère qu'ils sont corrélés de manière relativement égale avec les deux. Nous constatons également que les deux dimensions représentent un total de 50 % de la variance dans le corpus. Cela suggère que les corrélations observées sont particulièrement fortes.

De manière générale, le fait que les contours prosodiques ou les étiquettes pragmatiques soient utilisés comme variable primaire n'a pas eu d'effet significatif sur l'utilité des résultats. La figure 40 ci-dessous compare des graphiques produits avec les mêmes paramètres, à l'exception du choix de la variable primaire.

Figure 40 - Graphiques ACM SIMPLES50ptloc et SIMPLES50locpt



Ces exemples montrent des projections incluant les contours locaux les plus courants dans des prénoyaux simples. À gauche, les contours prosodiques ont été utilisés comme variable primaire. A droite, les étiquettes pragmatiques ont été utilisées comme variable primaire. Aucune des deux n'est nécessairement plus utile que l'autre, et toutes deux semblent avoir un problème commun : une majorité de points situés près de l'origine, et quelques points situés aux extrémités des axes.

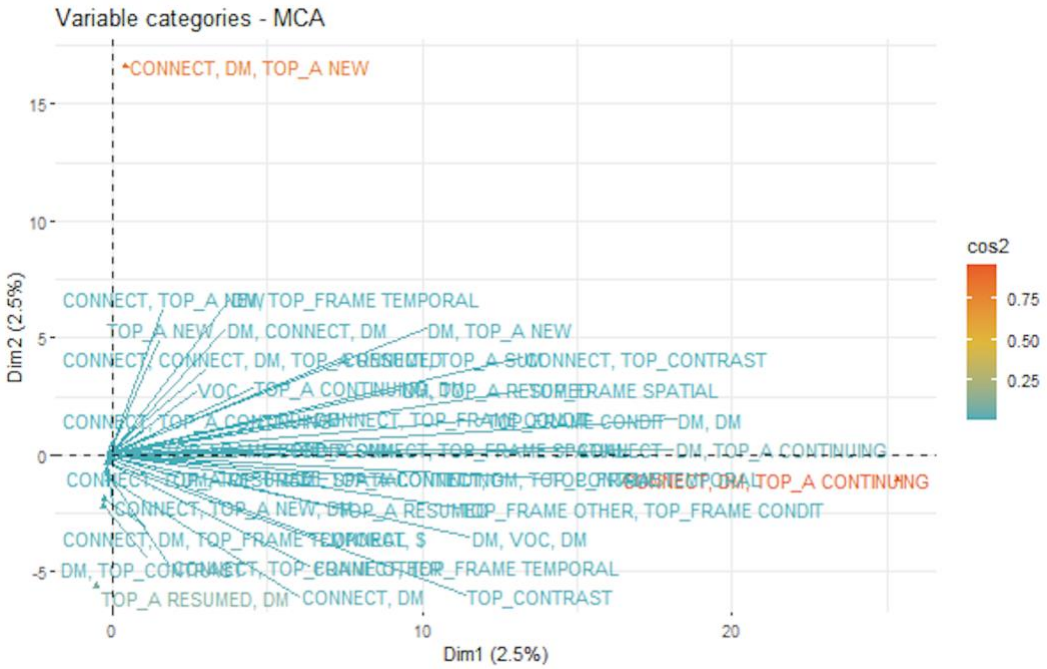
5.3.2 Validation des hypothèses

La section suivante est consacrée à l'interprétation des graphiques ACM générés dans cette étude dans le but de valider nos trois principales hypothèses.

5.3.2.1 Peut-on observer des différences prosodiques entre les différents types pragmatiques ?

Notre première hypothèse était que les différents types pragmatiques soient marqués par des différences prosodiques statistiquement observables. Dans un graphique ACM, une telle situation serait reflétée par une distribution relativement égale des types pragmatiques, qui se regroupent autour de différents types de contours prosodiques. Bien entendu, les graphiques ACM dont les dimensions représentent un pourcentage important de la variabilité des données doivent être privilégiés. A en juger par ces critères, les graphiques ACM générés dans le cadre de cette étude sont de qualité variable. Comme nous l'avons vu dans la section précédente, les graphiques ACM générés pour des prénoyaux complexes sont généralement les moins utiles.

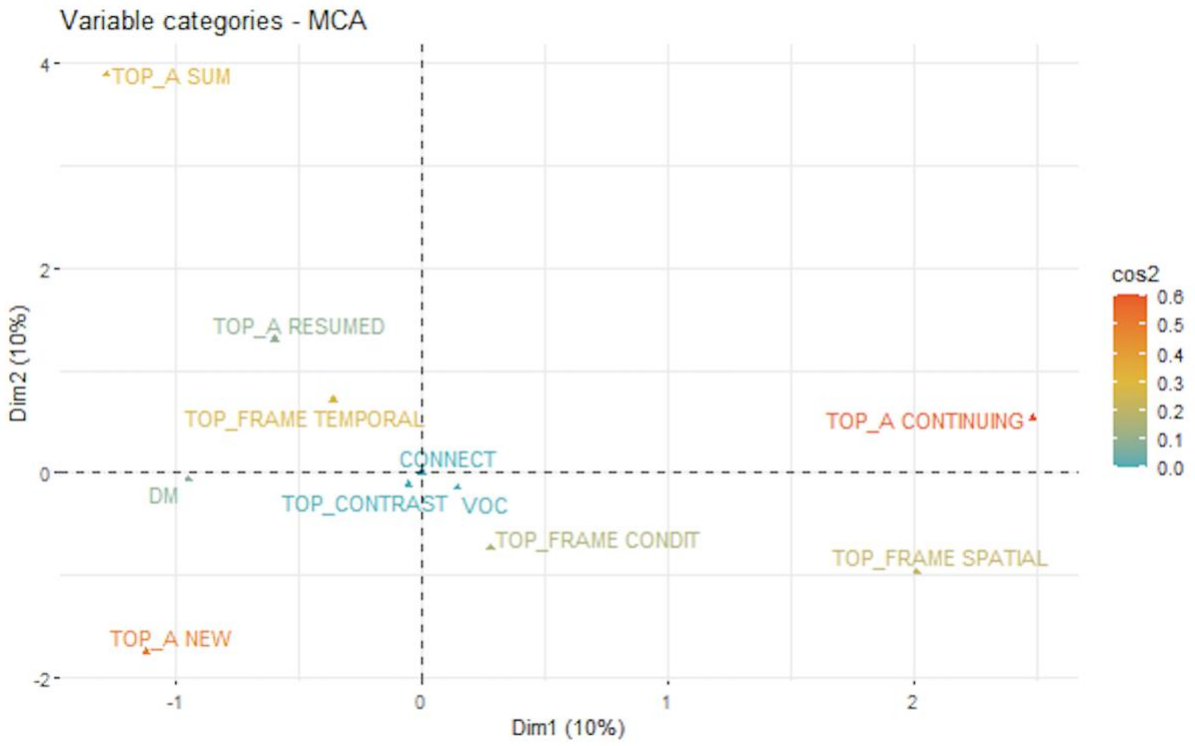
Figure 41 - Graphique ACM CSpt_generedepuisglo



Le graphique représenté dans la figure 41, généré en utilisant les étiquettes pragmatiques des prénoyaux complexes comme variable primaire et leurs contours globaux comme variable secondaire (cachée dans cette image), montre que presque toutes les étiquettes pragmatiques sont placées à l'origine. Seules deux étiquettes sont présentées comme étant prosodiquement distinctes, et c'est uniquement parce qu'elles n'apparaissent qu'une seule fois dans le corpus avec des contours prosodiques qui n'apparaissent pas ailleurs. Les prénoyaux complexes contiennent donc trop d'étiquettes pragmatiques possibles réparties sur un ensemble de données trop petit pour fournir des statistiques significatives. Les deux axes ne représentent également que 2,5% de la variation dans le corpus, ce qui rend ce graphique particulièrement inexploitable pour observer des relations significatives.

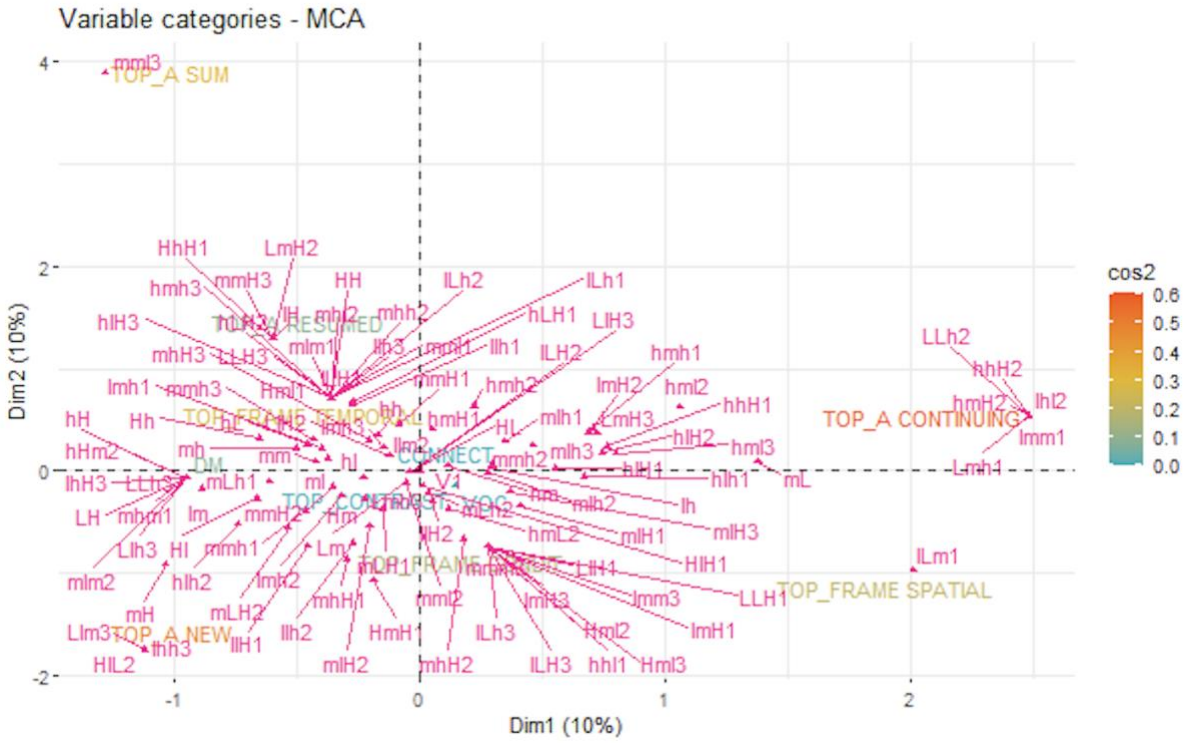
Un regard sur les graphiques ACM générés pour des prénoyaux simples est beaucoup plus utile. La plupart des étiquettes pragmatiques sont réparties de manière beaucoup plus homogène à partir de l'origine, et les étiquettes situées près des bords du graphique ne correspondent pas aux étiquettes les plus rares. Dans l'exemple représenté dans la figure 42, chaque axe représente 10% de la variation totale des données, une proportion modeste mais néanmoins non négligeable.

Figure 42 - Graphique ACM SIMPLESpt_generedepuisglo



Tout ceci suggère que les fonctions pragmatiques affichées sont en effet prosodiquement distinctes les unes des autres, *TOP_A CONTINUING*, *TOP_A NEW*, *TOP_FRAME SPATIAL* et *TOP_A SUM* étant les plus distinctes dans l'ensemble. Les étiquettes *CONNECT*, *VOC*, et *TOP_CONTRAST* semblent être plus similaires les unes aux autres et moins distinctes globalement.

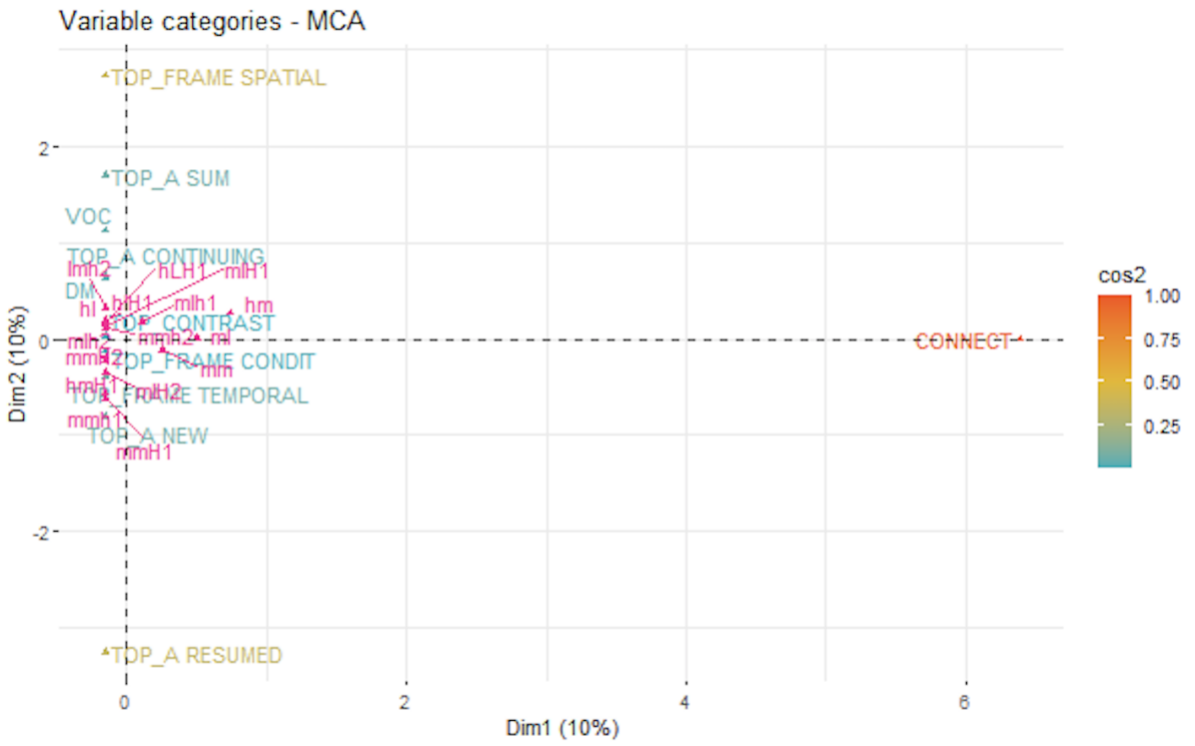
Figure 43 - Graphique ACM SIMPLESptglo



La visualisation de cette distribution avec les contours prosodiques visibles dans la figure 43 ci-dessus le confirme largement. Les contours situés près de l'origine sont associés à un éventail de contours beaucoup plus large, tandis que ceux situés près des extrêmes sont associés à un sous-ensemble plus restreint de contours. On peut interpréter cela comme signifiant qu'au moins certains types pragmatiques sont effectivement distincts les uns des autres. Cela est particulièrement vrai pour *TOP_A CONTINUING* qui est directement associé à un group distinct de contours. Il convient de noter que *TOP_A SUM* apparaît moins de 10 fois dans les prénoyaux simples, ce qui signifie que son positionnement doit être pris avec un certain degré de précaution.

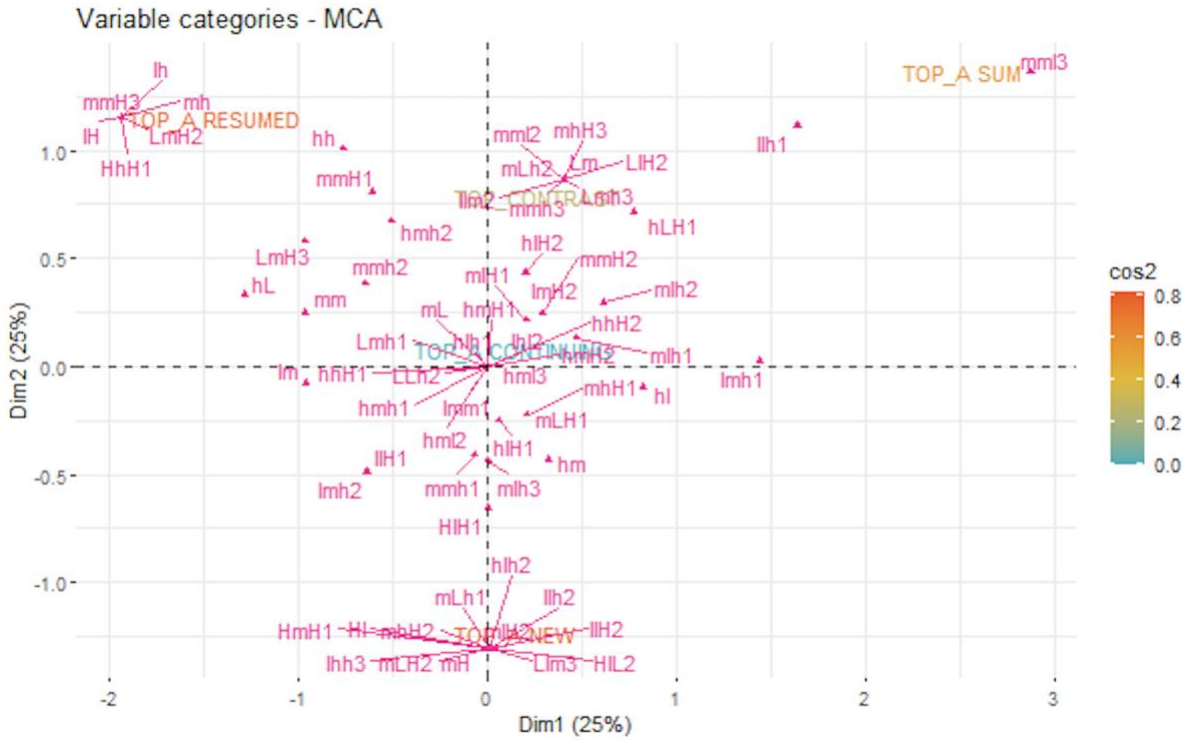
Les graphiques ACM qui incluent seulement les contours les plus courants semblent être moins utiles pour ce niveau d'analyse, comme le montre l'exemple dans la figure 44. Les étiquettes pragmatiques situées sur les bords du graphique ne sont associées à aucun contour, ce qui suggère que les contours associés n'étaient pas suffisamment communs pour être inclus.

Figure 44 - Graphique ACM SIMPLES50ptglo



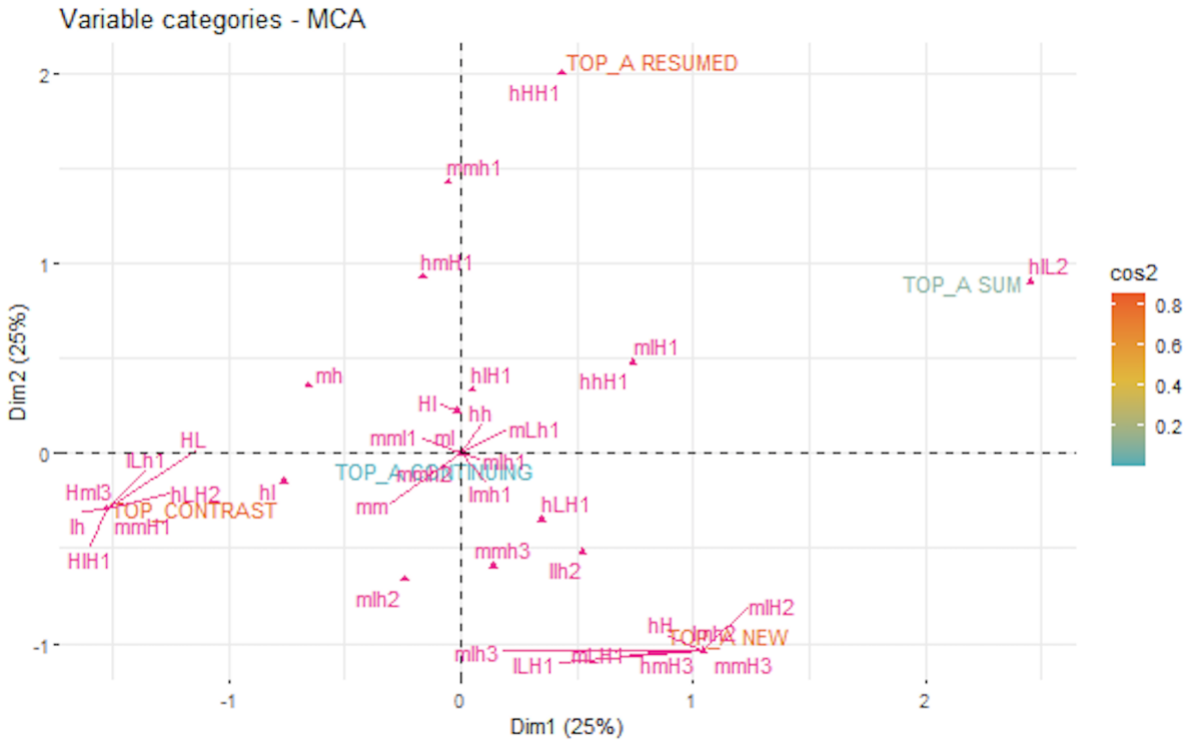
Les graphiques les plus utiles dans l'ensemble pour la validation de cette hypothèse étaient ceux produits en utilisant tous les contours prosodiques comme variable secondaire, et uniquement les topics comme variable primaire. Le graphique représenté dans la figure 45 représente la moitié de la variation totale du corpus, et montre un haut degré de distinction prosodique entre les différents topics. Chaque topic, à l'exception du relativement rare *TOP_A SUM*, est associé à un large éventail de contours prosodiques. Certains contours semblent être associés de manière plus ou moins égale à différents types pragmatiques. HlH1, par exemple, se situe entre *TOP_A NEW* et *TOP_A CONTINUING*, mais est très éloigné des autres étiquettes.

Figure 45 - Graphique ACM SimpleListeptglo



Une distribution similaire peut être observée en regardant le graphique équivalent produit pour les composants des prénoyaux complexes (figure 45), bien que l'on puisse noter que la variété des contours est moins grande dans l'ensemble. Ceci est dû au fait que les types pragmatiques représentés dans le graphique sont simplement moins fréquents dans les prénoyaux complexes.

Figure 46 - Graphique ACM InternesListeptglo

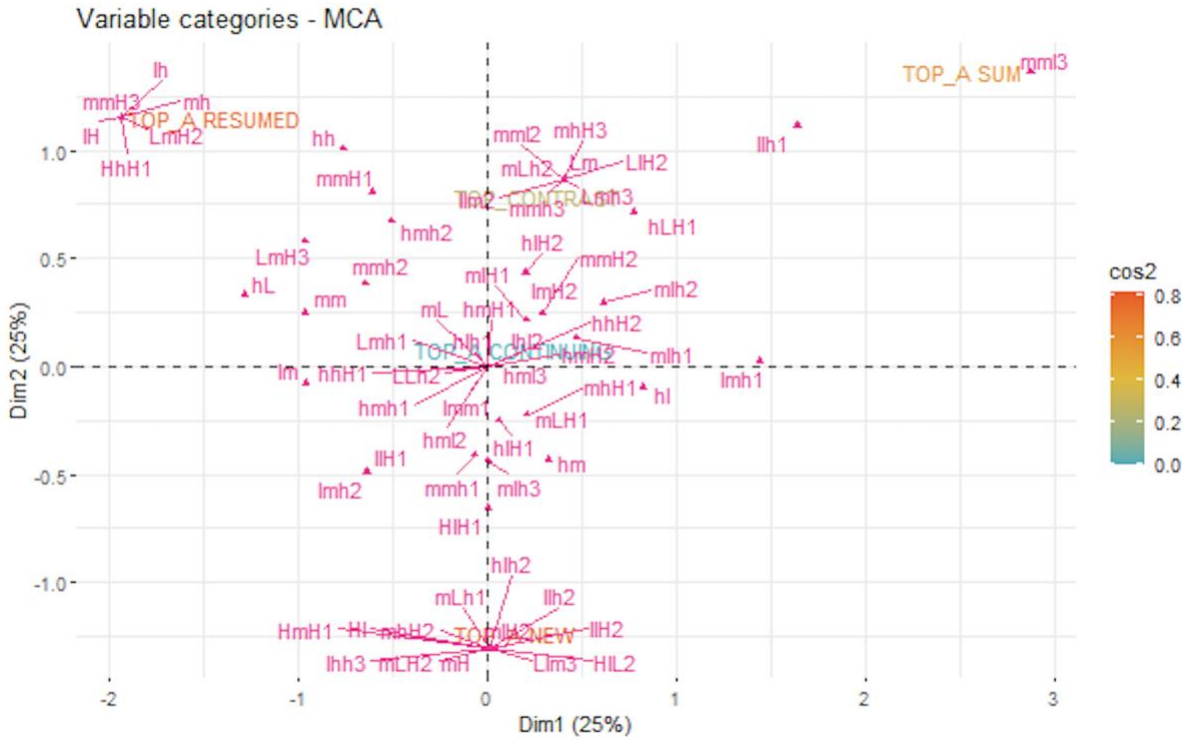


Dans l'ensemble, nous pouvons dire avec un degré de certitude relatif qu'au moins certains types pragmatiques sont prosodiquement distincts les uns des autres. Ceci est particulièrement apparent lorsque nous analysons des graphiques ACM ne contenant que des topics, qui sont tous associés à un ensemble distinct de contours prosodiques. Comme nous l'avons vu, cette hypothèse est difficile à valider au niveau des prénoyaux complexes, qui sont étiquetés avec un large éventail de types pragmatiques composites possibles, qui ont souvent un faible nombre d'occurrences. Pour améliorer la fiabilité de ces résultats, les chercheurs futurs devront utiliser un plus grand corpus de données, ce qui devrait diminuer le nombre de types pragmatiques qui n'apparaissent que plusieurs fois et avec un nombre très limité de contours prosodiques.

5.3.2.2 Peut-on identifier les contours prosodiques qui sont le plus fortement associés à certains types pragmatiques ?

Cette deuxième hypothèse est étroitement liée à la première, et a déjà été validée en partie dans la section précédente. Le graphique ACM dans la figure 47, déjà commenté dans la section précédente, montre clairement que chaque type de topic est associé à un ensemble différent de contours prosodiques. Pour rappel, ce graphique est généré en utilisant tous les contours prosodiques comme variable secondaire, et uniquement les topics comme variable primaire.

Figure 47 - Graphique ACM SimpleListeptglo



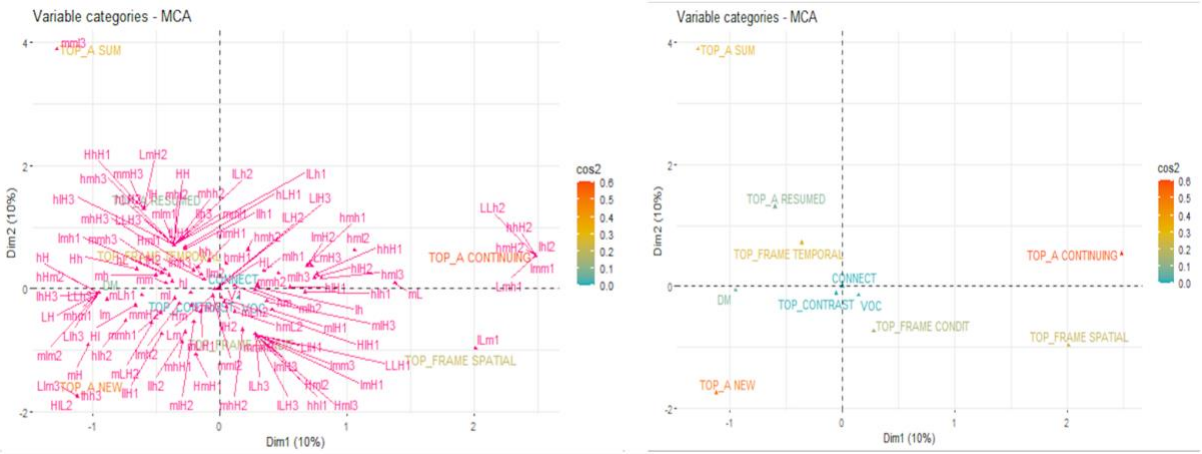
TOP_A CONTINUING est vaguement associé à un très large éventail de contours possibles, tandis que *TOP_A NEW*, *TOP_A RESUMED* et *TOP_CONTRAST* sont fortement associés à leurs propres groupes de contours prosodiques. *TOP_A SUM*, comme nous l'avons vu précédemment, est sous-représenté dans notre ensemble de données. Par conséquent, nous ne pouvons pas dire de manière définitive si ce type pragmatique se distinguerait des autres dans un ensemble de données plus équilibré. Parmi les cinq types pragmatiques représentés ici, *TOP_CONTRAST* et *TOP_A CONTINUING* semblent avoir un certain degré de similarité prosodique, se situant relativement près l'un de l'autre sur le graphique et ayant un grand nombre de contours prosodiques situés directement entre les deux.

Dans l'ensemble, ce graphique confirme largement notre deuxième hypothèse. Une question qui découle naturellement de l'inspection de ces données est de savoir si les contours prosodiques regroupés autour de chaque type pragmatique partagent ou non des caractéristiques communes. En d'autres termes, ces contours prosodiques appartiennent-ils à une classe de contours plus large ?

Il est difficile de fournir une réponse définitive à cette question. Toutefois, elle mérite certainement d'être approfondie. À première vue, il semblerait que la plupart des contours associés à TOP_A NEW se terminent par un ton bas ou très bas, alors que les contours associés avec TOP_A RESUMED terminent par un ton haut où très haut. Il semblerait

également que la plupart des contours associés à TOP_A CONTINUING se terminent par un ton moyen.

Figure 48 - Graphiques ACM SIMPLESptloc et SIMPLESpt_generedepuisloc



Si nous étendons notre analyse à tous les types pragmatiques présents dans les prénoyaux simples, nous constatons également que la plupart des types pragmatiques semblent être relativement distincts les uns des autres. Dans la figure 48 ci-dessus les topics sont situés assez loin les uns des autres, ce qui confirme notre observation précédente. Tous les cadres de discours sont également situés loin les uns des autres. Il semblerait également que les connecteurs, les vocatifs et les topics contrastifs soient tous prosodiquement assez similaires. Malheureusement, comme nous l'avons noté dans les sections précédentes, la fiabilité de ce graphique est minée par la fréquence relativement faible de certains types pragmatiques dans notre ensemble de données. Il convient également de rappeler que ce graphique ne représente que 20% de la variation totale de l'ensemble de données. Pour ces raisons, nous pouvons dire que ce graphique soutient principalement notre hypothèse, mais qu'il doit être considéré avec un certain recul.

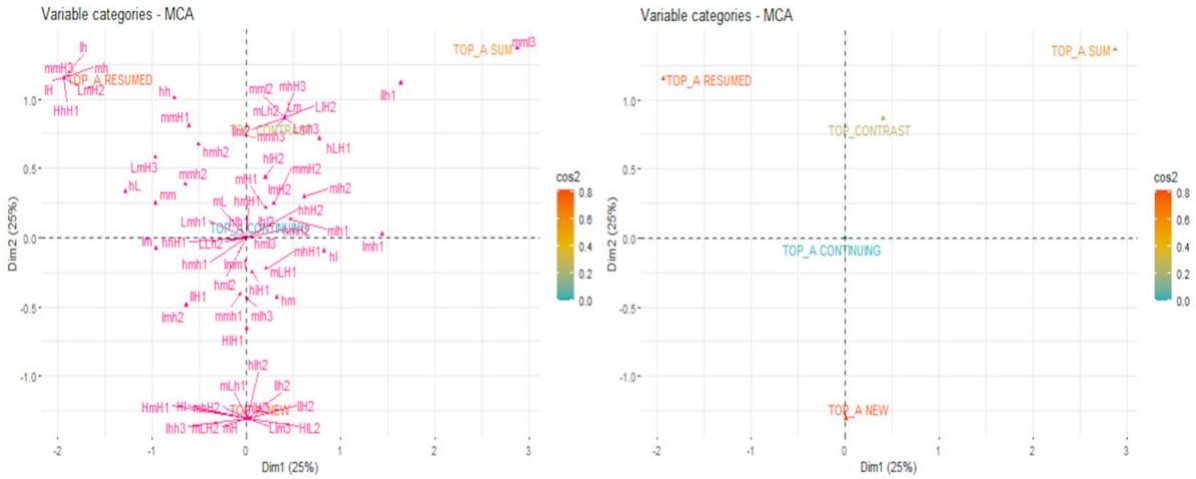
Dans l'ensemble, nous pouvons dire que notre deuxième hypothèse est largement validée. Ceci est particulièrement vrai pour les différents types de topics, qui sont tous très distincts les uns des autres. De futures études utilisant davantage de données permettront de confirmer ces résultats.

5.3.2.3 Les topics les plus accessibles sont-ils moins marqués prosodiquement ?

De toutes les hypothèses avancées, cette hypothèse est de loin la mieux soutenue par les données. Pour rappel, cette hypothèse repose sur l'idée que les topics qui ne sont pas immédiatement présents à l'esprit de l'auditeur sont plus susceptibles d'être marqués prosodiquement lorsqu'ils sont introduits par le locuteur. En d'autres termes, les topics

continus qui sont déjà au centre du discours seront moins marqués prosodiquement que les autres types de topics. Les résultats de nos traitements statistiques sont parfaitement conformes à ce profil, comme le montrent les graphiques la figure 49 ci-dessous. Pour rappel, ces graphiques ont été produits en utilisant les différents topics comme variable primaire, et les contours globaux comme variable secondaire.

Figure 49 - Graphiques ACM SimpleListeptglo et SimpleListept_generedepuisglo



Dans ces graphiques, nous voyons que l'étiquette *TOP_A CONTINUING* est située au milieu du graphique, et qu'elle est associée au plus grand nombre de contours prosodiques. Cela montre que *TOP_A CONTINUING* est le topic le moins marqué prosodiquement, car il coïncide avec une gamme extrêmement large de contours possibles. Et ce, malgré le fait que l'étiquette *TOP_A NEW* est en fait légèrement plus fréquente dans les prénoyaux simples. Si les contours prosodiques associés à chaque type pragmatique étaient le résultat du hasard, on s'attendrait à un plus grand nombre total de contours associés à l'étiquette *TOP_A NEW*. Les topics continus étant les plus accessibles dans l'esprit de l'auditeur, le fait que les topics continus soient les moins marqués par les contours type confirme parfaitement notre hypothèse finale.

5.4 Réexamen des données en utilisant des traits prosodiques

Cette section est consacrée à la présentation et à l'analyse des données générées après la soumission initiale de ce mémoire de recherche. La première sous-section est consacrée à la présentation et à l'analyse des graphiques ACM produits à l'aide des caractéristiques prosodiques extraites principalement des contours SLAM+. La deuxième sous-section est consacrée à l'exploration des p-values obtenues en appliquant le test exact de Fisher au même ensemble de données.

5.4.1 Graphiques ACM produits à partir des traits prosodiques

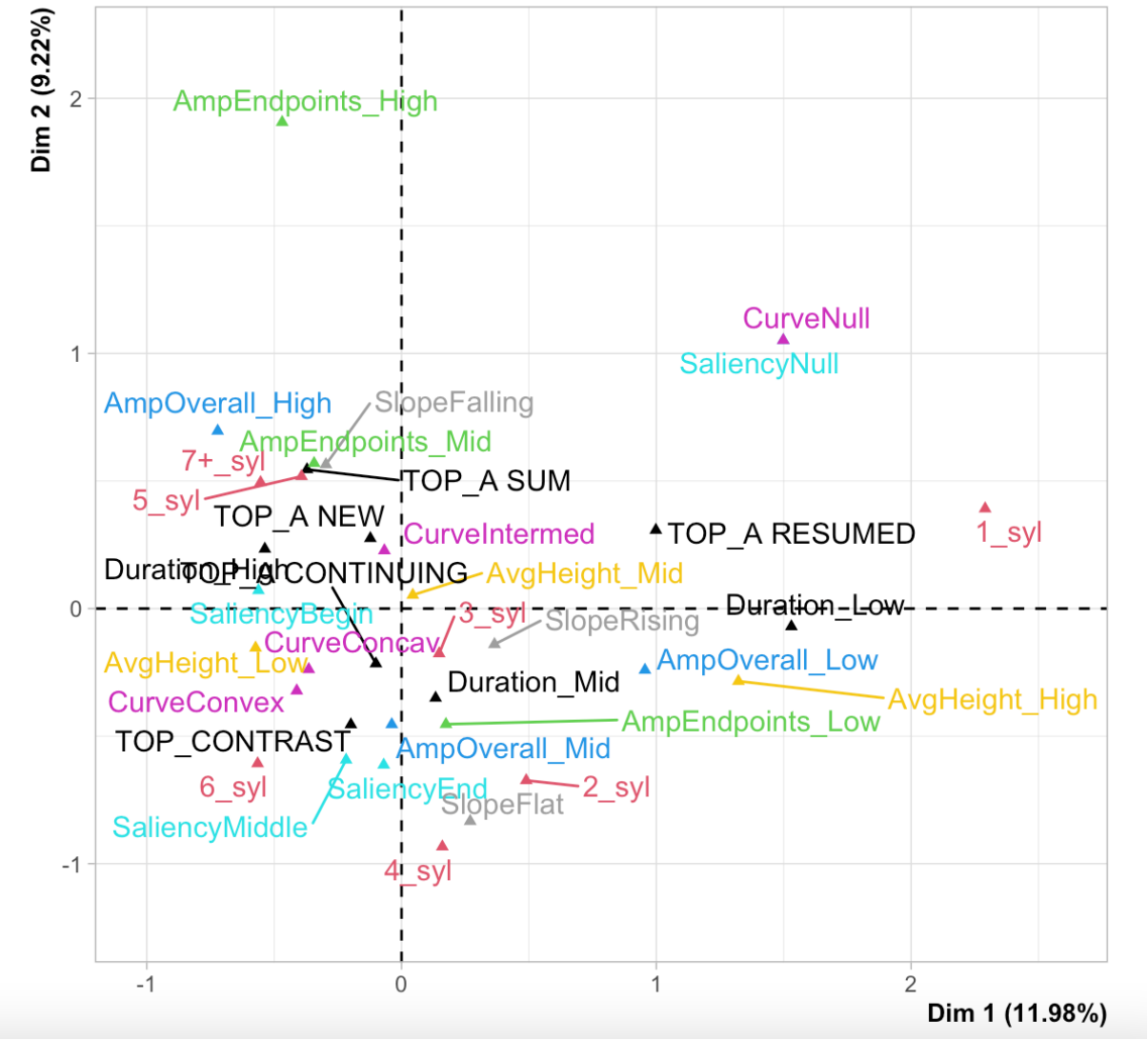
Deux graphiques ACM ont été produits au cours de cette phase de notre analyse : un pour les cinq catégories de topic situés dans les prénoyaux simples, et un pour ceux situés dans les prénoyaux complexes. Chaque graphique sera analysé séparément dans les sections 5.4.1.1 et 5.4.1.2.

En raison du nombre relativement important d'étiquettes impliquées, ces deux graphiques sont quelque peu encombrés. Ils sont néanmoins plus faciles à interpréter que la plupart des graphiques ACM présentés précédemment dans ce mémoire de recherche. Notez que chaque point dans ces deux graphiques correspond à chaque étiquette de notre ensemble de données, tandis que chaque couleur représente une catégorie de variables. Par exemple, les durées et les différents types pragmatiques sont représentés en noir, tandis que la hauteur moyenne est représentée en jaune.

Bien que le poids exact de chaque dimension varie entre les deux, chacun de ces graphiques représente environ un quart de la variation totale des jeux de données, un chiffre non négligeable. En raison du grand nombre de points de données, dont beaucoup se rapprochent de plusieurs types de topic, nous ne traiterons pas chaque variable en détail. Nous présenterons plutôt certaines observations qui nous ont semblé particulièrement pertinentes ou intéressantes pour ce mémoire de recherche.

5.4.1.1 Prénoyaux simples

Figure 50 - Graphique ACM produit à partir des prénoyaux simples



Dans le graphique ACM produit à partir des prénoyaux simples (Figure 50), les topics repris semblent être vaguement associés à une faible durée (Duration_Low), des segments monosyllabiques (1_syl) et une absence de saillance interne (CurveNull/SaliencyNull). Ils semblent également avoir une faible association avec des contours de faible amplitude, selon l'une de nos deux mesures (AmpOverall_Low). À en juger par ce graphique, les topics repris semblent être les plus distincts du point de vue prosodique, se trouvant loin des autres catégories de topic.

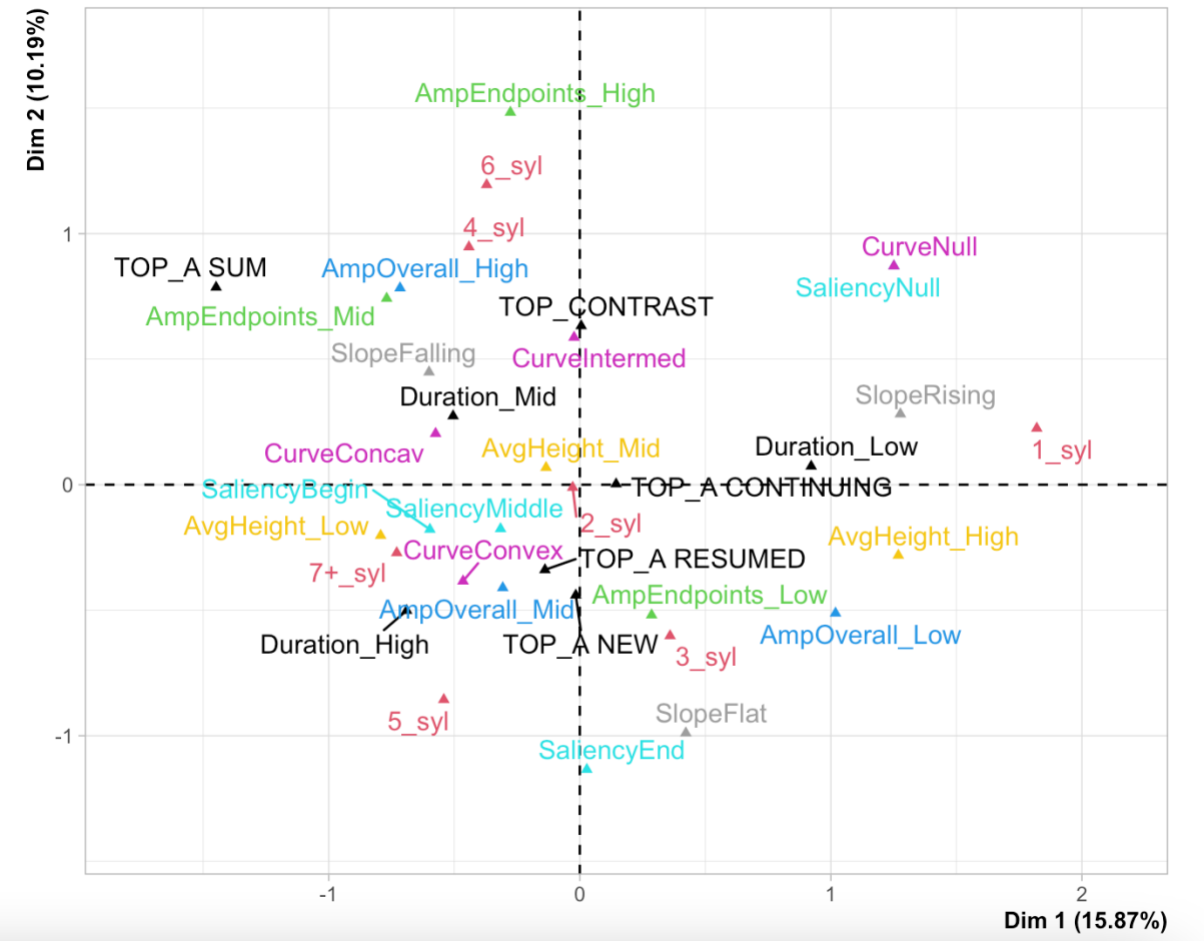
Les nouveaux topics et les topics récapitulatifs semblent être les plus étroitement liés à des contours descendants (SlopeFalling) et à des contours ayant une amplitude moyenne ou élevées selon la mesure utilisée (AmpOverall_High, EmpEndpoints_Mid). Ceux-ci sont associés à une durée élevée (Duration_High) et, logiquement, un relativement grand nombre

de syllabes (5_syl, 7+_syl). Il est intéressant de noter que ceux-ci ne s'associent pas avec des segments de six syllabes.

Les topics continus et contrastifs sont situés dans un cluster indiquant notamment une corrélation avec une hauteur moyenne basse (AvgHeight_Low) et à des saillances non-initiales (SaliencyMiddle, SaliencyEnd). Il faut néanmoins noter que les topics continus sont situés plus près de l'origine, non loin d'autres points situés dans d'autres quadrants. Ces observations peuvent donc être plus significatives en ce qui concerne les topics contrastifs.

5.4.1.2 Prénoyaux complexes

Figure 51 – Graphique ACM produit à partir des prénoyaux complexes



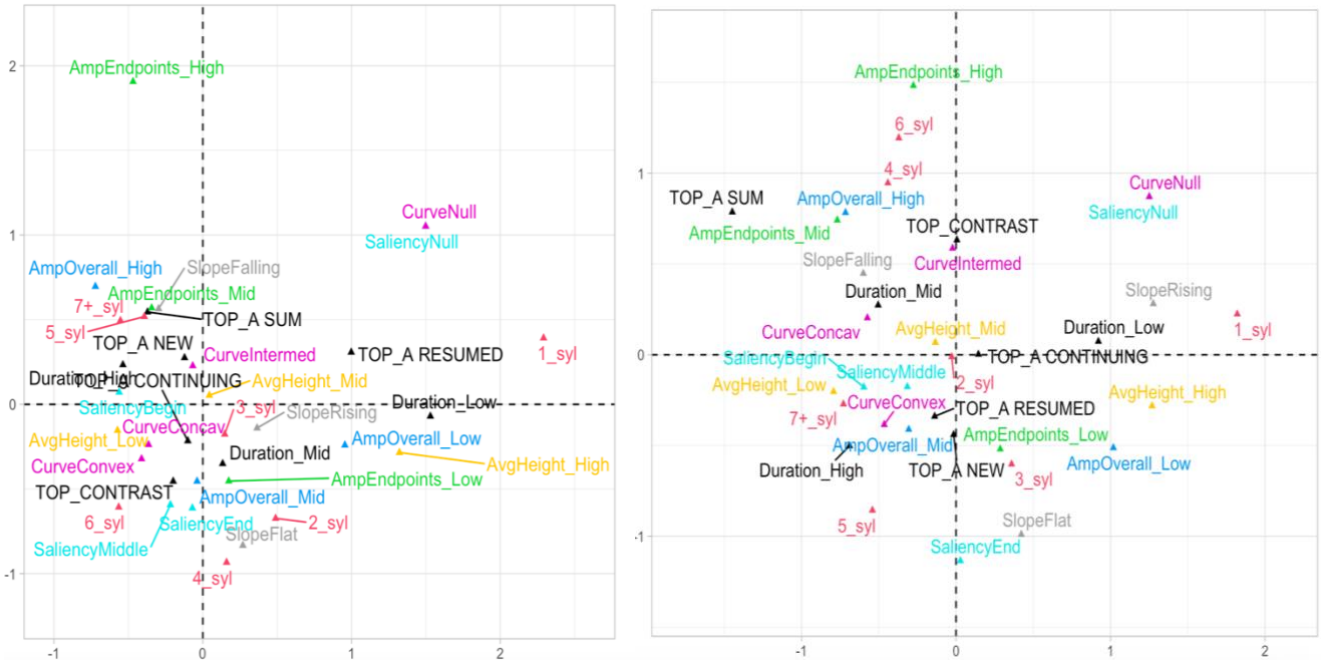
Pour les prénoyaux complexes (Figure 51) on peut également faire un certain nombre d'observations. Dans ce graphique, les topics récapitulatifs semblent être les plus distincts, et sont vaguement corrélés à des amplitudes modérées et élevées selon différentes mesures (AmpEndpoints_High, AmpOverall_High, AmpEndpoits_Mid). Ces topics peuvent également avoir une longueur modérée (4_syl, 6_syl, Duration_Mid) et une pente descendante (SlopeFalling). Cependant, ces points se situent beaucoup plus près des topics contrastifs,

qui sont situés très près des contours dont la saillance interne a une hauteur entre celles de ses valeurs F0 initiales ou finales (CurveIntermed).

Les nouveaux topics et les topics repris sont très proches les uns des autres. Cette section du graphique est très encombrée, ce qui rend difficile l'identification des principaux corrélats de ces topics. Néanmoins, on peut observer une certaine association avec des contours convexes (CurveConvex) et une variation faible ou modérée de la F0 selon la méthode de calcul utilisée (AmpOverall_Mid, AmpEndpoints_Low). Les topics continus, situés au milieu du graphique, semblent être caractérisés par des segments bisyllabiques (2_syl) avec une hauteur moyenne modérée (AvgHeight_Mid).

5.4.1.3 Comparaison des deux graphiques

Figure 52 – Comparaison des deux graphiques ACM produits à partir des traits prosodiques



Cette section a pour but de fournir une brève comparaison des deux graphiques ACM produits en utilisant les différentes caractéristiques prosodiques. Pour faciliter cette comparaison, les deux graphiques sont affichés côte à côte dans la Figure 52. Le graphique produit à partir des topics des prénoyaux simples est situé à gauche, tandis que celui produit à partir de ceux des prénoyaux complexes est situé à droite.

Dans les deux graphiques, on voit que le point correspondant aux topics continus est situé le plus près de l'origine. Nous considérons qu'il s'agit d'une évidence supplémentaire en faveur de notre hypothèse selon laquelle cette catégorie de topics est prosodiquement la moins marquée par des traits prosodiques distinctifs. En dehors de cela, la distribution des topics est très différente dans les deux graphiques. Dans les prénoyaux simples, les topics repris

semblent être les plus distinctifs, alors que les topics récapitulatifs le sont dans les prénoyaux complexes. On peut néanmoins noter que, dans les deux images, les topics récapitulatifs sont associés avec les étiquettes AmpOverall_High et AmpEndpoints_Mid. Ceci suggère que ces types de topics ont souvent des profils similaires en termes de variation de la F0, indépendamment du fait qu'ils apparaissent ou non à côté d'un connecteur ou d'un marqueur de discours.

Plus généralement, cependant, les associations entre les étiquettes pragmatiques et les caractéristiques prosodiques ne sont pas cohérentes entre les deux graphiques. Par exemple, l'étiquette CurveIntermed s'associe très étroitement aux nouveaux topics dans les prénoyaux simples, mais aux topics contrastifs dans les prénoyaux complexes. Les nouveaux topics semblent les plus similaires aux topics récapitulatifs parmi les prénoyaux simples, alors qu'ils sont les plus proches des topics repris dans les prénoyaux complexes. Ce genre de différences est fréquent et difficile à expliquer avec un haut degré de confiance. Il est tout à fait possible que le fait d'être précédé par un connecteur ou un marqueur de discours ait un impact substantiel sur la structure mélodique des topics. Si c'est le cas, il y a une différence fondamentale dans la façon dont ces deux catégories de prénoyaux sont réalisées prosodiquement. Cependant, il est également important de noter que les graphiques analysés dans ce mémoire de recherche ne représentent qu'une minorité de la variation dans notre corpus, et que d'autres similitudes et différences ne peuvent tout simplement pas être observées dans les images que nous avons présentées.

5.4.2 Le test exact de Fisher

Le Tableau 4 présente les p-values calculées pour chaque paire d'un des cinq topics et des 30 étiquettes prosodiques. Par convention, les p-values inférieures au seuil de 0,05 sont considérées comme significatives, indiquant un haut degré de probabilité que les deux étiquettes soient corrélées. Les p-values inférieures à ce seuil sont soulignées en **bleu et gras**. Nous avons également remarqué plusieurs valeurs légèrement supérieures à ce seuil, qui ont également de fortes chances d'être significatives. Les autres valeurs inférieures à 0,10 ont donc été soulignées en bleu.

Tableau 4 – p-values calculés à partir des traits prosodiques

	PRÉNOYAUX SIMPLES					PRÉNOYAUX COMPLEXES				
	CONTIN.	NEW	RESUMED	SUM	CONTRAST	CONTIN.	NEW	RESUMED	SUM	CONTRAST
AmpEndpoints_Low	0.466315	0.826108	0.688195	0.704456	0.280087	0.674068	0.113497	0.378941	1.0	0.960972
AmpEndpoints_Mid	0.304594	0.386757	0.872449	0.890661	0.661326	0.281352	0.956592	0.74966	0.2	0.565445
AmpEndpoints_High	0.965886	0.400599	0.137987	0.222822	0.906504	0.878833	0.811303	0.845819	1.0	0.067437
AmpOverall_Low	0.431881	0.83106	0.079734	0.763765	0.817672	0.387315	0.588716	0.186383	1.0	0.961176

AmpOverall_Mid	0.676929	0.490147	0.958222	0.222638	0.322142	0.695265	0.514913	0.909922	1.0	0.239494
AmpOverall_High	0.588729	0.362529	0.631564	0.878219	0.580056	0.710743	0.695866	0.733528	0.373333	0.407598
SaliencyBegin	0.370763	0.743821	0.919482	0.261129	0.419953	0.46928	0.961478	**0.010688**	1.0	0.840007
SaliencyMiddle	0.424516	0.461503	0.951121	0.97169	0.167756	0.649149	0.253144	1.0	0.133333	0.783395
SaliencyEnd	0.502312	0.846601	0.532276	0.543121	0.572722	0.955395	**0.027781**	1.0	1.0	0.65462
SaliencyNull CurveNull	0.931939	0.28723	**0.008926**	0.569631	0.989482	0.411486	0.86143	0.891601	1.0	0.193145
CurveConvex	0.616423	0.718066	0.992243	0.704456	**0.009867**	0.674068	0.113497	0.378941	1.0	0.960972
CurveIntermed	0.108825	0.644056	0.532276	1.0	0.965069	1.0	1.0	0.296216	1.0	0.445045
CurveConvav	0.420466	0.857593	1.0	0.254178	0.71412	0.672714	1.0	1.0	**0.04**	0.589485
AvgHeight_Low	0.847718	0.238374	0.756864	0.831536	0.413579	0.602032	0.188688	1.0	1.0	0.838754
AvgHeight_Mid	0.439642	0.795835	0.678882	0.357384	0.620811	0.648605	0.977375	0.502867	0.786667	0.157083
AvgHeight_High	0.38389	0.786976	0.282499	1.0	0.868098	0.649149	0.253144	0.503194	1.0	0.957045
SlopeFalling	**0.013383**	0.408761	0.998338	0.260157	0.947033	0.705849	0.774158	0.749307	0.546667	0.27766
SlopeRising	0.874045	0.62219	0.056889	0.843022	0.652524	0.785708	0.401405	0.614992	1.0	0.637254
SlopeFlat	0.979848	0.698303	0.163471	0.833143	0.072574	0.335128	0.57582	0.491933	1.0	0.888914
Duration_Low	0.743652	0.877312	**0.025186**	1.0	0.479544	0.20382	0.885473	0.920349	1.0	0.263044
Duration_Mid	0.285325	0.461503	0.529561	0.590771	0.936148	0.450647	0.82178	0.69832	1.0	0.353677
Duration_High	0.714902	0.40318	0.983576	0.299907	0.208207	0.976956	0.074579	0.145484	0.253333	0.984455
1_syl	0.974651	0.917508	**0.006655**	1.0	0.446631	0.158126	0.739654	0.917311	1.0	0.700147
2_syl	**0.041719**	0.894874	0.579764	1.0	0.760598	0.276329	0.755487	0.811556	1.0	0.724789
3_syl	0.320457	0.937114	0.388434	0.402717	0.644043	0.832897	0.4485	0.441522	1.0	0.724789
4_syl	0.079734	0.951121	0.802992	1.0	0.294135	0.878833	0.811303	0.503194	1.0	0.219256
5_syl	0.99572	**0.030316**	0.906804	0.222822	0.72698	0.602032	0.510503	0.662566	1.0	0.838754
6_syl	0.527003	0.90445	0.576297	1.0	0.391353	1.0	0.287708	1.0	1.0	0.264495
7+_syl	0.929634	**0.030414**	0.952478	0.210449	0.737483	0.673772	0.687978	0.320457	0.32	0.814538

En examinant ces données, on observe que, au niveau des prénoyaux simples, les topics continus sont fortement corrélés aux contours descendants (SlopeFalling), et ont tendance à être composés de deux ou de quatre syllabes (2_syl, 4_syl). Quant aux nouveaux topics, les principaux corrélats semblent être un nombre relativement important de syllabes (5_syl, 7+_syl), bien que paradoxalement ils ne soient pas corrélés avec les segments de six syllabes. Les topics repris sont associés le plus fortement aux contours qui n'ont pas de saillance interne (SaliencyNull/CurveNull), qui ont une faible durée (Duration_Low) et qui sont composés d'une seule syllabe (1_syl). Notez qu'il existe une relation intuitive entre ces trois caractéristiques : les segments d'une seule syllabe ont naturellement tendance à avoir une durée courte, laissant moins de place pour la production d'une saillance interne importante. Les topics contrastifs sont plus fortement associés aux contours convexes (CurveConvex) et, à un moindre degré, aux pentes plates (SlopeFlat). Il est intéressant de noter qu'il n'y a pas de corrélats significatifs des topics récapitulatifs à ce niveau d'analyse.

Parmi les prénoyaux complexes, on voit beaucoup moins de p-values en dessous de 0,1. Néanmoins, on peut observer des relations entre les nouveaux topics et les contours ayant

une saillance située à la fin (SaliencyEnd) et une durée élevée (Duration_High), entre les topics repris et les contours ayant une saillance au début, et entre les topics récapitulatifs et les contours concaves. Il semble également exister une relation entre les topics contrastifs et l'amplitude selon une des deux mesures (AmplitudeEndpoints_High).

Il convient également de noter qu'il existe peu de parallèles entre les caractéristiques prosodiques des topics apparaissant dans des prénoyaux simples et complexes. Par exemple, les topics repris dans les prénoyaux complexes sont fortement associés à une saillance initiale, alors que ceux qui apparaissent dans les prénoyaux simples ont tendance à ne pas avoir de saillance interne. De manière générale, le fait d'avoir une très faible p-value entre une paire d'étiquettes à un niveau d'analyse ne suggère pas une faible p-value entre les étiquettes équivalentes à l'autre niveau. Ceci soutient l'idée qu'il existe une différence prosodique fondamentale entre les topics qui sont précédés par un connecteur ou un marqueur de discours, et ceux qui ne le sont pas. Il est fortement probable que ce facteur ait un impact significatif sur les caractéristiques prosodiques d'un topic. Ce phénomène constitue une piste de recherche intéressante pour de futures études sur les topics du naija.

Une question qui se pose naturellement est de savoir si les p-values produites par le test exact de Fisher reflètent ou non les observations faites en examinant les graphiques MCA produits à partir des mêmes données. En examinant le graphique ACM représentant les topics apparaissant dans les prénoyaux simples (Figure 50 - Graphique ACM produit à partir des prénoyaux simples, page 78), nous avons observé que les étiquettes 5_syl et 7+_syl semblaient fortement corrélées avec les nouveaux topics et les topics récapitulatifs. Ces deux étiquettes prosodiques ont produit des p-values inférieures à 0,05 pour les nouveaux topics, mais pas pour les topics récapitulatifs. Il est néanmoins intéressant de noter que les p-values produites pour les topics récapitulatifs étaient relativement faibles, de l'ordre de 0,2. De même, la relation détectée entre les topics repris et les étiquettes 1_syl, Duration_Low, AmpOverall_Low, et SaliencyNull se reflétait parmi les p-values. Cependant, la relation observée entre les topics continus et les pentes descendantes ou les segments bisyllabiques n'a pas été détectée du tout lors de l'examen du graphique ACM. Inversement, notre observation précédente selon laquelle les nouveaux topics et les topics contrastifs sont associés à des hauteurs faibles et à des saillances non initiales n'est pas reflétée parmi les p-values.

Si l'alignement entre les p-values et les observations faites sur le graphique ACM peut être décrit comme inconsistante au niveau des prénoyaux simples, la divergence est encore plus grande lorsque nous considérons les prénoyaux complexes. Parmi les observations faites en inspectant le graphique ACM représenté dans la Figure 51 (page 79), aucune n'est reflétée

par une p-value inférieure à 0,1. La raison de ce résultat n'est pas entièrement claire, surtout si l'on considère que les dimensions représentées sur les deux graphiques ACM représentent un pourcentage similaire de la variation totale au sein des deux jeux de données. Quoi qu'il en soit, cela suggère qu'il faut être particulièrement prudent lors de l'interprétation de graphiques ACM représentant une minorité de la variation dans un ensemble de données. Alors que l'ACM était plutôt utile dans l'exploration des aspects prosodiques des prénoyaux simples, il était extrêmement trompeur lorsqu'il était appliqué à des topics situés dans des prénoyaux complexes. Pour cette raison, nous estimons que le test exact de Fisher est globalement plus fiable et plus adapté à ce type d'analyse.

Comme exercice final, nous avons décidé de générer une liste de toutes les paires d'étiquettes qui ont produit une p-value inférieure à 0,05. Cela nous a permis de générer un classement définitif des étiquettes les plus fortement corrélées. Le Tableau 5 – les p-values les plus significatives ci-dessous présente toutes les paires de chaque ensemble de données qui sont inférieures à ce seuil, classées par ordre décroissant de signification. Notez que cette liste comprend à la fois des paires composées d'une étiquette prosodique et d'une étiquette pragmatique, ainsi que des paires composées de deux étiquettes prosodiques. En effet, ces dernières constituent l'écrasante majorité de ce tableau. Notez également que parmi les topics situés dans les prénoyaux complexes, un peu moins de p-values ont été générées en dessous du seuil de 0,05.

Tableau 5 – les p-values les plus significatives

PRÉNOYAUX SIMPLES		PRÉNOYAUX COMPLEXES	
Pair d'étiquettes	**p-values**	**Pair d'étiquettes**	**p-values**
7+_syl-Duration_High	0	7+_syl-Duration_High	0
AmpEndpoints_Low-AmpOverall_Low	0	1_syl-SaliencyNull	0
AmpEndpoints_Low-SlopeFlat	0	1_syl-CurveNull	0
AmpEndpoints_High-AmpOverall_High	0	AmpOverall_High-SlopeFalling	0
CurveNull-SaliencyNull	0	CurveNull-SaliencyNull	0
CurveConvex-SaliencyBegin	3.00E-06	1_syl-Duration_Low	1.00E-06
AmpOverall_Low-SlopeFlat	6.00E-06	AmpEndpoints_Low-AmpOverall_Low	1.00E-06
1_syl-SaliencyNull	8.00E-06	AmpOverall_Low-SlopeFlat	1.00E-06
1_syl-CurveNull	8.00E-06	CurveConvex-SaliencyBegin	3.00E-06
AmpOverall_High-Duration_High	9.00E-06	7+_syl-SaliencyBegin	4.00E-06
1_syl-Duration_Low	1.50E-05	1_syl-AmpOverall_Low	4.00E-06
AmpOverall_High-AvgHeight_Low	7.30E-05	AmpEndpoints_High-AmpOverall_High	1.60E-05
Duration_Low-SaliencyNull	8.40E-05	AmpEndpoints_Low-SlopeFlat	1.70E-05
CurveNull-Duration_Low	8.40E-05	AmpOverall_Low-Duration_Low	0.000101
AmpOverall_Low-SaliencyNull	0.000109	Duration_Low-SaliencyNull	0.000147
AmpOverall_Low-CurveNull	0.000109	CurveNull-Duration_Low	0.000147
AmpEndpoints_Mid-AvgHeight_Low	0.000165	AmpOverall_Low-AvgHeight_High	0.000191
AmpEndpoints_High-SlopeFalling	0.000167	AmpEndpoints_Mid-AmpOverall_High	0.000235
AmpOverall_High-SlopeFalling	0.00017	1_syl-SlopeRising	0.000454
AmpEndpoints_Mid-AmpOverall_High	0.00022	AmpEndpoints_Mid-SlopeFalling	0.000553
AvgHeight_Mid-SlopeFalling	0.000404	1_syl-AvgHeight_High	0.001236
AmpOverall_Low-Duration_Low	0.000998	AmpEndpoints_High-SlopeFalling	0.001352
AvgHeight_Low-CurveConvex	0.001194	SaliencyNull-SlopeRising	0.001824
7+_syl-AmpOverall_High	0.0012	CurveNull-SlopeRising	0.001824
7+_syl-SlopeFalling	0.00136	Duration_High-SaliencyBegin	0.00283
7+_syl-AmpEndpoints_High	0.00158	4_syl-AmpOverall_High	0.004461

2_syl-Duration_Low	0.00176	Duration_Low-SlopeRising	0.006676
4_syl-SlopeFlat	0.001883	AmpOverall_Low-SaliencyNull	0.006723
AmpOverall_High-CurveConvex	0.00229	AmpOverall_Low-CurveNull	0.006723
7+_syl-SaliencyBegin	0.002341	7+_syl-CurveConvex	0.007149
4_syl-AmpEndpoints_Low	0.002485	SaliencyBegin-TOP_A RESUMED	0.010688
CurveConvex-SaliencyMiddle	0.002507	CurveConvex-Duration_High	0.011978
CurveConvex-SlopeFlat	0.003465	7+_syl-SlopeFalling	0.013738
6_syl-SaliencyBegin	0.003538	3_syl-Duration_Low	0.014514
Duration_High-SlopeFalling	0.005666	AmpOverall_Low-SlopeRising	0.015128
AmpOverall_Low-AvgHeight_High	0.005799	AvgHeight_High-SaliencyNull	0.018638
1_syl-TOP_A RESUMED	0.006655	AvgHeight_High-CurveNull	0.018638
2_syl-AmpOverall_Mid	0.008535	2_syl-AvgHeight_Low	0.020473
SaliencyNull-TOP_A RESUMED	0.008926	5_syl-Duration_Mid	0.020714
CurveNull-TOP_A RESUMED	0.008926	4_syl-Duration_Mid	0.021832
4_syl-Duration_Mid	0.009416	SaliencyBegin-SlopeFalling	0.023323
CurveConvex-TOP_CONTRAST	0.009867	AmpEndpoints_Low-AmpOverall_Mid	0.024668
1_syl-AmpOverall_Low	0.010459	AmpOverall_Mid-CurveConvex	0.024668
Duration_High-SaliencyBegin	0.012629	1_syl-AmpEndpoints_Low	0.026776
SlopeFalling-TOP_A CONTINUING	0.013383	SaliencyEnd-TOP_A NEW	0.027781
AmpEndpoints_High-AvgHeight_Mid	0.013385	AmpEndpoints_Mid-Duration_Mid	0.032816
AmpOverall_High-SaliencyBegin	0.013844	AvgHeight_Mid-SlopeFalling	0.032871
SaliencyEnd-SlopeRising	0.014639	AmpOverall_High-SaliencyBegin	0.038791
AmpEndpoints_Low-CurveConvex	0.01618	CurveConcav-TOP_A SUM	0.04
Duration_Low-SlopeRising	0.01656	AmpEndpoints_Low-SlopeRising	0.040394
AmpOverall_Mid-Duration_Mid	0.017647	SaliencyEnd-SlopeFlat	0.043215
AmpEndpoints_Mid-SlopeRising	0.018065	AmpOverall_High-Duration_Mid	0.045011
AmpEndpoints_Mid-SlopeFalling	0.019907	Duration_High-SlopeFalling	0.046912
AmpEndpoints_Low-SaliencyMiddle	0.02147	AvgHeight_High-SlopeRising	0.047993
2_syl-SlopeRising	0.02338	5_syl-AmpOverall_Mid	0.04826
AmpEndpoints_Low-AmpOverall_Mid	0.024159	7+_syl-AmpEndpoints_Mid	0.049994
CurveConvex-Duration_High	0.02467		
Duration_Low-TOP_A RESUMED	0.025186		
4_syl-AmpOverall_Low	0.028925		
5_syl-TOP_A NEW	0.030316		
7+_syl-TOP_A NEW	0.030414		
5_syl-SlopeFalling	0.035036		
AvgHeight_Low-SlopeRising	0.037105		
3_syl-SlopeRising	0.040751		
2_syl-TOP_A CONTINUING	0.041719		
AmpEndpoints_Low-AvgHeight_Mid	0.043192		
CurveIntermed-SlopeFalling	0.043768		
2_syl-SaliencyMiddle	0.045643		
AmpEndpoints_High-Duration_High	0.049064		

Les paires les plus significatives statistiquement sont vraies pour les deux ensembles de données, bien qu'elles ne soient pas particulièrement informatives en ce qui concerne la relation entre la prosodie et la structure de l'information dans le naija. Les topics composés de sept syllabes ou plus ont également une longue durée, tandis que les segments monosyllabiques ont, logiquement, une courte durée et manquent d'une saillance interne. La longueur est également fortement corrélée à un degré élevé de variation de F0. Sans surprise, il y a également une relation étroite entre les deux méthodes utilisées pour calculer la variation de F0. L'observation la plus intéressante sur le plan linguistique est sans doute que, dans les deux segments, les contours ascendants sont étroitement associés aux segments de faible durée, tandis que les contours descendants sont associés aux segments plus longs. La raison de ce phénomène n'est pas tout à fait claire et mérite une exploration plus approfondie. De même, il est prouvé que les contours longs sont caractérisés par une saillance interne initiale, alors que les contours montants ont tendance à avoir une saillance interne finale. Bien que la

raison exacte de ce parallèle ne soit pas encore claire, nous pensons qu'un contour descendant est perceptivement comparable à un contour avec une saillance initiale. En d'autres termes, le contour HL serait perçu de la même manière qu'un contour tel que mLH1, avec une forte saillance interne élevée au début.

En explorant ce phénomène, nous avons également remarqué que les saillances initiales élevées correspondaient souvent au premier élément lexical du segment. Cela pourrait indiquer que la différence entre HL et mLH1 est en partie morphosyntaxique, déterminée par le fait que le segment commence ou non par un morphème lexical ou fonctionnel.

6 Conclusion et perspectives

Bien que la plupart des graphiques ACM générés dans cette étude étaient pratiquement ininterprétables pour les raisons décrites précédemment, nous avons néanmoins réussi à générer des données utiles et intéressantes qui ont servi à valider nos trois hypothèses initiales. Cependant, nous pensons que plusieurs problèmes clé ont empêché ce projet d'atteindre son plein potentiel. Tout d'abord, la relative pauvreté des données. Seuls 23 des 80 monologues présents dans le corpus Gold ont été utilisés dans cette étude. Si les futurs chercheurs sont en mesure d'effectuer des annotations pragmatiques pour les 57 fichiers restants, il serait possible d'examiner environ quatre fois plus de prénoyaux. Cela permettrait également de réduire le nombre d'étiquettes pragmatiques ou de contours prosodiques stylisés qui n'apparaissent que quelques fois dans le corpus.

Un total de 224 types de contours prosodiques stylisés étaient présents dans le corpus, dont beaucoup n'apparaissaient qu'une fois ou un petit nombre de fois. Cela a très probablement nui à l'utilité des graphiques qui tenaient compte de tous ces contours. Pour cette raison, et parce que la présence de 224 points distincts rendait la lecture des graphiques ACM difficile, nous avons basé nos analyses principalement sur des graphiques qui ne prenaient en compte que les contours les plus fréquents. Cependant, cette approche présente également des inconvénients importants. En n'incluant que les contours les plus fréquents dont le nombre d'occurrences combinées représente 50% de l'ensemble des données, nous avons effectivement limité la couverture de nos analyses en excluant la moitié de nos données.

Nos approches successives, qui consistaient à décomposer chaque contour en un ensemble de caractéristiques prosodiques essentielles, nous ont également permis de faire des observations plus fondamentales sur les caractéristiques prosodiques de chaque type pragmatique. Le texte exact de Fisher, que nous avons appliqué au cours de cette phase de notre analyse, s'est également révélé être un outil utile pour quantifier les relations entre

différentes variables, mais a également mis en évidence certaines des faiblesses lorsqu'il s'agit d'interpréter les graphiques ACM. Ces tests ont également contribué à confirmer nos hypothèses selon lesquelles il existe des différences prosodiques importantes entre les différents types de topics. Ce mémoire de recherche avait pour but d'appliquer une approche statistique originale à l'étude de la relation entre la prosodie et la structure informationnelle, en appliquant ces méthodes à un corpus de naija. Cependant, ces approches pourraient théoriquement être appliquées à n'importe quelle autre langue, à condition qu'il existe un corpus de parole aligné. De telles études pourraient permettre de montrer quelles sont les observations faites qui sont spécifiques au naija.

Bibliographie

Arnold, Jennifer, Elsi Kaiser, Jason M. Kahn, et Lucy Kyoungsook Kim (2015). “Information Structure: Linguistic, Cognitive, and Processing Approaches”. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4491328/

Avanzi, Mathieu, Anne Lacheret-Dujour, Bernard Victorri. “ANALOR. A Tool for Semi-Automatic Annotation of French Prosodic Structure”, Mai 2008, Campinas, Brazil. pp.119-122. hal-00334656

BBC. “BBC starts Pidgin digital service for West Africa audiences.” (17 août 2017). https://www.bbc.com/news/world-africa-40975399

Bear, John et Price, Patti. 1990. “Prosody, Syntax, and Parsing”. *ACL '90: Proceedings of the 28th annual meeting on Association for Computational Linguistics*. Juin 1990. Pp. 17–22 https://doi.org/10.3115/981823.981826

Bernandy, Jean-Philippe et Themistocleous, Charalambos (2017). Modelling prosodic structure using Artificial Neural Networks

Boersma, Paul et van Heuven, Vincent. 2001. Speak and unSpeak with PRAAT. In *Glot International*, vol. 5, 341–347. Oxford, UK: Blackwell Publishing Ltd.

Brezina, Vaclav (2018). *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge University Press.

Bigi, Brigitte, Bernard Caron et Oyelere Abiola. Developing Resources for Automated Speech Processing of the African Language Naija (Nigerian Pidgin). 8th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, Nov 2017, Poznan, Poland. pp.441-445. https://hal.archives-ouvertes.fr/hal-01705707/document

Chomsky, Noam. (1957). *Syntactic Structures*. Mouton.

Desagulier, Guillaume (2017). *Corpus Linguistics and Statistics with R: Introduction to Quantitative Methods in Linguistics*. Springer.

Edionhon, Edosa James (14 octobre 2018). “West Africa’s pidgins deserve full recognition as official languages”. *Quartz Africa*. <https://qz.com/africa/1423524/west-africas-pidgin-can-be-official-language-bbc-shows/>

Courtin, Marine, Bernard Caron, Kim Gerdes, Sylvain Kahane. Establishing a Language by Annotating a Corpus: The Case of Naija, a Post-creole Spoken in Nigeria. annDH 2018 Annotation in Digital Humanities, Aug 2018, Sofia, Bulgaria. pp.7-11. https://halshs.archives-ouvertes.fr/halshs-01958330/document

Huber, Magnus. 1999. *Ghanaian Pidgin English in its West African Context: A sociohistorical and structural analysis*. John Benjamins Publishing Company. Varieties of English Around the World, G24.

Gerdes, Kim, Bruno Guillaume, Sylvain Kahane, Guy Perrier (2018). “SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD”.

Guibon, Gaël et al. (2020) “When Collaborative Treebank Curation Meets Graph Grammars: Arborator with a Grew Back-End”. Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 5291–5300. https://aclanthology.org/2020.lrec-1.651.pdf

Guillaume, Bruno (2021). “Graph Matching and Graph Rewriting: GREW tools for corpus exploration, maintenance and conversion”. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 168–175 April 19 - 23, 2021. https://aclanthology.org/2021.eacl-demos.21.pdf

Guiller, Kirian (2020). *Analyse syntaxique automatique du pidgin-créole du Nigeria à l’aide d’un transformer (BERT) : Méthodes et Résultats* (mémoire de master). http://www.tal.univ-paris3.fr/plurital/memoires/kirian_GUILLER-RD-1920.pdf

Lacheret-Dujour, Anne et Beaugendre, Frédéric. 1999. *La Prosodie du français*. CNRS Éditions.

Lacheret-Dujour, Anne, Sylvain Kahane, Paola Pietrandrea (eds.). 2018. *Rhapsodie: A Prosodic and Syntactic Treebank for Spoken French*.

Lacheret-Dujour, Anne, Guillaume Desagulier, Serge Fleury et Frédéric Isel (2018). “The distribution of prosodic features in the Rhapsodie corpus from general observations to discourse characterization”. Dans Anne Lacheret-Dujour, Sylvain Kahane, Paola Pietrandrea (eds.) *Rhapsodie: A Prosodic and Syntactic Treebank for Spoken French*. Benjamins 2018.

Liu, Luigi, Anne Lacheret-Dujour, Nicolas Obin. Automatic Modelling and Labelling of Speech Prosody: What’s New with SLAM+ ?. International Congress of Phonetic Sciences (ICPhS), Aug 2019, Melbourne, Australia. hal-02119926

McDonald, John H. 2014. *Handbook of Biological Statistics* (3rd ed.). Sparky House Publishing, Baltimore, Maryland. Pages 77-85. http://www.biostathandbook.com/fishers.html

Oyelere, Biola (2021). Forthcoming PhD thesis. Paris Nanterre University.

Florence Agbo, O., & Plag, I. (2020). “The Relationship of Nigerian English and Nigerian Pidgin in Nigeria: Evidence from Copula Constructions in Ice-Nigeria”, *Journal of Language Contact*, 13(2), 351-388. doi: https://doi.org/10.1163/19552629-bja10023

Nasri, M.K. et Caelen-Haumont, G. (1991) “Utilisation de règles prosodiques en reconnaissance de la parole”, *18e Journées d’étude sur la parole*, Société française d’acoustique (éd.), Montréal, pp. 276-280.

Pietrandrea, Paola and Sylvain Kahane (2018). “Macrosyntactic Annotation”. Dans Anne Lacheret, Sylvain Kahane, Paola Pietrandrea (eds.) *Rhapsodie: A Prosodic and Syntactic Treebank for Spoken French*. Benjamins 2018.

Pépiot, Erwan. 2014. “Male and female speech: a study of mean f0, f0 range, phonation type and speech rate in Parisian French and American English speakers”. Speech Prosody 7, Mai 2014, Dublin, Ireland. pp.305-309.

Pérennou, G. et Caelen-Haumont, G. (1982). “Utilisation de la prosodie pour la reconnaissance de la parole dictée”, dans A. Di Cristo *et al.* (1982 éd.) *Prosodie et reconnaissance automatique de la parole*, Publications du CRECO Communication Parlée. pp. 25-27.

Ranganath, Rajesh *et al*. 2009. “It’s Not You, it’s Me: Detecting Flirting and its Misperception in Speed-Dates”. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp. 334–342, Singapore.

Simard, Candide (2010). “The Prosodic Contours of Jaminjung, a Language of Northern Australia” (thèse doctorale).

Simard, Candide, Anne Lacheret-Dujour, et Biola Oyelere (2019). “Broad and Narrow focus marking in Naija (Nigerian Pidgin): the role of prosody”. ICPhS 2019, Australie. https://www.researchgate.net/publication/339051551_Broad_and_Narrow_focus_marking_in_Naija_Nigerian_Pidgin_the_role_of_prosody

Song, Yuchen (2020), *Extraction de lexiques syntaxiques à partir d'un treebank - Le cas du treebank SUD du naija, un pidgin créole de l'anglais* (mémoire de master). http://www.tal.univ-paris3.fr/plurital/memoires/song_yuchen-RD-1920.pdf

Tesnière, Lucien (1959). *Éléments de syntaxe structurale.*

Wallis, Sean (2021). *Statistics in Corpus Linguistics Research: A New Approach*. Routledge.