Chapitre 5

Annotations et mesures

Dans ce chapitre, nous présentons les traitements et annotations effectués sur les données collectées, ainsi que la méthodologie adoptée pour analyser la distribution syntaxique des pauses et les patterns d'accentuation lexicale chez les locuteurs de niveaux CECRL B1 et B2. L'ensemble des annotations est réalisé de manière automatique, et compilé dans une chaîne de traitements conçue dans le cadre de cette thèse. Nous avons nommé cet outil *Pauses and Lexical Stress Processing Pipeline (PLSPP)*, et l'avons publié en open-source sur gricad-gitlab.univ-grenoble-alpes.fr[1].

Nous avons tenu à ce que la totalité des traitements soit effectuée de manière automatique pour deux raisons. En premier lieu, nous souhaitons permettre, à terme, à des enseignants ou des évaluateurs sans expertise spécifique en informatique ou en phonétique d'analyser leurs propres corpus d'enregistrements. De plus, nous espérons que cet outil pourra être adapté et intégré dans un système d'évaluation automatique tel que SELF, afin de prendre en compte les schémas de pauses et d'accentuation lexicale dans l'évaluation de la production orale spontanée des apprenants.

PLSPP est constitué de cinq blocs modulables pouvant être adaptés en fonction des besoins (*cf.* figure 5.1). Les trois premiers blocs sont des modules de pré-traitement des données. Le premier bloc permet de segmenter les enregistrements en locuteurs et d'extraire des segments de parole qui seront analysés par les modules suivants. Le second fournit une transcription orthographique de ces segments et un alignement temporel de chaque mot au signal de parole. Le troisième effectue enfin une analyse syntaxique des énoncés. Les deux derniers modules constituent quant à eux le cœur de notre contribution, en générant une annotation des pauses et de l'accentuation lexicale.

[1] https://gricad-gitlab.univ-grenoble-alpes.fr/lidilem/plspp/

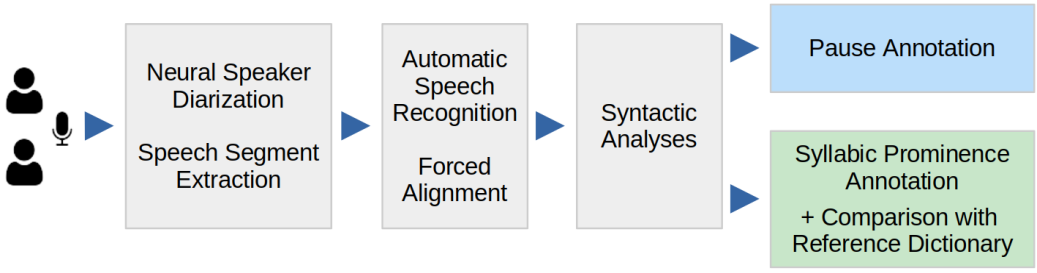

FIG. 5.1 : Architecture générale de PLSPP

La première section de ce chapitre présente les trois modules de pré-traitement ainsi que les métriques utilisées pour les évaluer. La deuxième et la troisième sections portent respectivement sur l'annotation des patterns de pauses et sur l'estimation de la position de l'accent lexical et du contraste accentuel, les métriques d'évaluation choisies, et les différentes mesures utilisées pour comparer les tendances entre locuteurs B1 et B2.

5.1 Modules de pré-traitement

5.1.1 Segmentation en locuteurs

Les enregistrements de parole des trois corpus que nous avons constitués sont des conversations entre deux ou trois locuteurs, il est donc avant tout nécessaire de les segmenter en locuteurs, de manière à pouvoir relier chaque énoncé au locuteur correspondant. Ces énoncés pourront ensuite être analysés indépendamment et par locuteur. Cette tâche de segmentation implique dans un premier temps de détecter l'activité de parole dans l'enregistrement *(voice activity detection)*, de détecter les changements de locuteurs *(speaker change detection)*, puis d'associer chaque segment de parole au locuteur correspondant *(speaker identification)*. Ces trois étapes sont rassemblées sous le terme de diarisation des locuteurs *(speaker diarization)*.

L'outil choisi pour effectuer cette segmentation est pyannote.audio (Bredin, 2023 ; Bredin et al., 2020). Il s'agit d'une boîte à outils open-source développée par l'Institut de Recherche en Informatique de Toulouse (IRIT), et permettant de combiner différents modules de traitement pour effectuer une diarisation en locuteurs. Pyannote

met à disposition des modèles pré-entraînés pouvant être utilisés *out of the box* et qui obtiennent des résultats jugés suffisants pour notre étude[2].

En sortie de pyannote, on obtient pour un enregistrement audio donné, une liste des temps de début et de fin de chaque segment de parole détecté et le locuteur identifié. L'étape suivante consiste à fusionner les segments consécutifs du même locuteur. Toutefois, il arrive qu'un locuteur B réagisse aux propos de son partenaire A sans pour autant lui prendre la parole, par exemple en disant "*I see*" ou "*yeah*". Si nous segmentons précisément ces énoncés, nous nous retrouvons avec un premier énoncé incomplet de A, puis un segment très court de B, et un autre segment contenant la suite de l'énoncé de A. Comme nous souhaitons observer la distribution syntaxique des pauses par la suite, il est pertinent de conserver des énoncés aussi complets que possible, même si quelques mots de l'interlocuteur sont présents. La difficulté ici est de savoir à partir de quand on considère que la prise de parole de l'interlocuteur doit constituer un segment à part entière.

Implémentation

Nous proposons de combiner la diarisation de Pyannote avec un script de compilation de segments de notre facture, afin d'obtenir des segments de parole proches de ce que nous pourrions considérer comme des tours de parole. Un troisième script vient ensuite découper l'enregistrement audio en fonction de ces segments pour obtenir une liste de fichiers qui seront analysés indépendamment par la suite.

Diarisation Dans notre pipeline, les modules de diarisation de Pyannote v2.1 sont appelés par le script `diarisationPyannote.py`. Ce script prend en entrée les enregistrements audio du corpus et renvoie un fichier texte par enregistrement, listant chaque segment de parole détecté et le locuteur identifié.

Compilation Le script `pyannote2TextGrid.py` convertit ensuite les fichiers obtenus au format TextGrid, avec une tier par locuteur, et fusionne les segments consécutifs. Un seuil de durée paramétrable permet de jouer sur ce découpage : il correspond à la durée de silence en secondes pour un locuteur donné à partir duquel on souhaite que le découpage se fasse. Plus le seuil est élevé, plus les segments consécutifs d'un locuteur auront tendance à être fusionnés, au risque toutefois de contenir de longs silences ou des réactions assez longues de l'interlocuteur. À l'inverse, plus le seuil est bas, plus les segments seront courts et contiendront peu de silences ou de réactions de l'interlocuteur. En contrepartie, les énoncés seront souvent incomplets. Ce seuil est

[2]Pyannote v2.1 obtient par exemple un taux d'erreur moyen de 18,9 % sur le corpus d'interactions en réunions professionnelles AMI-IHM *(Augmented Multi-Party Interaction - Individual Headset Microphone)* (Bredin, 2023).

Sortie brute de Pyannote

```
start=41.46s, stop=50.14s, speaker_A
start=52.29s, stop=56.00s, speaker_A
start=56.59s, stop=57.10s, speaker_A
start=57.03s, stop=57.78s, speaker_B
start=57.73s, stop=65.49s, speaker_A
start=64.85s, stop=75.08s, speaker_B
start=67.47s, stop=68.10s, speaker_A
```

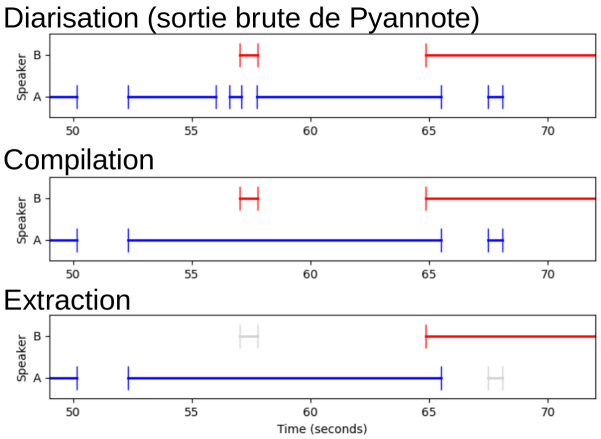

FIG. 5.2 : Exemple de segmentation en locuteurs, avec la sortie brute de Pyannote à gauche, sa visualisation graphique (en haut à droite), la compilation des segments consécutifs (seuil de durée fixé à 1 s*), et les segments extraits (sup. ou égal à* 8 s*)*

fixé à 1 s par défaut, ce qui signifie qu'un segment est coupé à partir d'1 s de silence pour le locuteur courant.

Extraction Le script `intervalles2wav.praat` extrait enfin chaque segment de parole en fichiers audio indépendants. Un paramètre permet de fixer la durée minimum des segments à extraire, par défaut 8 s, et un autre la marge de découpage avant et après le segment, par défaut 10 ms.

En sortie de ce module, chaque enregistrement se retrouve donc découpé en autant de fichiers audio qu'il contient de segments de parole de la durée minimum paramétrée. Chaque fichier sera ensuite analysé séparément par les modules suivants. La figure 5.2 illustre les trois étapes du module de segmentation : la diarisation de Pyannote, la compilation des segments consécutifs et l'extraction des segments d'une certaine durée.

Évaluation

Il convient ensuite de vérifier la qualité de la segmentation et de l'identification du locuteur, pour s'assurer que la majorité des mesures effectuées par la suite seront attribuées au bon locuteur. La qualité de la diarisation est communément évaluée au moyen du taux d'erreur de diarisation *(Diarization Error Rate, DER)*, qui se calcule de la façon suivante :

$$DER = \frac{\text{False alarm} + \text{Missed detection} + \text{Confusion}}{\text{Speech duration}} \tag{5.1}$$

avec *False alarm* correspondant à la durée de non-parole détectée comme parole, *Missed detection* la durée de parole détectée comme non-parole, et *Confusion* la durée de parole attribuée au mauvais locuteur. Le DER permet d'identifier en une seule valeur la qualité globale de la diarisation en termes de détection de parole et d'identification du locuteur.

Nous avons d'abord calculé le DER des sorties de Pyannote sur les 40 binômes du corpus *gold standard* présenté dans le chapitre précédent, et pour lesquels nous disposons d'une diarisation vérifiée manuellement. Nous avons ensuite interprété les valeurs obtenues à partir des proportions de *false alarm*, *missed detection* et *confusion*.

Nous avons également cherché à quantifier la présence de l'interlocuteur dans chaque segment de parole extrait par PLSPP, puisque ce sont ces segments de parole qui seront ensuite analysés par les modules de traitements suivants. En effet, il arrive que certains interlocuteurs réagissent sans pour autant prendre la parole du locuteur, ou bien qu'il y ait un chevauchement de parole. Cela aboutit à la présence de parole de l'interlocuteur dans les segments de parole du locuteur, et donc une potentielle mauvaise attribution de patterns de pauses ou d'accent lexical. Nous proposons le calcul d'un « indice d'interférence » I_L, où l'indice d'interférence I du locuteur L correspond à la proportion de parole D de l'interlocuteur survenant à l'intérieur des segments de parole attribués au locuteur L. Le calcul est effectué comme suit :

$$I_L = \frac{D_{autres,L}}{D_L} \tag{5.2}$$

I_L donnera donc un pourcentage correspondant à la durée de parole de L correspondant en réalité à l'interlocuteur.

5.1.2 Reconnaissance et alignement de la parole

L'étape suivante consiste à transcrire la parole des locuteurs et à aligner temporellement chaque mot de la transcription au signal. Ceci permettra, par la suite, de localiser les pauses dans leur contexte syntaxique, et de cibler les mots sur lesquels nous souhaitons effectuer les mesures acoustiques pour analyser les patterns accentuels.

Nous avons commencé par comparer les performances de plusieurs systèmes de reconnaissance de la parole sur un échantillon de 17 extraits des premiers enregistrements effectués pour le corpus CLES-FR. Ces 17 extraits sont des segments de parole extraits manuellement, d'une durée de 15 à 45 s chacun, transcrits manuellement.

```
REF: i think we can find * compromise to that ** import of technology in classroom        i think
HYP: i think we can find a compromise to that in  terms of technology in       the industry *****
                        I                   I      S                          S        S     D
```

FIG. 5.3 : Calcul du taux d'erreurs de mots (WER) pour un segment issu du corpus CLES-FR (WER=40%, REF : transcription manuelle, HYP : transcription automatique)

Les systèmes testés étaient Google Speech Cloud API[3] (v2.29), EML Transcription[4] (v1.19), Amberscript[5] (v1.3), Fraunhofer Speech Recognition[6] (v2.13), Radboud University LST[7] (v1.1), et SpeechBrain (Ravanelli et al., 2021) (v0.5.11). Chaque système a été testé avec le modèle de reconnaissance associé en anglais britannique et/ou américain ; et deux modèles open-source ont été utilisés dans le cas de SpeechBrain, l'un entraîné sur CommonVoice[8] et l'autre sur LibriSpeech[9].

Pour chaque système, nous avons calculé le taux d'erreur de mots *(Word Error Rate, WER)*. Il s'agit d'une métrique standard utilisée pour évaluer la précision des systèmes de reconnaissance automatique de la parole. Elle mesure le taux d'erreurs dans une transcription générée automatiquement par rapport à une transcription de référence. Le WER est défini comme suit :

$$WER = \frac{S + D + I}{N} \tag{5.3}$$

où N est le nombre de mots de référence, S est le nombre de substitutions (mots incorrectement reconnus), D est le nombre de suppressions (mots omis), et I est le nombre d'insertions (mots ajoutés). Le WER est exprimé en pourcentage, avec des valeurs plus faibles indiquant une meilleure performance du système.

À l'issue de cette analyse comparative préliminaire, dont les résultats sont présentés en annexe C, nous avons choisi d'utiliser le système Fraunhofer Speech Recognition, qui obtenait le taux d'erreur le plus faible parmi les systèmes gratuits. Cependant, quelques mois après ce travail préliminaire, un nouveau système appelé Whisper (Radford et al., 2022) a été publié, réduisant de moitié le taux d'erreur que nous avions jusqu'alors. Nous avons donc choisi d'utiliser ce système pour analyser nos corpus de parole spontanée. Une évaluation plus fine des résultats obtenus avec Whisper est présentée dans le chapitre 7.

[3] Google Speech Cloud (2022) : https://cloud.google.com/speech-to-text/
[4] EML Transcription (2022) : https://www.eml.org/
[5] Amberscript (2022) : https://www.amberscript.com/
[6] Fraunhofer Speech Recognition (2022) : https://www.idmt.fraunhofer.de
[7] LST (2022) : https://webservices.cls.ru.nl/
[8] https://huggingface.co/speechbrain/asr-wav2vec2-commonvoice-en (consulté avril 2022)
[9] https://huggingface.co/speechbrain/asr-crdnn-rnnlm-librispeech (consulté avril 2022)

Concernant l'alignement de la transcription au signal audio, nous avons également testé plusieurs systèmes sur un échantillon de notre corpus francophone. Les systèmes testés étaient WebMAUS v3.4 (Kisler et al., 2017), Montreal Forced Aligner v2.0 (McAuliffe et al., 2017) et Wav2Vec v2.0 (Baevski et al., 2020). Si les deux premiers systèmes ont eu tendance à générer des alignements plus précis au niveau des frontières de mots, ils sont apparus moins robustes à la spontanéité de la parole et aux nombreuses hésitations des locuteurs et finissent par être complètement décalés par rapport au signal de parole. De son côté, Wav2Vec est resté généralement robuste aux hésitations, mais les frontières de mots ont tendance à raccourcir légèrement la première ou la dernière syllabe. Nous avons toutefois choisi ce dernier pour nos analyses.

Implémentation

La reconnaissance automatique de Whisper et l'alignement forcé de Wav2Vec ont été implémentés dans PLSPP par le biais d'un outil appelé WhisperX (Bain et al., 2023), qui combine justement ces deux systèmes. À partir d'un enregistrement audio, WhisperX fournit une transcription orthographique avec un *timestamp* pour chaque mot, indiquant son temps de début et de fin dans l'enregistrement.

Transcription et alignement WhisperX est exécuté par le script `myWhisperxTG.py`. Celui-ci prend en entrée les segments de parole précédemment extraits et renvoie la transcription alignée au mot en format TextGrid pour chaque fichier audio. Le script accepte plusieurs arguments : le modèle utilisé, le type de processeur (par défaut CPU et GPU en parallèle) et plusieurs paramètres techniques ajustables en fonction du serveur utilisé.

Évaluation

L'évaluation de la qualité de la reconnaissance de la parole avec Whisper a fait l'objet d'une évaluation approfondie sur 40 locuteurs et 3 h de parole grâce à la partie manuellement transcrite du corpus CLES-FR. Le taux d'erreur de mots a été calculé pour chaque segment (n=349) et chaque locuteur. De plus, nous avons calculé le taux d'insertions (IR), de délétions (DR), et de substitutions (SR) de mots par locuteur, afin de déterminer quels types d'erreurs sont les plus fréquents.

Évaluer la qualité de l'alignement temporel des mots au signal s'est révélé toutefois plus compliqué car cela nécessite de disposer d'un alignement de référence. Or, nous ne disposons pas d'un alignement manuel ou corrigé pour la portion de corpus *gold standard*. Nous disposons, en revanche, de deux enregistrements de parole spontanée monologuée de locuteurs francophones de l'anglais, provenant d'une étude

effectuée en parallèle de cette thèse (Frost et al., 2024), et dont l'alignement temporel des mots a été vérifié manuellement. Il s'agit de l'enregistrement de deux enseignants francophones donnant un cours en anglais lors d'une école d'été en météorologie à l'université Grenoble Alpes. Les deux enregistrements sont courts, 3 min48 s et 3 min34 s, mais présentent une parole proche de celle de notre corpus principal, à savoir de l'anglais spontané produit par des locuteurs francophones en contexte formel. Une transcription orthographique a d'abord été obtenue avec Whisper, puis corrigée manuellement, et enfin alignée temporellement au signal audio avec WebMAUS. Cet alignement a ensuite été édité manuellement par un phonéticien et constitue ainsi notre alignement de référence.

Pour évaluer la qualité de l'alignement automatique de Wav2Vec à partir de cet alignement de référence, nous nous sommes inspirés de métriques proposées par V. Martin et al. (2024) pour évaluer la qualité de l'alignement automatique des phonèmes. Les auteurs calculent un rappel R correspondant à la durée d'alignement correct des phonèmes divisé par la durée totale des phonèmes de l'alignement de référence, et une mesure de précision P correspondant à la durée d'alignement correct sur la durée totale d'alignement. Nous proposons d'effectuer les mêmes mesures au niveau des mots. Ainsi, R renseigne sur la proportion de durée des mots de l'alignement de référence qui a été correctement alignée, et P indique la proportion de durée d'alignement automatique correspondant effectivement aux mots cibles de l'alignement de référence. Plus les deux valeurs sont proches de 1, plus l'alignement automatique est proche de l'alignement de référence. La qualité de l'alignement automatique de Wav2Vec 2.0 a par ailleurs été comparée à celles de WebMaus v3.4 et Montreal Forced Aligner v2.0.

5.2 Analyses syntaxiques

Deux types d'analyses syntaxiques sont effectuées à partir de la transcription orthographique issue du module précédent : un étiquetage morphosyntaxique pour déterminer la catégorie grammaticale de chaque mot, et une analyse par constituants pour obtenir un arbre syntaxique et regrouper les mots en syntagmes et en propositions.

Étiquetage morphosyntaxique Il est effectué par Spacy (Honnibal et al., 2020). Le script correspondant est `spacyTextgrid_v2.py`. Il prend en entrée le fichier TextGrid contenant la transcription alignée et renvoie le même fichier avec une tier supplémentaire indiquant la catégorie de chaque mot. Les paramètres sont le nom du modèle (par défaut `en_core_web_md`) et le nom de la tier contenant l'alignement des mots.

Analyse par constituants Elle est effectuée par Berkeley Neural Parser (Kitaev et al., 2019) via le script `text2benepar.py`. Celui-ci prend en entrée le texte brut de la transcription et génère un fichier texte contenant le résultat de l'analyse par constituants. Il prend en arguments le modèle d'analyse, par défaut `benepar_en3`[10].

Nous n'avons pas effectué d'évaluation spécifique sur la qualité de ces analyses syntaxiques et il s'agit là d'une limite importante sur laquelle nous revenons dans le chapitre 10.

5.3 Annotation des pauses

Comme indiqué par plusieurs études précédentes (de Jong, 2016 ; Kahng, 2018 ; Kallio et al., 2022 ; Suzuki & Kormos, 2020, entre autres), la distribution des pauses est dépendante de la syntaxe de l'énoncé. On aura ainsi tendance à observer les pauses en frontière de constituants plutôt qu'à l'intérieur de ceux-ci ; et plus le nombre de pauses intra-constituant est élevé, plus la parole a tendance à être jugée disfluente. Nous souhaitons donc localiser les pauses et les catégoriser en fonction de leur contexte syntaxique.

5.3.1 Détection des pauses & caractérisation syntaxique

À ce stade, nous disposons de l'alignement temporel des mots au signal de parole, et par extension, les intervalles éventuels entre les mots, causés par la présence de silences, d'hésitations ou de tout autre élément non transcrit par le système de reconnaissance de la parole. Ces intervalles « vides » seront considérés comme des pauses si leur durée est supérieure ou égale à un seuil minimum défini par l'utilisateur. Ainsi, comme Fauth et Trouvain (2018), nous entendrons par « pause » toute interruption du flux de parole d'une certaine durée, qu'il s'agisse de pauses silencieuses ou pleines, de faux départs, de répétitions ou d'allongements.

Étant donné l'importante variabilité des seuils de durée minimum (et parfois maximum) des pauses dans la littérature, nous avons souhaité éviter de contraindre l'annotation de PLSPP en fonction d'une valeur prédéfinie. Aussi, l'annotation est effectuée sur l'ensemble des intervalles « vides » présents dans l'alignement de la transcription, et ce n'est que lors du traitement ou de la visualisation des résultats que l'utilisateur peut définir un seuil de durée minimum et maximum, pour ne considérer comme pauses que les intervalles d'une certaine durée.

[10] Les arbres syntaxiques peuvent être visualisés directement avec un outil tel que `RSyntaxTree` de Yōichirō Hasebe : https://yohasebe.com/rsyntaxtree.

Annotation des pauses L'étiquetage des pauses en fonction de leur position dans l'énoncé est effectué sur la base de l'analyse syntaxique par constituants issu du module précédent, qui segmente et hiérarchise l'énoncé en propositions et en syntagmes. Le script `pausesAnalysis.py` prend en entrée les transcriptions alignées au format TextGrid et les analyses par constituants au format texte, et renvoie un tableau listant tous les intervalles « vides », leur durée et leur contexte syntaxique : mots précédant et suivant ainsi que leur catégorie, type du plus grand constituant se terminant et commençant ainsi que le nombre de mots qu'ils contiennent et leur profondeur syntaxique à partir de la racine de l'arbre. L'étiquette du constituant peut ensuite être interprétée comme frontière de proposition, de syntagme ou de mot à partir des catégories de constituants de PennTree Bank, *cf.* annexe D. L'annotation automatique des pauses par PLSPP s'arrête ici. À partir de là, l'utilisateur peut définir un seuil à partir duquel considérer les intervalles comme pauses, et faire des mesures à partir de leurs informations contextuelles.

5.3.2 Mesures de fréquences des pauses

L'alignement effectué par Wav2Vec2.0 présente la spécificité de placer un intervalle « vide » entre chaque mot, même lorsque celui-ci est loin d'être perceptible par l'auditeur. En considérant chacun de ces intervalles comme « pause potentielle », nous pouvons ainsi catégoriser l'ensemble des frontières de mots en fonction du type de frontière syntaxique : inter-proposition, inter-syntagme, ou intra-syntagme, si aucune frontière de constituant n'est présente. La figure 5.4 illustre cet étiquetage de l'ensemble des intervalles inter-mots, dans un segment du corpus CLES-JP. On peut ensuite filtrer ces intervalles en fonction de leur durée pour ne conserver que celles à considérer comme pauses, selon le seuil de durée choisi.

Dans cette étude, nous considérons un seuil de durée minimum fixe de 180 ms pour prendre en compte les pauses brèves tout en évitant les phénomènes de coarticulation (Heldner & Edlund, 2010). Pour permettre une meilleure comparabilité de nos résultats avec les études précédentes, nous considérerons également le seuil de 250 ms, plus commun dans le domaine de l'évaluation de la fluence en L2. Un seuil de durée maximum de 2 s est également paramétré de manière à ignorer les pauses très longues, pouvant résulter d'erreurs d'alignement.

Nous avons vu différentes mesures de fréquence des pauses dans la revue de la littérature (*cf.* chapitre 3.2.3) : le nombre de pauses par minute, par mot, par syllabe, ou encore par tour de parole. La fréquence des pauses par unité de temps est influencée par le débit de parole : plus le locuteur parle vite, plus le nombre de mots par

[11] https://plspp.univ-grenoble-alpes.fr/pausesviz/CLESJP?req=doshisha2024_001_JNS_03A-03B_A_12

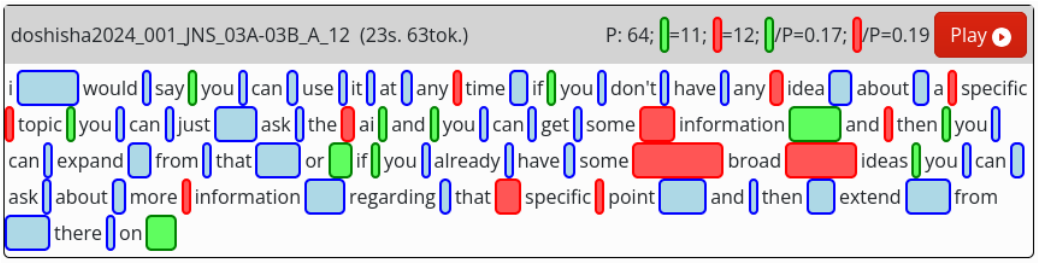

FIG. 5.4 : Transcription d'un segment audio du corpus CLES-JP, affichant chaque intervalle vide générée par l'aligneur automatique. La couleur des intervalles correspond au type de frontière syntaxique identifié (vert pour inter-proprosition, bleu pour inter-syntagme et rouge pour intra-syntagme). La longueur des intervalles représente leur durée. Notons que seuls les intervalles d'une durée supérieur au seuil défini par l'utilisateur sont considérés comme des pauses par la suite (cliquer ici[11] pour accéder à la visualisation en ligne)

minute augmente et par conséquent le nombre de pauses éventuelles. Pour neutraliser l'influence du débit de parole, nous avons choisi de calculer la fréquence des pauses par mot, ou plus exactement par token issu de la phase de transcription et d'alignement (globalement équivalent aux mots du dictionnaire). Nous calculerons d'abord la fréquence des pauses $F_{p,i}$ par type de frontière syntaxique i (inter-proposition (BC), inter-syntagme (BP) et intra-syntagme (WP)) :

$$F_{p,i\in\{BC,BP,WP\}} = \frac{N_{p,i}}{N_i} \tag{5.4}$$

avec $N_{p,i}$ le nombre de pauses p de catégorie i, et N_i le nombre de frontières syntaxiques de catégorie i. La valeur obtenue indique à quelle fréquence deux propositions sont séparées par une pause chez un locuteur donné. Nous compléterons cette mesure par la proportion de pauses $P_{p,i}$ de chaque catégorie i :

$$P_{p,i\in\{BC,BP,WP\}} = \frac{N_{p,i}}{N_p} \tag{5.5}$$

avec N_p le nombre total de pauses, toutes catégories confondues. Cette valeur indique par exemple la proportion de pauses intra-syntagmes chez un locuteur, quel que soit le nombre de pauses produites.

Comparer les groupes de locuteurs revient à comparer les scores obtenus par locuteur. Étant donné le nombre parfois limité de locuteurs (notamment pour le corpus japonophone et anglophone) et la non-normalité des distributions, la comparaison se fera au moyen du test de rangs non-paramétrique Wilcoxon-Mann-Whitney (Bauer, 1972). Ce test se concentre sur la différence de tendance générale entre deux distributions, mais a pour avantage d'être robuste à la taille et au type de distribution des

données. Nous nous basons sur ce test pour vérifier la significativité de la différence entre les distributions. Nous indiquerons également la tendance centrale de chaque distribution en indiquant leur valeur médiane, ainsi que la taille d'effet pour quantifier le degré de différence entre elles. Pour cela, nous proposons de calculer le delta de Cliff (Cliff, 1993), qui indique à quel point les valeurs d'une distribution A sont supérieures ou inférieures à celles de la distribution B. Le delta obtenu varie entre -1 et 1, 0 indiquant que les deux distributions sont identiques, 1 indiquant que toutes les valeurs de la première distribution sont supérieures à celles de la deuxième. Nous utiliserons les seuils d'interprétation de Romano et al. (2006) : la différence est grande à partir de $0{,}474$, moyenne à partir de $0{,}33$, et petite à partir de $0{,}147$; inférieure à cette valeur, la différence est négligeable. Nous indiquerons également l'intervalle de confiance à 95 %.

5.3.3 Score de distribution syntaxique des pauses

Nous proposons de calculer un score de distribution syntaxique des pauses (DSP) qui représente en une seule valeur le niveau syntaxique auquel les pauses ont tendance à survenir chez un locuteur, et ce indépendamment du nombre de pauses produites. Le calcul est effectué comme suit : pour un échantillon de parole donné, on compte le nombre de pauses N_p de chaque catégorie (inter-proposition (BC), inter-syntagme (BP) et intra-syntagme (WP)) que l'on pondère par une constante w de manière à favoriser les pauses de haut niveau et pénaliser celles de bas niveau. Enfin, on normalise par le nombre total de pauses N_p. Ce calcul peut se noter de la façon suivante :

$$DSP = \sum_{i \in BC,BP,WP} (p_i \cdot w_i) = \frac{N_{p,BC} \cdot w_{BC} + N_{p,BP} \cdot w_{BP} + N_{p,WP} \cdot w_{WP}}{N_p} \tag{5.6}$$

Nous proposons de fixer arbitrairement les poids w_{BC} à 1, w_{BP} à 0,5 et w_{WP} à -1, de manière à faire varier le score entre -1 et 1. La présence de pauses inter-proposition et inter-syntagme participe ainsi à élever le score, avec plus de poids pour les premières, tandis que les pauses intra-syntagme tireront le score vers le bas. Plus le score est haut, plus les pauses ont tendance à être placées en frontières de haut niveau. Un score négatif indique que la majorité des pauses est placée intra-syntagme.

5.3.4 Amélioration de l'approche

Selon nous, considérer les pauses en fonction de leur position vis-à-vis des propositions ou des syntagmes présente deux limitations importantes. La première est le

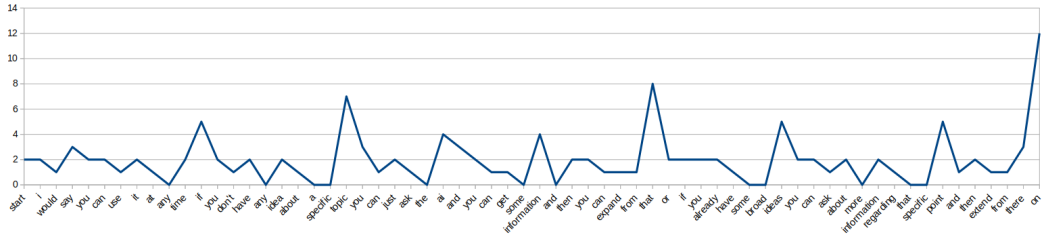

FIG. 5.5 : Nombre de constituants se fermant ou s'ouvrant après chaque mot

fait que le nombre de frontières intra-syntagmes est assez limité en anglais, et réduit ainsi la probabilité d'y trouver une pause. Dans l'exemple précédent, figure 5.4, on ne trouve ainsi que 12 frontières intra-syntagme (en rouge), contre 41 frontières inter-syntagmes (en bleu). La deuxième limitation est le fait que toutes les frontières qui ne sont ni inter-proposition ni intra-syntagme sont considérées au même niveau "inter-syntagme", alors qu'il y a en réalité toute une hiérarchie de syntagmes imbriqués les uns dans les autres – ce qui est, par ailleurs, également le cas pour les propositions. Nous proposons donc une approche complémentaire pour contourner ces limitations.

Au lieu de catégoriser le niveau des frontières syntaxiques en fonction de leur catégorie, nous proposons de les considérer de manière continue. Plus le nombre de constituants qui s'ouvrent ou se ferment est élevé à un endroit donné, plus la frontière syntaxique est considérée comme importante. Ainsi, la fermeture de trois syntagmes imbriqués les uns aux autres donnera une frontière plus importante que la fermeture d'un seul syntagme. On peut alors calculer une valeur représentant ce niveau de frontière – nous avons ainsi considéré la somme des constituants qui se ferment ou qui s'ouvrent après chaque mot. La figure 5.5 montre le niveau de frontière pour le même segment que la figure précédente. On y distingue des pics qui correspondent aux principales frontières syntaxiques de l'énoncé. Plus la valeur est haute, plus la frontière est importante, et une pause y survenant est susceptible d'aider à la compréhension.

On peut calculer le score de distribution syntaxique en remplaçant les catégories de pauses par des seuils d'importance de frontière. Nous proposons de définir arbitrairement trois seuils de la manière suivante : *high* pour une frontière d'importance 4 ou plus, *medium* pour 2 ou 3, *low* pour 0 ou 1. Le calcul est le même que précédemment : on fait une somme pondérée de la fréquence de pauses par niveau, avec les mêmes poids w_{high}, w_{medium} et w_{low} de 1, 0,5 et -1.

En résumé, nous proposons d'effectuer les mesures suivantes pour chaque locuteur :

- **Analyse globale**
  - F_p : fréquence des pauses (nombre de pauses par token)
  - $\overline{d_p}$: durée moyenne des pauses

- **Analyse structurelle**
  - $F_{p,i}$: fréquence des pauses par catégorie de frontière syntaxique
  - $P_{p,i}$: proportion des pauses par catégorie
  - DSP_i : score de distribution syntaxique des pauses basé sur les propositions et les syntagmes
  - DSP_n : score de distribution syntaxique des pauses basé sur le niveau d'importance des frontières

5.3.5 Évaluation de l'étiquetage

Pour évaluer la précision de la détection et de l'annotation des pauses, nous proposons de les comparer aux annotations manuelles du corpus de Mareková et Beňuš (2024), conçu à l'université Constantine le Philosophe à Nitra (Slovaquie). Ce corpus est composé de 72 dialogues de 24 binômes d'étudiants de langue maternelle slovaque, lors d'un jeu de rôle où les locuteurs sont amenés à décrire un trajet sur une carte. Le corpus totalise 8 h18 min de parole spontanée conversationnelle, entièrement transcrites et annotées manuellement en pauses inter- et intra-propositionnelles. Nous proposons de comparer le nombre et la catégorie des pauses annotées avec les résultats de nos annotations automatiques. Plus concrètement, nous avons calculé les scores de précision d'annotation automatique des pauses inter- et intra-proposition.

5.4 Annotation de l'accent lexical

L'accent lexical joue un rôle important pour la segmentation du flux de parole et l'accès lexical (Cutler, 2015 ; Cutler & Jesse, 2021). La qualité de sa réalisation est corrélée avec les jugements de compréhensibilité des auditeurs, et ce pour les débutants comme pour les locuteurs de niveau avancé (Isaacs & Trofimovich, 2012 ; Saito et al., 2015).

L'accentuation lexicale en anglais est réalisée par une combinaison de facteurs prosodiques et segmentaux qui font varier la qualité de la syllabe. La syllabe accentuée est en général plus haute en f_0, en intensité, et de durée plus longue que les syllabes non-accentuées, qui ont tendance au contraire à être réduites (f_0 et intensité plus basses, durée plus courte). Au niveau segmental, la voyelle accentuée est pleine et parfois diphtonguée, tandis que la voyelle réduite est centralisée et tend vers schwa relâché /ə/. Par ailleurs, les mots lexicaux (noms, adjectifs, verbes, adverbes) ont tendance à être accentués, alors que les mots grammaticaux ont tendance à être réduits.

Pour estimer la position de l'accent lexical et le niveau de contraste accentuel entre les syllabes, nous proposons de mesurer « le poids » relatif de chaque syllabe en termes de f_0, d'intensité et de durée. Nous parlerons le plus souvent de degré de « proéminence syllabique » pour faire référence à ce poids des syllabes, plutôt que de parler d'accent, qui fait intervenir des aspects segmentaux et perceptifs en plus des trois dimensions prosodiques analysées ici.

5.4.1 Détection des noyaux syllabiques

La première étape de l'analyse consiste à identifier les noyaux syllabiques sur lesquels seront ensuite effectuées les mesures acoustiques. Le noyau d'une syllabe correspond à son maximum d'intensité, il est généralement porté par une voyelle en anglais, et peut être précédé d'un onset et suivi d'une coda, tous deux consonantiques. Nous avons envisagé une méthode acoustique et une méthode phonologique pour localiser ces noyaux syllabiques. La méthode acoustique consiste à se baser sur les maximum locaux d'intensité situés à l'intérieur des frontières de mots alignés par Wav2Vec. Une segmentation en syllabes est alors effectuée à partir de la position de ces pics d'intensité, et les mesures acoustiques sont par la suite réalisées localement au niveau de chaque pic. La méthode phonologique consiste quant à elle à utiliser un alignement forcé de chaque phonème du mot à partir d'un dictionnaire phonologique. Les mesures acoustiques sont alors réalisées au niveau des intervalles vocaliques.

Les deux principales différences entre ces approches concernent la représentation du noyau syllabique et le nombre de noyaux détectés. Avec l'approche acoustique, les noyaux sont représentés par des points correspondant aux maximums locaux d'intensité, tandis qu'il s'agit d'intervalles avec l'approche phonologique. L'approche acoustique n'est pas dépendante du mot cible, et peut générer un nombre variable de noyaux syllabiques, quel que soit le nombre de syllabes attendues. Dans l'approche phonologique, au contraire, le nombre de syllabes est déterminé par le mot cible, d'après un dictionnaire phonologique, et ne dépend pas de la prononciation du mot par le locuteur.

Détection acoustique des noyaux syllabiques `SyllableNucleiv3.praat` prend en entrée les fichiers audio et génère un fichier TextGrid avec chaque noyau syllabique détecté aligné au signal. Il prend en paramètre les mêmes options que le script original, notamment un band-pass de 300 Hz à 3300 Hz activé par défaut.

Détection phonologique des noyaux syllabiques Ce module ajouté à partir de la version 2 de PLSPP recourt au Montreal Forced Aligner (MFA, McAuliffe et al., 2017) pour aligner le texte brut transcrit par Whisper. MFA permet de réaliser un alignement phonémique en plus de l'alignement des mots. En contrepartie, le système est

moins robuste aux disfluences et aux écarts entre la transcription et le signal audio, et a tendance à produire des alignement incohérents avec des enregistrements de parole disfluente. Ce module semble donc moins adapté à la parole spontanée.

5.4.2 Mesures du degré de proéminence syllabique

Nous proposons dans un premier temps de nous concentrer sur l'accentuation des mots polysyllabiques lexicaux. À ce stade des traitements, nous disposons d'un alignement des mots et de leurs syllabes au signal de parole, ainsi que la catégorie grammaticale issue de l'analyse morphosyntaxique. Nous sommes donc en mesure d'identifier le patron accentuel attendu pour chaque mot du corpus, en recourant à un dictionnaire phonologique de référence. Nous choisissons d'utiliser le *Carnegie Mellon University Pronouncing Dictionary*[12].

Pour chaque mot polysyllabique lexical, nous proposons de mesurer le degré de proéminence syllabique à partir de la f_0, de l'intensité et de la durée de chaque syllabe. La syllabe qui obtient le score maximum sera considérée comme la syllabe accentuée, et les autres seront pour l'instant considérées comme non accentuées (modèle binaire). Chaque dimension prosodique sera par ailleurs normalisée par locuteur et représentée en centile. Ainsi, une f_0 de 50 centiles indiquera une valeur médiane pour le locuteur en question, comparable à la valeur 50 de n'importe quel autre locuteur. Plus la valeur tend vers 100, plus la f_0 est élevée. Cette méthode de normalisation permet de tenir compte de la distribution des mesures pour chaque locuteur, tout en permettant de comparer les valeurs entre elles (50 représente la valeur médiane pour tous les locuteurs sur toutes les dimensions). En contrepartie, il est nécessaire d'avoir suffisamment de mesures pour chaque locuteur, sans quoi des centiles différents peuvent renvoyer aux mêmes valeurs absolues.

Normalisation par locuteur Elle est effectuée de la même manière pour les trois dimensions acoustiques : chaque valeur absolue est convertie en centile pour le locuteur et la dimension en question. La valeur ainsi obtenue s'étend de 0 à 100, avec 50 indiquant la valeur médiane de la dimension donnée pour le locuteur, et 100 la valeur maximale.

Annotation au niveau syllabique Effectuée dans la première version de PLSPP par le script `stressAnalysis.py`. Celui-ci prend en entrée les fichiers TextGrid contenant la transcription alignée, l'analyse morphosyntaxique et les noyaux syllabiques acoustiques (pics d'intensité) ; les fichiers audio, et le dictionnaire de référence CMU Pronouncing Dictionary. Pour chaque pic d'intensité, la f_0 est mesurée à partir du

[12] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

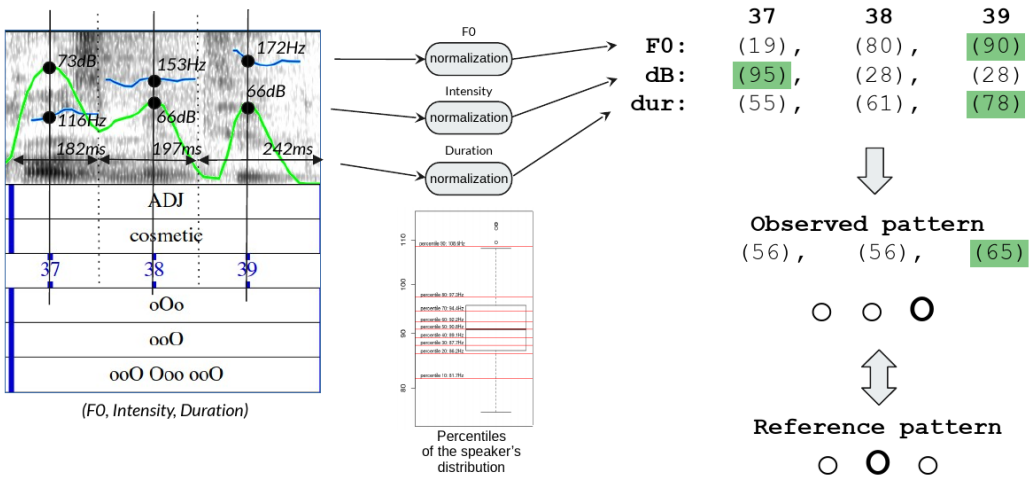

FIG. 5.6 : Extraction des paramètres prosodiques (PLSPP v1). À gauche un aperçu du fichier TextGrid de sortie avec les mesures acoustiques absolues indiquées en surimpression, la courbe bleue indique la f_0, la courbe verte indique l'intensité. À droite sont affichées les mesures normalisées. "Observed pattern" correspond au pattern accentuel observé (moyenne des trois dimensions prosodiques), "Reference pattern" correspond pattern attendu de l'accent primaire. La syllabe proéminente est marquée "O" et les autres syllabes sont marquées "o".

point le plus proche, ou bien par interpolation linéaire si aucune valeur n'est trouvée. La durée est quant à elle estimée à partir des noyaux voisins ou des frontières de mot. En sortie sont générés les fichiers TextGrid avec trois tiers supplémentaires : pour chaque mot cible, le pattern de référence, le pattern observé global consistant une moyenne des trois dimensions, et le pattern observé sur chacune des trois dimensions acoustiques (*cf.* figure 5.6).

Cette version est actuellement la plus robuste car elle s'appuie sur une combinaison de l'alignement au mot de Wav2Vec et de la détection acoustique des noyaux syllabiques. Toutefois, les mesures sont effectuées de manière ponctuelle au niveau des maximums d'intensité, et ne prennent donc pas en compte la variation de f_0 à travers la voyelle, et les mesures de durée sont plus facilement impactées par la structure syllabique et les allongements de consonnes, notamment les fricatives.

Annotation au niveau vocalique À partir de la deuxième version de PLSPP, les mesures acoustiques sont faites au niveau de l'intervalle vocalique de chaque syllabe. Le script `stressAnalysis_mfa.py` suit la même structure que son équivalent dans la version 1, à la différence qu'il boucle sur la tier des phonèmes plutôt que celle des noyaux syllabiques acoustiques. Pour chaque voyelle, les mesures de f_0 et d'intensité sont effectuées sur une fenêtre glissante de taille paramétrable (par défaut 10 ms, comme Ferrer et al., 2015) et les valeurs moyenne, minimum et maximum ainsi que

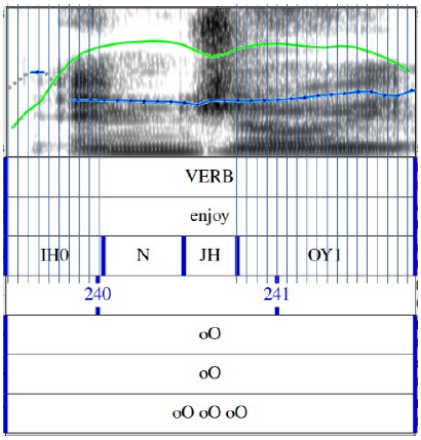

FIG. 5.7 : Extraction des paramètres prosodiques avec PLSPP v2. Les barres bleues ajoutées en surimpression représentent les frames de 10 ms pour le calcul de la f_0 (courbe bleue) et de l'intensité (courbe verte)

l'écart type sont enregistrées (*cf.* figure 5.7). La version 4 intègre également des mesures de qualité vocalique (F_1, F_2, et F_3) pour mesurer le degré de centralisation ou de diphtongaison des voyelles.

5.4.3 Comparaison des locuteurs

Deux scores sont ensuite calculés pour chaque locuteur : un score de position de l'accent, représentant le pourcentage de mots pour lesquels la syllabe proéminente correspond à la syllabe accentuée selon le dictionnaire de référence ; et un contraste prosodique moyen $\overline{C}$ calculé à partir de la différence entre la valeur prosodique de la syllabe censée être accentuée et la moyenne des autres syllabes, sur l'ensemble des mots produits par un locuteur. Le contraste prosodique pour un mot w peut être calculé comme suit :

$$C_w = P_{s,w} - \overline{P_{u,w}} \tag{5.7}$$

avec $P_{s,w}$ la valeur prosodique de la syllabe censée porter l'accent lexical, et $\overline{P_{u,w}}$ la moyenne des valeurs des autres syllabes du mot. Ce contraste pourra être calculé globalement (moyenne des trois dimensions prosodiques) ou par dimension. Il indique ainsi à quel point la syllabe accentuée se démarque acoustiquement des autres. La valeur obtenue varie entre -100 et +100, 0 indiquant qu'il n'y a pas de différence prosodique entre la syllabe accentuée et les autres syllabes, une valeur négative signalant que la proéminence se situe sur une autre syllabe que celle censée être accentuée.

En résumé, nous proposons d'effectuer les mesures suivantes pour chaque locuteur :

- N_{mots}, N_{poly} et N_{ann} : nombre de mots, nombre de mots polysyllabiques lexicaux, nombre de mots annotés
- S : score de position de l'accent
- C : contraste prosodique mesuré entre la syllabe accentuée et les autres syllabes du mot (C_{f_0}, C_{int}, C_{dur}, contrastes par dimension)
- $\overline{C}$: contraste prosodique moyen sur l'ensemble des mots annotés ($\overline{C_{f_0}}$, $\overline{C_{int}}$, $\overline{C_{dur}}$, contrastes moyens par dimension)

5.4.4 Évaluation de l'accentuation mesurée

Comment savoir si la syllabe proéminente identifiée par le système correspond effectivement à la syllabe accentuée perçue par l'auditeur ? Nous proposons trois approches différentes pour aborder cette question :

a) Demander à des auditeurs anglophones natifs de noter manuellement les syllabes qu'ils perçoivent accentuées dans des enregistrements de locuteurs non-natifs et comparer avec les annotations automatiques ;

b) Demander à des locuteurs natifs et non-natifs où doit être placé l'accent primaire sur une série de mots cibles, puis comparer leur conscience accentuelle avec leur production ;

c) Annoter automatiquement des enregistrements de parole plus ou moins contrôlée produite par des locuteurs natifs ;

a) Évaluation perceptive par des auditeurs natifs

La première approche a consisté à vérifier si les annotations automatiques d'accentuation de PLSPP sont cohérentes avec le jugement d'auditeurs natifs. En d'autres mots : est-ce que les auditeurs anglophones natifs perçoivent l'accent au même endroit que PLSPP ?

Cette étude a été menée par un étudiant du *Spoken Language Processing Laboratory* de l'université Dōshisha, et a fait l'objet d'une publication dans les actes du congrès bi-annuel de l'*Acoustical Society of Japan* (Kimura et al., 2024). Nous avons

recruté 10 évaluateurs anglophones natifs pour annoter manuellement six enregistrements de locuteurs japonophones. Les évaluateurs sont originaires des États-Unis et vivent dans la région de Tōkyō depuis plus de 5 ans au moment de l'expérimentation. Les enregistrements de parole ont été réalisés sur six élèves entre 9 et 11 ans d'une école primaire privée de la préfecture de Kyōto, enregistrés pendant une récitation de texte. Le texte est une description d'un personnage historique de 300 mots, commun à l'ensemble des locuteurs, et qui a fait l'objet d'un entraînement préalable. La transcription du texte avec les mots à évaluer mis en relief était fournie aux évaluateurs au format papier, en 6 exemplaires, et les évaluateurs devaient noter à la main la position de l'accent tel qu'ils le percevaient pour chaque enregistrement. Lorsqu'aucun accent n'était perçu, les participants pouvaient tracer une barre au-dessus du mot.

Pour un mot donné, si l'accent est noté sur une syllabe qui ne correspond pas à la syllabe censée porter l'accent primaire d'après le dictionnaire de référence utilisé par PLSPP, on compte une erreur. Après avoir calculé le taux de corrélation inter-annotateur, nous avons comparé le nombre d'erreurs relevées par évaluateur et par locuteur, et l'avons comparé au nombre d'erreurs identifiées automatiquement par PLSPP v2. Par ailleurs, un score de contraste similaire à celui présenté plus haut a été calculé pour chaque mot à partir des valeurs prosodiques mesurées par le système, puis comparé à un score d'accentuation humain issu de la moyenne des jugements des évaluateurs. Nous appellerons ce score C' pour le différencier de C (qui est une simple différence de centiles entre les syllabes). C' est calculé de la manière suivante :

$$C'_w = \frac{P_{s,w}}{P_{s,w} + \overline{P_{u,w}}} \tag{5.8}$$

où w est le mot courant, $P_{s,w}$ correspond à la valeur prosodique de la syllabe accentuée attendue (moyenne des centiles de f_0, d'intensité et de durée), et $\overline{P_{u,w}}$ la valeur prosodique moyenne des autres syllabes du mot. On obtient alors une valeur entre 0 et 1 ; 0,5 indiquant un contraste nul entre la syllabe accentuée et les autres syllabes, et 1 indiquant un contraste positif maximal.

b) Annotation automatique et conscience phonologique

La deuxième approche a consisté à comparer l'annotation automatique de PLSPP avec les patterns accentuels conscientisés par les locuteurs. En d'autres termes, nous avons cherché à savoir si PLSPP détecte une proéminence acoustique sur la syllabe que le locuteur pense accentuer. Nous avons également investigué l'influence des tendances d'accentuation de la langue maternelle du locuteur en comparant des locuteurs de différentes langues maternelles.

Cette étude a été coordonnée par Mariko Sugahara, enseignante-chercheuse au département d'anglais de l'université Dōshisha, qui travaille depuis plusieurs années sur la conscientisation de l'accent lexical en anglais par les apprenants japonophones et coréanophones. Comme nous l'avons vu dans le chapitre 3, l'anglais possède un accent lexical avec une certaine tendance à l'accentuation en initiale. Le japonais, dans sa variété la plus répandue *Tokyo/Keihan*, possède également un accent lexical, mais plutôt à tendance médiale. Quant au coréen de Séoul, à l'instar du français, il ne possède pas d'accent lexical, et les locuteurs coréanophones ont tendance à avoir plus de difficultés que les japonophones à maîtriser l'accentuation lexicale de l'anglais.

Une liste de 57 mots cibles en anglais a été enregistrée dans des phrases porteuses par 12 locuteurs anglophones natifs, 14 locuteurs japonophones et 11 locuteurs coréanophones (tous de niveau CECRL B1 à B2), puis annotée automatiquement avec PLSPP v2. En parallèle, ces mêmes locuteurs ont passé un test de conscience phonologique, lors duquel il leur était demandé d'indiquer sur une liste de mots la voyelle qui porte l'accent primaire selon eux. Les mots sélectionnés consistent en 19 triplets composés d'un verbe à 3 syllabes portant l'accent sur l'initiale (ex. *dominate*), sa forme en *-ing* (accent primaire sur l'initiale, ex. *dominating*), et son dérivé substantif en *-ion* (accent primaire sur la 3ème syllabe, ex. *domination*). Cette approche permet de vérifier si la syllabe proéminente identifiée automatiquement correspond à la syllabe considérée accentuée par les locuteurs, indépendamment d'une référence prescriptive externe.

Un taux de correspondance $Corr_{PLSPP-loc}$ entre l'annotation automatique de PLSPP et l'accent théorique selon le locuteur a été calculé pour chaque groupe de locuteurs et chaque item du triplet. Une observation des mesures acoustiques de chaque syllabe a également permis d'étudier le poids donné à l'accent secondaire vis-à-vis de l'accent primaire.

c) Annotation de parole produite par des locuteurs natifs

La troisième approche a consisté à considérer la production des locuteurs anglophones natifs comme référence en termes d'accentuation lexicale, en établissant le postulat selon lequel ces derniers accentuent systématiquement la syllabe censée porter l'accent primaire. Le taux d'erreur d'accentuation rapporté par PLSPP correspondrait ainsi directement au taux d'erreur d'annotation.

Nous avons comparé les scores de position de l'accent et le contraste acoustique moyen $\overline{C}$ obtenus par locuteur dans des enregistrements de parole contrôlée de différents types. Deux corpus ont été utilisés ici : le premier est un corpus de phrases porteuses lues par 17 locuteurs natifs, dont une partie a été utilisée dans l'étude décrite plus haut, et le second est constitué de 92 textes lus en studio par sept locuteurs

natifs professionnels, dans le cadre de l'enregistrement de manuels scolaires d'anglais (Nakanishi et al., 2023a, 2023b, 2024a, 2024b).

Si cette approche est limitée par le postulat de départ, elle permet néanmoins de quantifier le degré d'accentuation purement acoustique (en termes de contraste de f_0, d'intensité et de durée des syllabes) produit par les locuteurs natifs en situation de parole contrôlée. En effet, comme nous l'avons présenté dans le chapitre 3, ces trois paramètres prosodiques ne sont pas seuls en jeu dans le processus d'accentuation lexicale. Celle-ci est également influencée par des paramètres de qualité vocalique, mais aussi par le contexte lexical et la nécessité plus ou moins grande de désambiguïsation. En outre, il va de soi que les mesures effectuées sur de la parole spontanée seront considérablement moins précises qu'en parole contrôlée, mais cela fera l'objet du chapitre 8.

5.5 Récapitulatif des versions de PLSPP

PLSPP est un outil d'annotation automatique des pauses et de l'accent lexical développé de manière modulaire pour permettre de s'adapter facilement à différents types de données. Si l'annotation des trois corpus CLES de parole spontanée, qui constitue le cœur de notre travail de recherche, a été réalisée avec la première version de PLSPP (fond bleu sur la figure 5.8), d'autres versions ont par la suite été développées dans le cadre d'études parallèles, mais se sont révélées moins robustes pour l'analyse de la parole spontanée.

PLSPP se décline aujourd'hui en quatre versions utilisées selon les besoins et le type de parole analysée :

- **PLSPP v1** est à ce jour la version la plus adaptée pour analyser la parole spontanée. Elle se base sur une identification acoustique des noyaux syllabiques et a été utilisée pour analyser les corpus du CLES (Coulange & Kato, 2023 ; Coulange et al., 2023, 2024a), ainsi que d'autres corpus de parole spontanée comme celui des locuteurs slovaquophones de l'université de Nitra, de locuteurs sinophones d'un corpus de l'université de Gröningen et de locuteurs hispanophones de l'université de Barcelone n'ayant pas encore fait l'objet de publications ;
- **PLSPP v2** se base sur une identification phonologique des noyaux syllabiques, les annotations de l'accent sont plus précises car limitées aux intervalles vocaliques, mais l'alignement est moins robustes aux disfluences de la parole. Cette version est donc moins adaptée à la parole spontanée. Elle a été utilisée pour

l'analyse de phrases porteuses, de textes lus ou récités par des locuteurs japonophones, coréanophones et anglophones natifs (Kimura et al., 2024 ; Sugahara et al., 2023, 2024) ;

- **PLSPP v3** est une évolution de v2 permettant l'analyse des mots monosyllabiques. Elle permet entre autres de mesurer le contraste accentuel entre les mots lexicaux et grammaticaux, et a été utilisée sur des textes lus par des locuteurs japonophones et anglophones natifs (Nakanishi & Coulange, 2024) ;
- **PLSPP v4**, enfin, intègre des mesures de qualité vocalique pour analyser le degré de réduction et de diphtongaison des voyelles, et permet de croiser les mesures acoustiques avec des mesures physiologiques obtenues par des capteurs complémentaires. Cette version a été utilisée sur de la parole de locuteurs lusophones (Brésil) et anglophones natifs, en combinaison avec des mesures d'aperture de mâchoire réalisées avec un articulographe électromagnétique (**ErickssonAl2025** ; Raso et al., 2024).

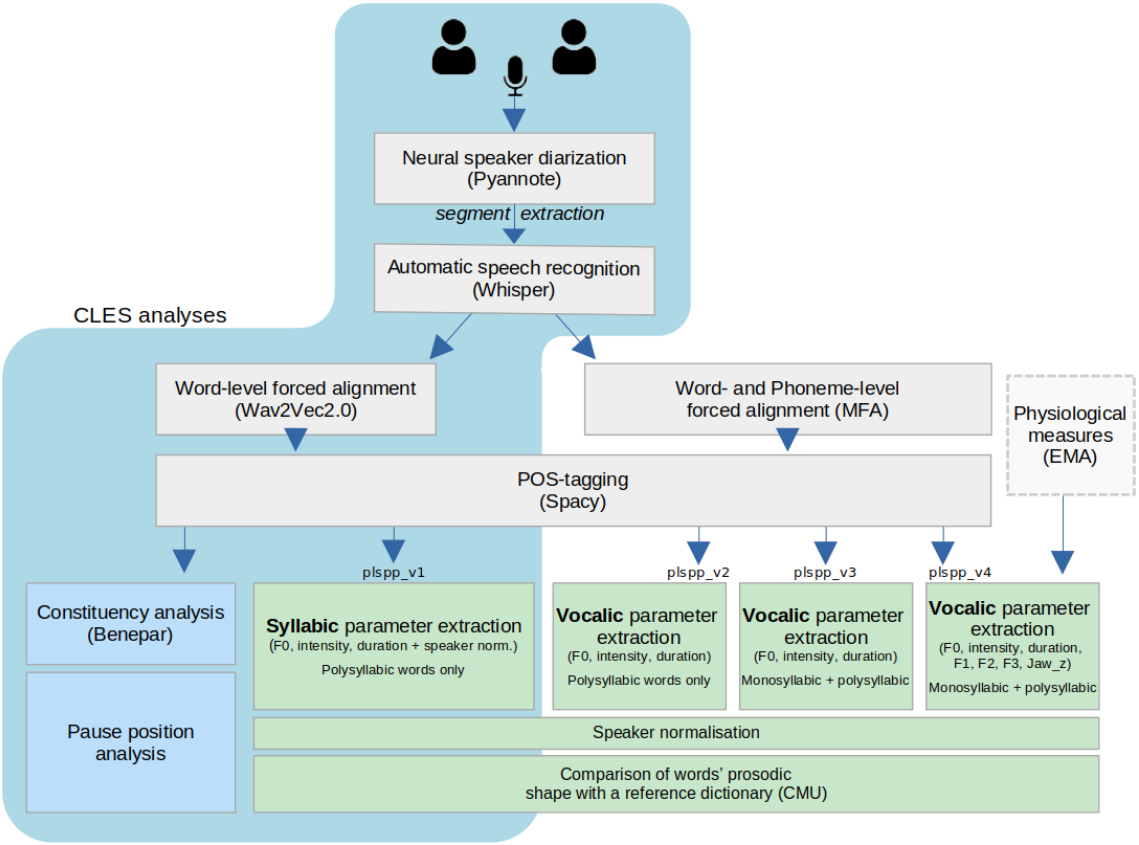

FIG. 5.8 : Architecture détaillée de PLSPP et ses différentes versions

5.6 Interface de visualisation des annotations

Pour simplifier le parcours et la lecture des annotations de PLSPP, une interface web de visualisation a été développée en parallèle de la création de la pipeline. L'interface est hébergée sur un serveur de l'université Grenoble Alpes au moment du dépôt de cette thèse, mais son code reste disponible et téléchargeable depuis gricad-gitlab.univ-grenoble-alpes.fr[13].

L'interface a trois objectifs principaux :

- Afficher un résumé interactif des patterns accentuels en fonction de paramètres de filtrage des locuteurs et des mots cibles, tout en permettant d'écouter facilement les mots recherchés ;
- Écouter et afficher la transcription des segments de parole annotés avec le détails des annotations de patterns accentuels ;
- Visualiser les pauses dans leur contexte syntaxique et permettre de moduler l'affichage grâce à différents filtres et paramétrages de seuils de durée notamment.

L'interface de visualisation se compose de trois pages principales : une vue globale des annotations de l'accent lexical, une page de visualisation des segments de parole avec les annotations d'accentuation, et une page de visualisation des segments avec les annotations de pauses.

6.1 Vue globale

Dans la vue globale (*cf.* figure 5.9), l'utilisateur peut directement charger des fichiers de sortie de PLSPP depuis son ordinateur, ou bien sélectionner un corpus déjà annoté dans l'onglet *Dataset*. N.B.: les corpus qui ne sont pas publics ne sont accessibles que par les utilisateurs disposant des droits appropriés.

Les options de filtrage des locuteurs (1) sont générées automatiquement à partir des colonnes disponibles dans le fichier *speaker.csv* généré par PLSPP. Par défaut, celui-ci ne contient que la liste des locuteurs identifiés dans le corpus. L'utilisateur peut y ajouter des informations de profil comme la langue maternelle ou le niveau d'anglais, de manière à filtrer les données sur la base de ces critères. Les résultats de la page sont mis à jour automatiquement à chaque modification de filtre.

[13] https://gricad-gitlab.univ-grenoble-alpes.fr/lidilem/plsppviz

Les options de filtrage des mots cibles (2) permettent de rechercher des patterns réguliers de mots (à partir d'une expression régulière), de filtrer par catégorie grammaticale et par gabarit accentuel théorique (position de l'accent attendue). Ces filtres sont générés automatiquement à partir du fichier de données généré par PLSPP *(stressTable.csv)*.

Plusieurs visualisations sont générées : une distribution des mots annotés en fonction de leur nombre de syllabes (3), la proportion de mots dont la position de l'accent est reconnue correcte (4), la distribution des patterns accentuels observés pour chaque gabarit théorique (5), et le degré de contraste syllabique moyen et par dimension prosodique (6).

Enfin, un tableau (7) listant les 299 premiers mots résultant du filtrage, avec leur locuteur, leurs gabarits théorique et observé, et une visualisation du poids de chaque syllabe sur chacune des dimensions f_0, intensité et durée. En cliquant sur le mot, l'utilisateur peut écouter le segment de parole associé, en cliquant sur le locuteur, il peut écouter le mot dans un contexte de 4 secondes.

6.2 Visualisations par segment de parole

La page *Stress patterns* (*cf.* figure 5.10), permet d'afficher la liste des segments de parole d'un locuteur donné, avec la transcription orthographique et les annotations accentuelles fournies par PLSPP. L'utilisateur peut écouter un mot (annoté ou non) en cliquant dessus, ou écouter le segment entier en cliquant sur le bouton *Play* en haut à droite du segment.

La page *Pause patterns* (*cf.* figure 5.11) est similaire à la précédente, à la différence que plusieurs paramètres sont personnalisables : le seuil de durée minimum et maximum des pauses, le nombre minimum de mots par segment, et un certain nombre de filtres temporaires pour afficher un top 15 des segments du corpus avec des caractéristiques spécifiques (maximum de mots, de pauses, de pauses inter-propositionnelles ou intra-syntagmes).

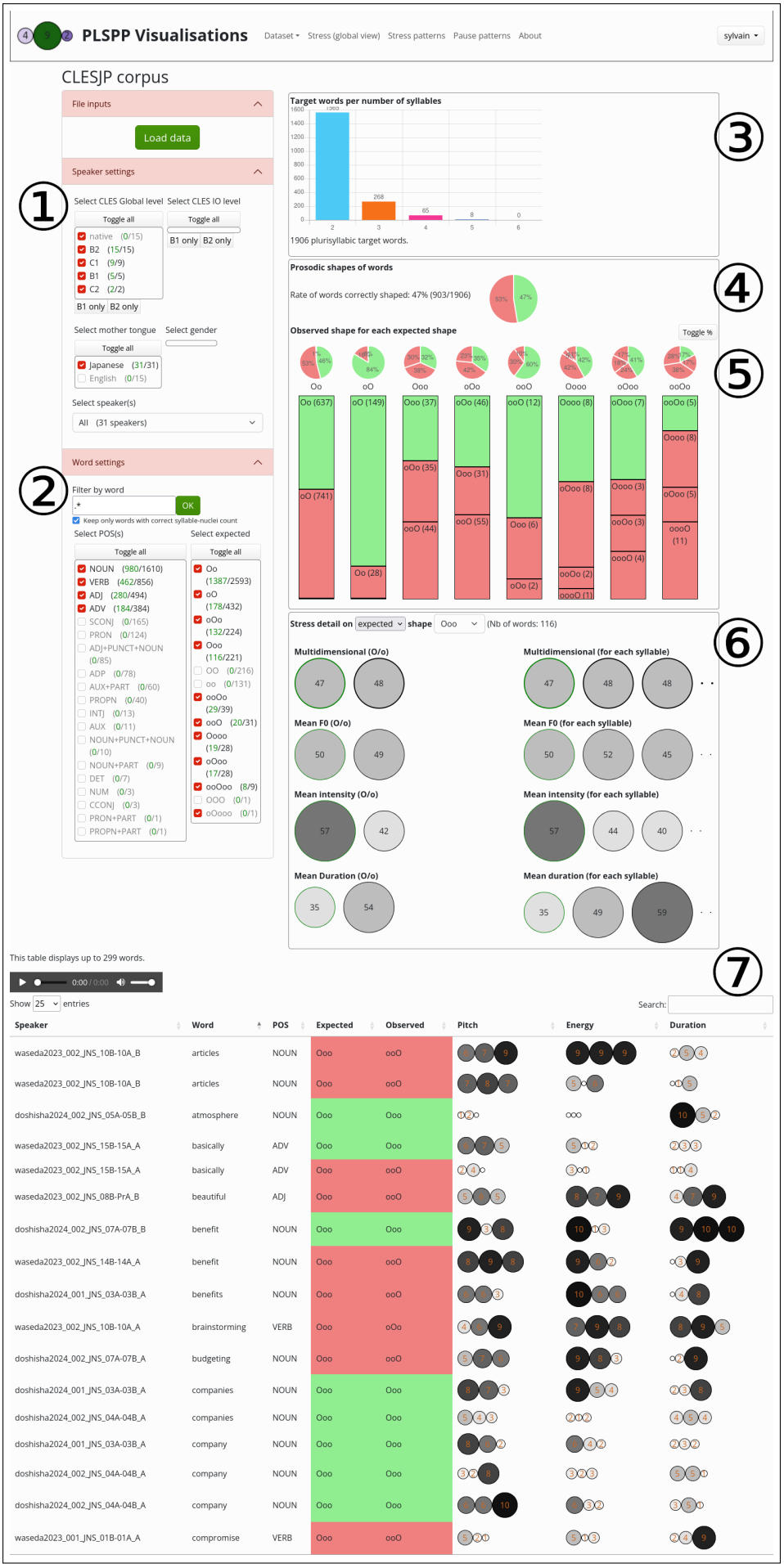

Fig. 5.9 : Aperçu de l'interface de visualisation de PLSPP, vue globale

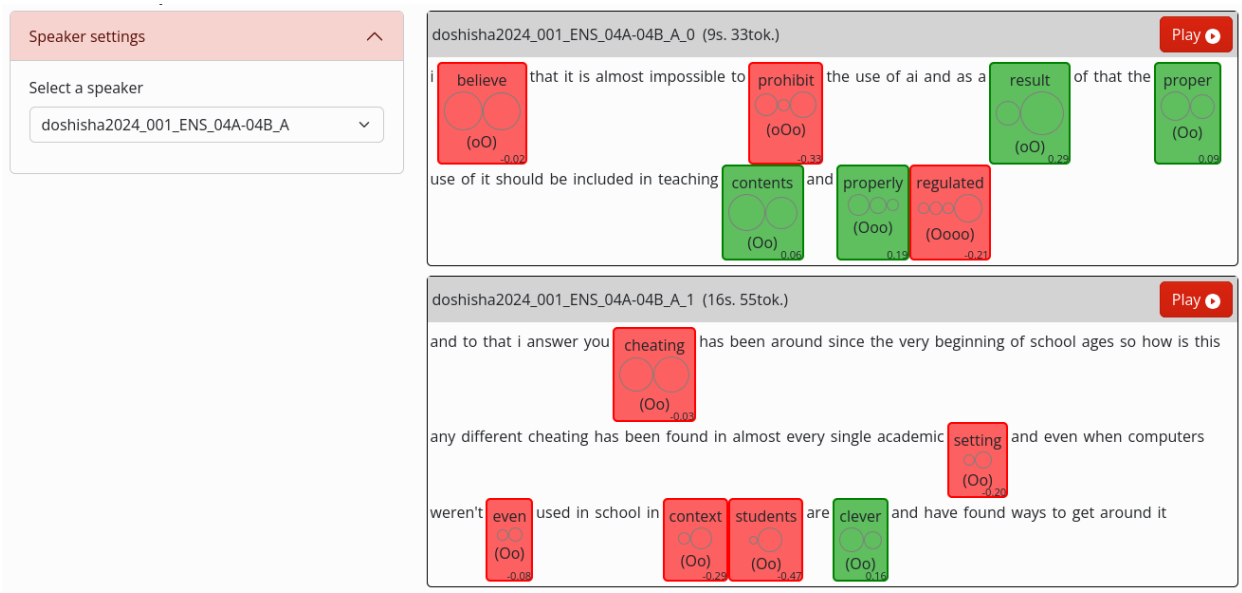

FIG. 5.10 : Aperçu de l'interface de visualitation de PLSPP, page Stress patterns

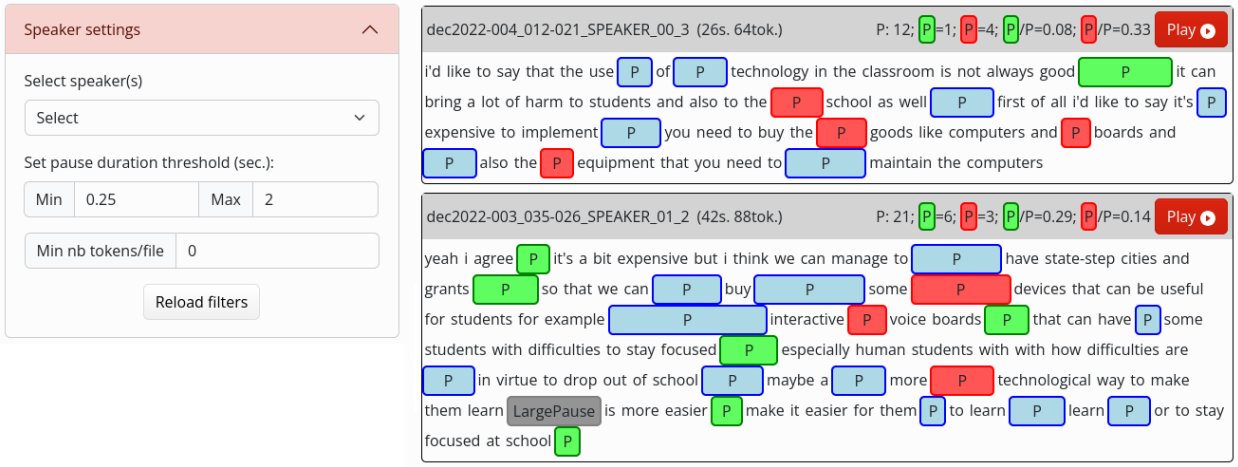

FIG. 5.11 : Aperçu de l'interface de visualitation de PLSPP, page Pause patterns

Conclusion

Dans ce chapitre, nous avons présenté les différents modules de traitement pour annoter les pauses et les proéminences syllabiques dans nos corpus de parole spontanée CLES. Chaque module est en charge d'un type de traitement spécifique. Les trois premiers permettent d'annoter les conversations et extraire des segments de parole par locuteur, de les transcrire et d'analyser leur syntaxe, et d'aligner temporellement l'ensemble de ces annotations. Les deux derniers modules génèrent ensuite une annotation des pauses et de l'accentuation lexicale.

L'évaluation de l'outil se fera module par module, à l'aide de différents corpus de référence. La liste ci-dessous récapitule les métriques d'évaluation et les données de référence utilisées pour chaque module.

1. **Segmentation en locuteurs**
    - Métriques :
        - DER, taux d'erreur de diarisation
        - I_L, indice d'interférence du locuteur
    - Données : corpus CLES-gold
2. **Reconnaissance et alignement de la parole**
    - Métriques :
        - WER, taux d'erreur de mots
        - SR, DR, IR, taux de substitution, de délétion et d'insertion
        - P, R, précision et rappel de l'alignement mot-signal
    - Données : corpus CLES-gold, corpus Frost et al. (2024)
3. **Analyses syntaxiques**
    - Analyse grammaticale par constituants évaluée avec l'annotation des pauses
4. **Annotation des pauses**
    - Métriques :
        - P, R, précision et rappel de détection des pauses inter- et intra-proposition
    - Données : corpus de Mareková et Beňuš (2024)

5. **Annotation de l'accent lexical**

    - Métriques :
        - Comparaison des scores de position de l'accent entre PLSPP et 10 évaluateurs humains
        - Comparaison des scores de contraste prosodique de PLSPP avec la moyenne des évaluations humaines
        - $Corr_{PLSPP-loc}$, correspondance entre la position de l'accent identifiée par PLSPP et le jugement de position théorique par le locuteur
        - S_{L1}, scores de position de l'accent chez des locuteurs natifs en parole lue
        - $\overline{C_{L1}}$, $\overline{C_{f_0,L1}}$, $\overline{C_{int,L1}}$, $\overline{C_{dur,L1}}$, contrastes prosodiques moyens globaux et par dimension observés chez des locuteurs natifs en parole lue
    - Données : corpus de Kimura et al. (2024), Sugahara et al. (2024), et Nakanishi et Coulange (2024)

Une fois l'outil d'annotation évalué, nous l'emploierons pour analyser les patterns de pauses et d'accentuation lexicale en parole spontanée entre les locuteurs de niveau CECRL B1 et B2, à travers les trois corpus CLES présentés dans le chapitre précédent. La liste suivante récapitule l'ensemble des mesures effectuées à partir des annotations automatiques pour comparer les productions des différents groupes de locuteurs.

- **Analyses de la fluence**
    - F_p : fréquence des pauses (nombre de pauses par token)
    - $\overline{d_p}$: durée moyenne des pauses
    - $F_{p,i}$: fréquence des pauses par catégorie de frontière syntaxique
    - $P_{p,i}$: proportion des pauses par catégorie de frontière syntaxique
    - DSP_i : score de distribution syntaxique des pauses basé sur les propositions et les syntagmes
    - DSP_n : score de distribution syntaxique des pauses basé sur le niveau de profondeur des frontières
- **Analyses du rythme**
    - N_{mots}, N_{poly} et N_{ann} : nombre de mots, nombre de mots polysyllabiques lexicaux, nombre de mots annotés

- S : score de position de l'accent
- C : contraste prosodique mesuré entre la syllabe accentuée et les autres syllabes du mot (C_{f_0}, C_{int}, C_{dur}, contrastes par dimension)
- $\overline{C}$: contraste prosodique moyen sur l'ensemble des mots annotés ($\overline{C_{f_0}}$, $\overline{C_{int}}$, $\overline{C_{dur}}$, contrastes moyens par dimension)