THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : LLSH- Langues, Littératures et Sciences Humaines
Spécialité : Sciences du langage Spécialité Informatique et sciences du langage
Unité de recherche : Laboratoire de Linguistique et Didactique des Langues Etrangères et Maternelles

Évaluation automatique de la parole spontanée en anglais langue étrangère: le rôle des pauses et de l'accent lexical dans la compréhensibilité du locuteur

Automated Assessment of Spontaneous Speech in English as a Foreign Language: the Role of Pauses and Lexical Stress in Speaker Comprehensibility

Présentée par :

Sylvain COULANGE

Direction de thèse :

Monica MASPERI PROFESSEURE, Université Grenoble Alpes	Directrice de thèse
Solange ROSSATO MAITRE DE CONFERENCES, Université Grenoble Alpes	Co-encadrante de thèse
Tsuneo KATO PROFESSEUR, Doshisha University	Co-encadrant de thèse

Rapporteurs :

Nicolas Ballier
PROFESSEUR, Université Paris-Cité
Nadine Herry Benit
PROFESSEURE, Université Paris Nanterre

Thèse soutenue publiquement le **10 avril 2025**, devant le jury composé de :

Monica MASPERI, PROFESSEURE, Université Grenoble Alpes	Directrice de thèse
Nicolas Ballier, PROFESSEUR, Université Paris-Cité	Rapporteur
Nadine Herry Benit, PROFESSEURE, Université Paris Nanterre	Rapporteure
Isabelle Darcy, PROFESSEURE, Indiana University	Examinatrice
Philippe Boula de Mareüil, DIRECTEUR DE RECHERCHE, LIMSI	Examinateur
François Pellegrino, DIRECTEUR DE RECHERCHE, Université Lumière Lyon 2	Examinateur
Antonio Romano, PROFESSEUR, Torino University	Examinateur

Remerciements

J'aimerais avant tout remercier les personnes qui ont dirigé ce travail de recherche. Monica, merci de m'avoir permis de réaliser ce projet de doctorat, qui est toutefois bien peu de chose par rapport à tout ce que m'ont apporté ces 10 ans au sein du projet Innovalangues. Et oui, 10 ans déjà. Merci pour ces 10 années, et d'avance pour les suivantes, qui seront, je l'espère, nombreuses encore. Solange, au-delà de la supervision de cette thèse, c'est davantage pour ma supervision scientifique depuis mon master de français langue étrangère en 2016, suivi par celui d'industries de la langue en 2019, que je souhaite te remercier. Je te dois mon approche de la science en général, ainsi que la plupart de mes choix de parcours de ces neuf dernières années. Katō sensei, thank you for accepting to supervise my work at a time when we didn't know each other, nor did our laboratories or universities. Thank you for inviting me to spend a year at your laboratory at Dōshisha University. I could not have expected that this year would have such a significant impact on my scientific life, as the list of collaborative studies in Annex H demonstrates. This is all thanks to you.

Mes remerciements se dirigent ensuite vers mon comité de suivi individuel de thèse. Merci à Elisabetta Carpitelli d'avoir accepté d'y participer. Tu auras décidément eu un lien particulier avec toute ma fratrie, puisque tu as été également très proche de mes deux sœurs, à des périodes et pour des raisons différentes. Mais il doit y avoir quelque chose de commun dans tout cela finalement. Merci également à Takaaki Shōchi pour tes nombreux conseils et ton approche franco-japonaise du monde académique.

Merci à Nadine Herry Benit et Nicolas Ballier d'avoir accepté d'évaluer mon travail. Merci également à l'ensemble de mon jury. J'espère que cette thèse aura su susciter votre intérêt.

En général, cette partie des remerciements arrive à la fin, mais je tiens à la placer ici. Merci à ma femme de m'avoir accompagné tous les jours de ce doctorat (et de m'avoir supporté). En ta compagnie, toutes ces années de doutes sont devenues un plaisir que je ne demande qu'à poursuivre.

Some researchers have had a profound influence on this work, and I would like to dedicate a few words to them as well. Pavel Trofimovich, I will keep deep inside my heart this moment we spent together under the shade of a tree in the Chartreuse mountains. What you said to me left a lasting impact on me and my PhD journey, which was just starting at the time. Beyond this, your writings have guided my entire thesis, and I sincerely hope we will have the opportunity to collaborate in the future. John Levis, my next words are for you, as you also had a significant impact on my work. Thank you for your early

comments on my research in 2022 and for everything you have written over the past two decades. I aspire to write papers or books similar to yours one day. I would also like to express my gratitude to Talia Isaacs and Nivja de Jong for your precious comments and suggestions. I feel very lucky to have had the opportunity to speak with you during my PhD journey.

Je tiens également à remercier Marie-Hélène Fries et le bureau de Direction CLES pour leur confiance, ainsi que les examinateurs CLES des campus de Grenoble et Valence de m'avoir ouvert leurs portes d'examen pour que j'y pose mes nombreux micro. Et bien sûr, merci aux 304 étudiants qui ont participé aux enregistrements sur les campus de Grenoble et Valence, ainsi que dans les universités Waseda à Tōkyō et Dōshisha à Kyōto.

Thank you to all my collaborators for believing in my work and for agreeing to collaborate or share your data with me : Mariko Sugahara, who was the first to use PLSPP after me ; Noriko Nakanishi, for our numerous collaborations ; Takayuki Konishi, for initiating our data collection in Japan ; Nobuaki Minematsu, for your invaluable suggestions and comments ; Dan Frost, for your early feedback on PLSPP ; and Donna Erickson, Lucia Mareková, and May Wu, for generously sharing your data with me. I would also like to express my gratitude to my younger collaborators, Tatsuya Kimura, Manato Nishioka, and Nathanaël Berthet.

Merci à Gérard Bailly de m'avoir mis sur la voie de l'évaluation dynamique de la compréhensibilité, et de m'avoir facilité l'utilisation de la plateforme Prolific. Merci aux 60 participants de l'expérience d'évaluation dynamique réalisée dans cette thèse.

Ils arrivent à la fin, mais sans eux non plus cette thèse n'aurait pas été possible : Roslyn Young, Éric Lepoint et Dan Frost (à nouveau), merci pour vos nombreux éclairages sur la prononciation de l'anglais et sur l'importance de la prosodie. Merci à Chris Mitchell, Marieke De Koning et Alex Carr pour leurs précieux commentaires sur l'enseignement et l'évaluation de la prononciation de l'anglais. Et enfin merci à tous les doctorants et jeunes chercheurs que j'ai côtoyés pendant ces trois années (et plus pour certains), avec un sentiment particulier pour Triscia Biagiotti, Solène Évain et Kōsuke Hinai.

Enfin, merci à ma famille pour son soutien.

Sommaire

Introduction

Dans une étude menée auprès de 459 enseignants d'anglais dans sept pays européens, Henderson et al. (2012) ont constaté que l'enseignement de la prononciation en langue étrangère (L2) est souvent négligé, tant en classe que dans la formation des formateurs. Cette situation engendre de grandes disparités dans les méthodes d'évaluation utilisées, et les enseignants disposent rarement des outils et de la formation nécessaires pour évaluer la prononciation de manière précise et systématique. Ils en viennent souvent à concevoir des grilles d'évaluation « maison », dont le manque de références communes entraîne des incohérences dans les notes obtenues par les apprenants (Frost & O'Donnell, 2018). Par ailleurs, Gilquin et al. (2022) observent que les critères d'évaluation diffèrent selon les évaluateurs, et notamment selon qu'ils sont locuteurs natifs ou non de la langue évaluée. Les auteurs constatent que les évaluateurs non natifs ont tendance à juger plus sévèrement et à accorder davantage d'importance à la précision lexicale ou grammaticale, tandis que les évaluateurs natifs privilégient généralement l'intelligibilité globale du discours. Le degré de familiarité de l'évaluateur avec l'accent du locuteur peut également influencer le jugement : par exemple, un évaluateur habitué à entendre un anglais parlé par des locuteurs japonais éprouvera souvent moins de difficultés à comprendre qu'un évaluateur qui n'y est pas accoutumé (Didelot et al., 2019 ; Kim & di Gennaro, 2012 ; Minematsu et al., 2004).

Ce manque de formation et d'outillage des enseignants pour enseigner et évaluer la prononciation est relevé depuis plusieurs années (Amengual-Pizarro & García-Laborda, 2017 ; Baker, 2011 ; Burgess & Spencer, 2000 ; Derwing & Munro, 2015 ; Gilquin et al., 2022 ; Piccardo, 2016 ; Rogerson-Revell, 2021). À cela s'ajoute un manque général de temps et de ressources humaines : dans l'enseignement secondaire comme à l'université, les classes de langues comptent souvent 20 à 30 élèves pour un enseignant, avec seulement 2 à 4 heures de cours hebdomadaires. Dans ce contexte, il est difficile pour chaque apprenant de bénéficier de retours personnalisés sur sa prononciation (Muñoz, 2014).

La prononciation se retrouve alors souvent reléguée à un petit encadré à la fin des leçons de manuels, à un “ ‘*add it on we have time’ language feature*” (Levis, 2018, p. 1). Pourtant, elle constitue un élément central de l'apprentissage des langues, qui influence profondément non seulement la capacité du locuteur à se faire comprendre, mais aussi sa faculté à comprendre les autres (Levis, 2018). Cette importance est d'ailleurs reconnue dans les tests certificatifs d'anglais, où la prononciation occupe une place centrale dans l'évaluation, contrastant avec la faible importance qui lui est accordée dans l'enseignement en classe (Gilquin et al., 2022 ; Henderson et al., 2012).

En parallèle, de nombreuses applications spécialisées dans l'enseignement des langues ont saisi cette opportunité pour proposer des fonctionnalités d'évaluation automatique de la prononciation. La majorité de ces applications se concentrent toutefois encore largement sur l'évaluation de la parole lue et proposent des feedbacks limités, souvent mal alignés avec les besoins réels des apprenants (Evanini & Zechner, 2019). En outre, l'évaluation que proposent ces outils repose généralement sur la comparaison à un modèle standardisé de la langue, où toute déviation constitue un accent à réduire ou éliminer (Saito, 2021). Comme le soulevaient déjà Neri et al. (2002), ces outils sont plus souvent le résultat d'une course technologique qu'une démarche répondant à des besoins pédagogiques identifiés.

En effet, les descripteurs de compétence en production orale, tels que ceux proposés par le Cadre Européen Commun de Référence pour les Langues (CECRL) ou ceux des tests certificatifs d'anglais comme le TOEFL et l'IELTS, se basent non pas sur un degré de déviation par rapport à une norme, mais sur la capacité de l'apprenant à être compris sans effort par son interlocuteur. Dans le domaine de l'acquisition des langues secondes (L2), cet effort de compréhension de la part de l'auditeur est généralement désigné par le terme de « compréhensibilité ». Parmi les paramètres clés de cette compréhensibilité, la fluence et le rythme de la parole occupent une place centrale, et le niveau CECRL B2 semble constituer un seuil déterminant à cet égard.

Cette thèse s'inscrit dans le prolongement de recherches fondamentales et appliquées menées à l'Université Grenoble Alpes entre 2013 et 2020 (ANR-11-IDFI-0024), ayant conduit à la création de SELF, un dispositif d'évaluation des compétences en langues à visée formative. Déployé dans une trentaine d'établissements en France et à l'international, SELF évalue actuellement trois habiletés langagières : la compréhension de l'oral, la compréhension de l'écrit et l'expression écrite courte. Cependant, l'évaluation automatisée de la production orale spontanée reste un défi majeur. Cette thèse constitue une première étape vers l'élaboration d'un module d'évaluation diagnostique de la production orale en anglais.

Trois objectifs principaux guident ce travail :

- Concevoir un outil d'évaluation automatique de la production orale spontanée, ciblant spécifiquement des phénomènes linguistiques susceptibles d'impacter la compréhensibilité du locuteur.
- Étudier la variation de ces phénomènes chez des locuteurs de niveau CECRL B1 et B2.
- Analyser l'impact de ces phénomènes sur la perception de l'effort de compréhension par des auditeurs natifs.

Le premier chapitre de cette thèse explore les méthodes et critères d'évaluation de la prononciation, aussi bien dans les principaux tests certificatifs d'anglais que dans les récents outils d'évaluation automatique. Le deuxième chapitre s'intéresse aux concepts d'intelligibilité et de compréhensibilité, en détaillant les méthodes utilisées pour les évaluer et les différents facteurs influençant leur jugement. Le troisième chapitre porte sur les notions de fluence et de rythme, avec une attention particulière au rôle des pauses et de l'accentuation lexicale. Ces trois premiers chapitres nous permettront d'élaborer notre problématique et nos questions de recherche, ainsi que d'exposer nos différentes hypothèses.

La deuxième partie de la thèse est consacrée à la méthodologie de recherche et à la présentation des données. Le chapitre 4 décrit les trois corpus de parole spontanée constitués dans le cadre de ce travail. Le chapitre 5 présente l'outil d'annotation automatique développé, ainsi que les métriques d'évaluation et la méthodologie d'analyse des résultats. Enfin, le chapitre 6 expose le protocole mis en place pour évaluer l'impact des phénomènes linguistiques étudiés sur la perception de l'effort de compréhension.

La troisième partie de la thèse est dédiée à la présentation des résultats. Le chapitre 7 détaille les performances des différents modules de traitement automatique. Le chapitre 8 présente les résultats de l'analyse des annotations effectuées sur les trois corpus de parole spontanée. Le chapitre 9 présente les résultats de l'évaluation de la perception de l'effort de compréhension. Enfin, le chapitre 10 propose une discussion approfondie des choix méthodologiques effectués et des résultats obtenus.

Première partie

Contexte théorique

Chapitre 1

Évaluation de la prononciation

Ce chapitre s'intéresse à la façon dont la prononciation est évaluée dans les tests certificatifs d'une part, et dans les systèmes d'évaluation automatique d'autre part. Nous proposons d'examiner dans un premier temps les descripteurs de compétences et les grilles d'évaluation de la prononciation que proposent les principaux tests certificatifs de l'anglais. Nous devrions ainsi mieux comprendre ce qu'il est attendu des apprenants à différents niveaux de compétence en langue, et identifier les bases sur lesquelles reposent (ou sont censées reposer) les jugements des évaluateurs. Nous nous intéresserons ensuite aux techniques d'évaluation automatique de la prononciation, en examinant les critères sur lesquels se basent les systèmes d'aujourd'hui, et en quoi ils diffèrent des évaluations humaines.

1.1 Évaluation humaine

1.1.1 Descripteurs du CECRL

Le Cadre Européen Commun de Référence pour les Langues (CECRL) est une initiative du Conseil de l'Europe pour définir des descripteurs de compétences détaillés afin de faciliter l'enseignement et l'évaluation des langues étrangères. La première édition de l'ouvrage (Conseil de l'Europe, 2001) propose une échelle dédiée à la « Maîtrise du système phonologique ». Cette échelle a toutefois été largement critiquée pour son manque de précision et ses descripteurs vagues basés sur l'intuition de l'évaluateur (ex. « net accent étranger » en A2, « prononciation clairement intelligible » en B1) et prenait pour modèle la prononciation d'un locuteur natif sans pour autant la définir (« prononciation et intonation claires et naturelles » au niveau B2). L'échelle complète est donnée en annexe A.1.

Lors de la mise au point de la nouvelle édition des descripteurs en 2018, ces limitations ont été reconnues par les concepteurs des nouveaux descripteurs (Piccardo, 2016) qui qualifient ces descripteurs phonologiques d'« échelle la moins réussie » du CECRL et de seule échelle avec une norme native (p. 133). Les nouveaux descripteurs abandonnent la comparaison au modèle natif et se focalisent sur l'intelligibilité comme base théorique principale du contrôle phonologique. Les auteurs définissent l'intelligibilité comme « l'accessibilité du sens pour les auditeurs, incluant également la difficulté de compréhension perçue par les auditeurs (habituellement désignée comme compréhensibilité) » (Conseil de l'Europe, 2018, p. 140). On accepte maintenant un accent qui n'affecte pas la compréhension au niveau C2, ainsi que l'influence d'autres langues connues par l'apprenant. Il y a maintenant trois échelles : « Maîtrise générale du système phonologique », « Articulation des sons » et « Traits prosodiques » (Conseil de l'Europe, 2018, p. 142). Le tableau complet est donné en annexe A.2.

Des termes relatifs à la prosodie, comme l'accent ou le rythme, sont maintenant mentionnés dès le niveau A1 : « très forte influence de l'accent, du rythme, et/ou de l'intonation de l'une ou l'autre des langues qu'il parle » ; au niveau B1 « l'intonation et l'accentuation des énoncés et des mots sont presque corrects » ; au niveau B2 « peut en général [...] placer correctement l'accent », « l'accent a tendance à subir l'influence de l'une ou l'autre des langues qu'il/elle parle, mais l'impact sur la compréhension est négligeable ou nul » ; au niveau C1 « peut prononcer un discours fluide et intelligible en ne faisant que de rares erreurs d'accent, de rythme et/ou d'intonation qui n'affectent ni la compréhension ni l'efficacité ». Du côté de la prononciation des phonèmes, il est fait mention de « produire correctement des sons dans la langue cible » (A1), de prononciation « en général intelligible » (A2), mais aussi de « mauvaise prononciation systématique des phonèmes » (A2) ou des « erreurs de prononciation de sons et de mots » (B1, B2) sans plus de détails, laissant une part importante à l'interprétation de l'évaluateur. Notons qu'à partir du niveau B2, l'influence des caractéristiques phonologiques sur la compréhension devient négligeable, et que le locuteur devient capable de « prédire avec une certaine précision les traits phonologiques de la plupart des mots non familiers (par ex. l'accent tonique en lisant) ».

Si la première édition du CECRL restait limitée au niveau de la prononciation, la nouvelle édition présente quant à elle des descripteurs plus détaillés, séparant la réalisation des phonèmes et les aspects prosodiques. Bien qu'ils ne soient pas exempts de critiques[1], ces descripteurs apportent déjà une base commune et solide pour évaluer la prononciation des apprenants.

[1] Didelot et al. (2019) critiquent notamment l'absence de considération des représentations sociales de l'auditeur sur la perception de l'intelligibilité, et le fait que le point de vue de l'auditeur de manière générale est peu pris en compte.

1.1.2 Descripteurs du CLES

Le Certificat de Compétences en Langues de l'Enseignement Supérieur (CLES) est une certification universitaire française établie par le Ministère de l'Enseignement Supérieur et de la Recherche. Le CLES est déployé aujourd'hui en 10 langues et proposé par une trentaine de centres CLES accrédités en France (rapport d'activité 2023[2]). Chaque niveau du CECRL est évalué indépendamment : le candidat doit choisir un niveau cible à valider lors de la passation de l'examen. Il existe des sessions CLES pour les niveaux B1, B2 et C1. Le CLES évalue quatre habiletés : la compréhension de l'écrit, la compréhension de l'oral, l'expression écrite ainsi que l'expression orale en monologue pour le niveau B1, et en interaction pour les niveaux B2 et C1.

Au niveau B2, l'épreuve consiste en une interaction orale sous la forme d'un jeu de rôle d'une dizaine de minutes à deux ou trois participants. Chaque participant se voit attribuer un rôle en faveur ou contre un sujet polémique, comme l'usage de la cigarette électronique ou des tests cliniques sur les animaux par exemple. Les candidats disposent de deux minutes de préparation avant la discussion, puis doivent échanger leurs points de vue et argumenter pour arriver à un compromis dans un temps imparti de dix minutes. Ils sont évalués en direct par un ou deux évaluateurs accrédités présents dans la salle. L'évaluation est faite sur huit critères : la capacité à prendre position et négocier, la pertinence et la variété des arguments, la capacité à interagir, l'aisance, la phonologie, la cohérence du discours, la précision grammaticale et enfin la pertinence et la variété lexicale (*cf.* grille d'évaluation en annexe A.3). Pour chacun des critères, l'évaluateur peut attribuer le niveau B2, ou à défaut B1 ou « non validé ». Le niveau B2 en interaction orale n'est validé que si l'ensemble des huit critères est validé au niveau B2.

Concentrons-nous sur les deux critères qui relèvent de l'évaluation de la prononciation : l'aisance et la phonologie. Le premier fait référence à la capacité de l'étudiant à « exprimer ses idées avec fluidité sans faire de longues pauses (hésitations tolérées) », et « exprimer ses idées malgré des pauses pour chercher ses mots ». Le critère phonologie est décrit par une « prononciation et intonation suffisamment claires pour être aisément compris(e), même si un accent subsiste » et « globalement compréhensible malgré l'accent étranger et/ou des erreurs de prononciation ».

Sur le site du CLES, on peut lire qu'il est attendu du candidat de niveau B2 qu'il soit « significativement plus fluide et fasse moins d'erreurs » qu'au niveau B1, et soit

[2]Disponible en ligne à l'adresse suivante : https://www.certification-cles.fr/medias/fichier/rapport-d-activite-2023-certification-cles_1705953556233-pdf

« aisément compréhensible »[3]. Le niveau B2 semble donc caractérisé par une certaine fluidité de parole et d'aisance de compréhension côté auditeur.

Le CECRL et le CLES proposent des descripteurs communs à toutes les langues, mais qu'en est-il pour les descripteurs spécifiquement rédigés pour l'anglais L2 ?

1.1.3 Descripteurs du TOEFL

Le *Test of English as a Foreign Language* (TOEFL) est un test certificatif pour évaluer l'anglais langue seconde et développé par l'organisme privé *Educational Testing Service*. Il se décline en plusieurs versions adaptées à des publics allant du primaire à l'université. Nous nous intéresserons ici à deux de ces versions : le *TOEFL iBT*, qui évalue les compétences de l'apprenant en situation académique et qui est le test le plus répandu, et le *TOEFL ITP Assessment Series*, présenté comme un test à visée formative utilisé par certaines universités pour mieux adapter les enseignements aux besoins des apprenants. Les deux versions se passent sur ordinateur.

TOEFL iBT

Le TOEFL iBT met en avant l'évaluation de la production orale de manière asynchrone : les candidats sont enregistrés en centre d'examen lors de la passation du test, et cet enregistrement est évalué ultérieurement de manière semi-automatique. Le TOEFL dit garantir la qualité de l'évaluation en permettant aux évaluateurs humains de se concentrer sur le contenu de la production, sans être biaisés par les apparences : "*No matter who you are, or how you sound, you can be 100% confident that the only thing our test raters score is all your hard work and English skills.*" (ETS.org[4]).

La section de production orale du TOEFL iBT se compose de 4 questions qui simulent des situations de la vie réelle de l'étudiant. Elles peuvent porter sur une thématique précise, mais aucune connaissance sur le sujet n'est requise. Le temps total estimé pour la section de production orale est de 16 min. Après chaque question, le candidat dispose d'un temps de préparation de 15 à 30 s, puis doit enregistrer sa réponse au microphone pendant 45 à 60 s selon l'exercice.

- **Question 1 :** Le candidat doit se positionner par rapport à un cas présenté, en exprimant ses préférences et en argumentant son discours. L'énoncé est écrit à l'écran ; le candidat dispose de 15 s de préparation et 45 s pour donner sa

[3] https://www.certification-cles.fr/se-preparer/grilles-d-evaluation/grilles-d-evaluation-1196363.kjsp, consulté le 24/11/2024)

[4] https://www.ets.org/toefl/test-takers/ibt/scores.html, consulté le 22/07/2024

réponse. Exemple d'énoncé tiré d'une vidéo tutoriel : "*Some people think it is more fun to spend time with friends in restaurants or cafés. Others think it is more fun to spend time with friends at home. Which do you think is better? Explain why.*" (ETS.org[5])

- **Questions 2 à 3 :** Elles combinent la production orale avec la compréhension de l'oral et de l'écrit :
  - **Question 2 :** Le candidat lit un court texte à propos de la vie étudiante, par exemple une annonce écrite sur un panneau d'annonce à l'université, puis il écoute une conversation entre deux personnes à propos de ce texte, où l'un des locuteurs donne son avis. Le candidat doit alors résumer l'avis de la personne en 60 s, après un temps de préparation de 30 s.
  - **Question 3 :** Le candidat lit un texte à propos d'une notion académique donnée, puis écoute un bref extrait de cours sur le même sujet, et doit ensuite expliquer la notion présentée et comment l'exemple donné dans la vidéo illustre ce concept. Temps de préparation 30 s, temps de réponse 60 s.
- **Question 4 :** Le candidat écoute un nouvel extrait de cours et doit le résumer en listant les points mentionnés par l'enseignant. Temps de préparation 20 s, temps de réponse 60 s.

Plusieurs conseils sont donnés aux candidats : parler de manière continue pendant 45 secondes, sans se répéter et sans parler trop vite ; éviter les faux départs et les arrêts brutaux qui rendent le flux de parole saccadé ; connecter et varier ses arguments, bien noter les arguments donnés par la personne de la conversation et les mentionner dans la réponse.

La grille complète d'évaluation de la production orale du TOEFL iBT est disponible en ligne[6] et donnée en annexe A.4. Elle est séparée en deux parties : *Independent Speaking Rubric* pour les questions 1 et 4, et *Integrated Speaking Rubric* pour les questions 2 et 3. Chaque question est évaluée sur quatre critères de manière holistique sur une échelle de 0 à 4. Il y a trois critères différents : *Delivery*, pour la qualité de la prononciation et la fluidité de la parole ; *Language use*, pour la précision lexicale et grammaticale ; et *Topic development*, pour la précision et la clarté de la réponse formulée par le candidat. Concentrons-nous sur la rubrique *Delivery* (*cf.* tableau 1.1). Les descripteurs mettent en avant l'effort requis par l'auditeur pour comprendre et l'« intelligibilité » du locuteur ("*intelligibility*" , terme toutefois non défini). Au niveau

[5] https://www.ets.org/toefl/test-takers/ibt/about/content/speaking.html, consulté le 22/07/2024
[6] https://www.ets.org/pdfs/toefl/toefl-ibt-speaking-rubrics.pdf, consulté le 29/11/2024

Score	Delivery
4	Generally well-paced flow (fluid expression). Speech is clear. It may include minor lapses, or minor difficulties with pronunciation or intonation patterns, which do not affect overall intelligibility.
3	Speech is generally clear, with some fluidity of expression, though minor difficulties with pronunciation, intonation, or pacing are noticeable and may require listener effort at times (though overall intelligibility is not significantly affected).
2	Speech is basically intelligible, though listener effort is needed because of unclear articulation, awkward intonation, or choppy rhythm/pace ; meaning may be obscured in places.
1	Consistent pronunciation, stress and intonation difficulties cause considerable listener effort ; delivery is choppy, fragmented, or telegraphic ; frequent pauses and hesitations.
0	Speaker makes no attempt to respond OR response is unrelated to the topic.

Tab. 1.1 : *Grille d'évaluation de la production orale du TOEFL iBT Independent Speaking Rubric, section Delivery*

2, le locuteur est intelligible mais demande des efforts à l'auditeur (articulation peu claire, intonation étrange, rythme saccadé). Au niveau 3, il commence à être un peu plus fluide (difficultés mineures de prononciation, intonation et rythme qui peuvent demander un certain effort de la part de l'auditeur mais affectent peu l'intelligibilité). Au niveau 4, le discours est généralement fluide avec des difficultés mineures qui n'affectent pas l'intelligibilité. Les descripteurs sont pratiquement mot pour mot identiques dans la partie *Integrated Speaking Rubric*.

TOEFL ITP Assessment Series

Le *TOEFL ITP Assessment Series* est un test à visée formative destiné à évaluer les compétences des étudiants pour mieux adapter l'enseignement qui leur est proposé. La production orale est évaluée par le *TOEFL ITP Speaking test*[7]. Celui-ci dure environ 15 min et est composé d'une tâche de lecture à voix haute après écoute d'un modèle, deux questions à réponse ouverte sur un sujet familier, et une question portant sur une conversation enregistrée entre deux étudiants. Dans chaque cas, l'énoncé est écrit à l'écran et lu à voix haute. Une fois la lecture terminée, un chronomètre s'active pour le temps de préparation, puis un autre pour l'enregistrement. Il n'est possible de faire une pause qu'entre les questions. Seule la conversation de la dernière question n'est pas transcrite. Comme pour le TOEFL iBT, le temps de préparation est limité (entre 30 s et 60 s selon les questions) et le temps d'enregistrement est fixé entre 45 s et 60 s.

[7] Une version démo est disponible en ligne : https://www.ets.org/toefl/itp/prepare.html

La grille d'évaluation de la production orale pour le TOEFL ITP est accessible en ligne[8] et donnée en annexe A.5. Elle décrit quatre niveaux de compétence de A2 à C1. Les descripteurs semblent être un mélange de ceux du TOEFL iBT et du CECRL de 2018. Au niveau A2, le locuteur est intelligible sur des sujets familiers, mais requiert un certain effort de la part de l'auditeur ; forte influence de la L1 sur la prononciation et l'accent lexical, discours entrecoupé de nombreuses pauses et faux-départs. Au niveau B1, l'accent, l'intonation et le rythme commencent à être maîtrisés mais restent parfois influencés par la L1. En B2, la parole est globalement fluide et bien rythmée *("well-paced")* malgré quelques hésitations ; l'accent et l'intonation sont maîtrisés malgré quelques erreurs. En C1, enfin, la parole est fluide, sans effort ni hésitation ; l'accent et l'intonation sont utilisés de manière stratégique.

Autant pour le TOEFL iBT que ITP, les premiers niveaux (respectivement 1 et A2) sont caractérisés par des difficultés de prononciation, d'accentuation et d'intonation dues à une forte influence de la L1, un rythme saccadé *("choppy, fragmented, telegraphic")* et de nombreuses pauses et hésitations demandant un effort important pour comprendre. Ces difficultés sont présentes dans une moindre mesure au niveau 2 ou B1, mais toujours avec une forte influence de la L1. On constate un changement clair au niveau 3 ou B2, où la parole devient globalement fluide avec une bonne maîtrise de l'accent, de l'intonation et du rythme, et seulement des erreurs mineures qui s'estompent encore au dernier niveau.

1.1.4 Descripteurs du TOEIC

Le *Test of English for International Communication* (TOEIC) est un test également produit par *Educational Testing Service*. Nous nous intéresserons ici à deux versions différentes : le *TOEIC Speaking Test*, qui met l'accent sur la communication en milieu professionnel, et le *TOEIC Bridge Speaking Test*, adapté pour les plus petits niveaux. Les deux se passent en ligne et en autonomie.

Le *TOEIC Speaking Test* se compose de 11 tâches et dure environ 20 min. Deux tâches sont des lectures à voix haute (45 s de préparation, 45 s d'enregistrement), deux autres sont une description d'image (45 s de préparation, 30 s d'enregistrement), les trois suivantes sont une simple question (3 s de préparation, 15 s ou 30 s de réponse), suivies de trois autres questions relatives à un court texte (45 s de lecture), et la dernière tâche demande au candidat d'exprimer son opinion (45 s de préparation, 60 s d'enregistrement).

[8] https://www.ets.org/pdfs/toefl/toefl-itp-speaking-descriptors.pdf, consulté le 29/11/2024

Le TOEIC met à disposition une grille de descripteurs de niveaux[9] ainsi qu'une grille d'évaluation pour chaque type de tâche[10], données en annexe A.6. La grille de descripteurs présente huit niveaux de compétence (de 1 à 8). On notera des difficultés constantes de prononciation, d'accentuation et d'intonation au niveau 4, une prononciation peu claire ou une intonation ou un accent inappropriés au niveau 6, ou encore de longues pauses et de fréquentes hésitations aux niveaux 4 et 5. Il est également fait mention de difficultés pour comprendre (niveaux 2 à 4) qui s'estompent peu à peu pour laisser place à une parole "*generally intelligible*" (niveau 5), "*intelligible*" (niveau 6) et "*highly intelligible*" (niveau 7).

En parallèle de ces huit descripteurs, on trouve une grille d'évaluation pour chaque type de tâche. Pour la lecture à voix haute, la performance du candidat est évaluée sur une échelle de quatre niveaux (de 0 à 3) et selon deux critères : *Pronunciation* et *Intonation and Stress*. Il est fait ici mention d'intelligibilité et du degré d'influence de la L1, de l'utilisation plus ou moins appropriée de pauses, d'emphase et de l'intonation. L'évaluation des réponses à la description d'images et aux questions est également effectuée sur quatre niveaux, mais plus orientée sur l'adéquation de la réponse en termes de contenu, de choix de vocabulaire et de structures utilisées. Il est toutefois toujours fait mention d'intelligibilité du locuteur et de l'effort de compréhension demandé à l'auditeur. Quant à l'évaluation de l'expression d'opinion, enfin, elle est effectuée sur six niveaux plus détaillés, qui reprennent mot pour mot la grille du TOEFL iBT (*cf.*annexe A.6).

Le *TOEIC Bridge Speaking Test* évalue les compétences communicationnelles aux niveaux débutant et intermédiaire. Les candidats répondent à des questions simples sur des sujets familiers et utilisent des phrases pour décrire des événements de la vie quotidienne. Ils peuvent être amenés à expliquer brièvement leur opinion ou leurs projets, et raconter des histoires simples. Il est mentionné dans le livret de l'examinateur[11] qu'il est attendu des candidats de pouvoir prononcer les mots de manière à être compris par un locuteur anglophone, en utilisant l'intonation, l'accent et les pauses pour rythmer *(* "*pace*" *)* la parole et faciliter la compréhension *(* "*contribute to comprehensibility*" *)*.

Le test est composé de huit tâches et dure environ 15 min. Il comprend deux tâches de lecture à voix haute, deux descriptions d'image, une tâche de type *listen and retell*, un enregistrement de message vocal, une narration d'histoire sur la base d'une suite d'images et la formulation d'une recommandation sur la base d'un texte court donné à l'écrit.

[9] https://www.ets.org/pdfs/toeic/toeic-speaking-writing-score-descriptors.pdf, (22/11/2024)
[10] https://www.ets.org/pdfs/toeic/toeic-speaking-writing-examinee-handbook.pdf (idem)
[11] https://www.ets.org/pdfs/toeic/toeic-bridge-speaking-writing-examinee-handbook.pdf, consulté le 22/11/2024

On trouve ici encore une grille d'évaluation pour chaque type de tâche, mais toutes relativement similaires. Elles proposent 4 à 5 niveaux de compétence, et se concentrent sur le degré de complétion de la tâche et la pertinence du vocabulaire et des structures utilisées. Au niveau de la prononciation, l'évaluation reste très subjective et se contente globalement de varier les adverbes : "*mostly unintelligible*" (niveau 1), "*sometimes unintelligible*" (niveau 2), "*generally intelligible*" (niveau 3). L'effort demandé à l'auditeur est aussi clairement mentionné : "*requires listener effort to understand*" (niveau 1), là encore avec différents adverbes selon les niveaux. La grille d'évaluation de la lecture à voix haute contient un peu plus d'éléments relatifs à la prononciation : "*intonation and stress are somewhat appropriate*" (niveau 1), "*mostly appropriate*" (niveau 3) ; "*lapses and/or other language influence are present*" (niveau 2), "*other-language influence does not affect overall intelligibility*" (niveau 3). De manière générale, l'évaluation porte sur l'intelligibilité du locuteur et l'effort requis pour le comprendre. Précisons qu'il est mentionné page 37 du livret de l'examinateur que "*intonation and stress refer to your ability to use emphases, pauses, and rising and falling pitch to convey meaning to a listener*".

1.1.5 Descripteurs de IELTS

L'*International English Language Testing System* (IELTS) est une certification internationale cogérée par l'université de Cambridge, le British Council et la société australienne IDP Education Limited. L'IELTS évalue les compétences du candidat sur les quatre habiletés sur un test d'une durée approximative de 2 h 45 min. Le test de production orale est une interview en face-à-face avec un examinateur durant 11 à 14 min. L'interview se compose de trois parties : une présentation personnelle du candidat, une discussion à partir d'une *task-card* présentant un sujet à aborder (monologue de 2 min puis questions-réponses), et une discussion approfondie sur ce sujet.

La version publique de la grille d'évaluation de la production orale du test IELTS est accessible en ligne[12] et donnée en annexe A.7. Elle décrit neuf niveaux selon quatre critères : fluidité et cohérence, ressources lexicales, variété et précision grammaticale, et prononciation.

Concernant la fluidité, les descripteurs mentionnent principalement l'influence des pauses dans les niveaux 2 à 4 (niveau 2 : "*pauses lengthily before most words*", niveau 3 : "*speaks with long pauses*", niveau 4 : "*noticeable pauses*"), à partir de 5, la parole est fluide dans les contextes simples. Les répétitions, auto-corrections et hésitations sont mentionnées des niveaux 4 à 9 (niveau 4 : "*frequent repetition and self-correction*", niveau 6 : "*occasional repetition, self-correction or hesitation*", niveau 9 :

[12] https://assets.cambridgeenglish.org/webinars/ielts-speaking-band-descriptors.pdf

"*rare repetition or self-correction ; any hesitation is content-related rather than to find words or grammar*").

Les descripteurs de prononciation sont limités et restent difficiles à interpréter. Il est fait mention de "*pronunciation features*" et "*mispronunciations*" sans plus de détails. La perception côté auditeur apparaît brièvement (niveau 4 : "*mispronunciations are frequent and cause some difficulty for the listener*" , niveau 9 : "*effortless to understand*"), de même pour l'intelligibilité (niveau 8 : "*L1 accent has minimal effect on intelligibility*" *).* Il n'est fait nulle part mention d'accentuation ou d'intonation.

Premières conclusions

La prononciation apparaît comme un critère clé de l'évaluation de la production orale dans les principaux tests certificatifs de l'anglais. Ces tests, ainsi que le CECRL, mettent fortement l'accent sur l'intelligibilité du locuteur et l'effort demandé à l'auditeur pour le comprendre. La fluidité de la parole est un aspect qui revient souvent (CECRL, TOEFL, TOEIC, CLES), notamment l'utilisation des pauses (TOEFL, TOEIC, IELTS, CLES) ; l'accent *(stress)* est souvent mentionné (CECRL, TOEFL, TOEIC), ainsi que l'intonation (CECRL, TOEIC, IELTS, CLES). Si ces paramètres sont explicités, le jugement du caractère approprié de leur utilisation est quant à lui souvent laissé à l'évaluateur (CLES B2 : « exprime ses idées avec fluidité », TOEFL iBT niveau 4 : "*generally well-paced flow*" , IELTS niveau 8 : "*wide range of pronunciation features*" , CECRL A2 : « les traits prosodiques (par ex. l'accent tonique) des mots familiers et quotidiens et des énoncés simples sont convenables », CECRL B2 : « peut généralement placer correctement l'accent »). Lorsque des précisions sont données pour aider l'évaluateur, elles restent relativement limitées. Par exemple, le caractère fluide de la parole est déterminé par l'absence de longues pauses pour le CLES, ou de répétitions, auto-corrections et hésitations non liées au contenu pour IELTS. De manière générale, même lorsqu'il est fait mention d'une influence notable de la L1, ce sont avant tout les phénomènes affectant l'intelligibilité qui sont considérés comme problématiques.

À l'issue de ce tour d'horizon, nous avons donc une idée plus précise des paramètres à cibler lors de l'évaluation de la prononciation : tout ce qui affecte l'intelligibilité du locuteur, et en premier lieu la fluidité (relative à la présence de pauses et d'hésitations), et les traits prosodiques comme l'accentuation et l'intonation. Qu'en est-il maintenant des systèmes d'évaluation automatique de la prononciation ? Ciblent-ils les mêmes paramètres ? Et sur quelles bases reposent leurs jugements ?

1.2 Évaluation automatique

Les premiers systèmes d'évaluation automatique de la prononciation sont arrivés dans les années 90 avec les débuts de la reconnaissance automatique de la parole. Le système Autograder (Bernstein et al., 1990) est pionnier dans le domaine : il présente une liste de questions à choix multiple, pour lesquelles l'apprenant est amené à lire à voix haute l'une des options de réponse. La machine identifie alors l'option qui a été prononcée, et donne un score basé sur le nombre de mots correctement reconnus. Un peu plus tard, les systèmes VILTS (*Voice Interactive Language Training Systems*, Neumeyer et al., 1996, Franco et al., 1997) permettent de donner n'importe quel texte à la machine pour le faire lire à l'apprenant. Les scores sont calculés à partir de mesures segmentales (reconnaissance des phonèmes) et suprasegmentales (durée des phonèmes, des syllabes ou débit de parole). Les premiers systèmes qui évaluent la parole spontanée arrivent dans les années 2000, et se focalisent sur la fluence de la parole (Cucchiarini et al., 2002), mais la parole spontanée reste toutefois marginale par rapport à la parole lue.

Après un fort engouement pour l'évaluation automatique de la prononciation à la fin des années 90, la discipline s'essouffle à cause d'une mauvaise fiabilité des systèmes alors commercialisés (Witt, 2012). Elle revient toutefois rapidement à la charge avec la généralisation des smartphones et l'amélioration des systèmes de reconnaissance de parole à la fin des années 2000. Notons la création de la conférence *Speech and Language Technology for Education* (SLaTE) au sein de l'*International Speech Communication Association* (ISCA) en 2007, qui se consacre spécifiquement à l'apprentissage des langues et les technologies de la parole (Ellis & Bogart, 2007).

Vingt ans après ces débuts, à quoi ressemblent les systèmes et comment évaluent-ils la prononciation ?

Commençons par distinguer deux types d'évaluation. L'évaluation certificative d'une part *(high-stake assessment)*, dont l'objectif premier est de déterminer le niveau du candidat et peut être la condition d'obtention d'un diplôme, d'un emploi, voire d'un visa ; et l'évaluation formative d'autre part *(low-stake assessment)*, qui a pour but d'identifier les difficultés de l'apprenant et lui fournir un feedback pour l'aider à progresser. Il apparaît dans la littérature que ces deux types d'évaluation se distinguent par les techniques qu'elles mettent en œuvre pour évaluer la prononciation : la première choisit généralement d'entraîner des modèles à prédire le score global du locuteur sur la base d'évaluations humaines, tandis que la deuxième cherche plutôt à identifier et mesurer des phénomènes cibles afin de proposer un feedback formatif à l'utilisateur.

1.2.1 Évaluation certificative

Ces dernières années, de plus en plus de tests certificatifs se sont équipés de systèmes d'évaluation automatique de la production orale. Ces outils sont conçus pour prédire le score d'un candidat à partir d'un enregistrement audio, on parle de *machine scoring* (Davis & Papageorgiou, 2021). Un grand nombre de productions d'apprenants est évalué par des experts sur des critères similaires à ceux mentionnés dans la section précédente, et un modèle est entraîné à prédire les scores donnés par les évaluateurs à partir de mesures automatiques diverses. Educational Testing Service fait partie des leaders mondiaux en la matière avec des entraînements sur plusieurs centaines de milliers d'enregistrements de candidats (Loukina & Yoon, 2019). Une fois le modèle entraîné, le système est capable de prédire le score d'un nouvel enregistrement qui n'a pas été évalué manuellement.

Les mesures utilisées sont avant tout des paramètres en lien avec la fluence, mais de plus en plus souvent combinés avec des paramètres spectraux (Evanini & Wang, 2013 ; Fontan et al., 2018), lexicaux (Yoon et al., 2012) ou syntaxiques (Bhat & Yoon, 2015 ; L. Chen & Zechner, 2011 ; Loukina et al., 2015). Parmi les paramètres de fluence utilisés, on retrouve généralement le débit de parole et d'articulation, la fréquence de pauses pleines et silencieuses et leur durée moyenne, le nombre de syllabes par unité rythmique, la durée des syllabes ou des voyelles, ou encore les variations d'intonation et d'intensité. Ces techniques combinent souvent de nombreux paramètres (77 pour Coutinho et al., 2016, 75 pour Loukina et al., 2015), mais ce sont souvent les mêmes qui obtiennent la meilleure corrélation avec les jugements humains : le débit de parole et d'articulation, la proportion de pauses et leur durée moyenne, la longueur des segments entre pauses.

Dès les premiers systèmes de ce type dans les années 2000, la corrélation humain/machine est comparable à la corrélation inter-évaluateurs : 0,8 pour Neumeyer et al. (2000) et Cucchiarini et al. (2002), 0,7 pour Moustroufas et Digalakis (2007). On tourne autour des mêmes valeurs aujourd'hui, même en parole spontanée : 0,8 pour Fu et al. (2020), 0,8-0,9 pour Shen et al. (2021), 0,8 encore pour Saito et al. (2022). Si la corrélation avec les évaluations humaines est élevée, ces outils restent toutefois limités lorsqu'il s'agit d'évaluer des paramètres de plus hauts niveaux, comme l'organisation du discours, la précision grammaticale ou la cohérence de l'énoncé (Isaacs, 2018). Aussi sont-ils souvent combinés avec des jugements humains pour garantir une meilleure fiabilité des résultats, tout en bénéficiant des avantages de l'évaluation automatique – on parle de *hybrid human-machine scoring*. On considère généralement trois approches : l'approche hybride confirmatoire, l'approche contributive parallèle et l'approche contributive divergente (Davis & Papageorgiou, 2021). Dans la première, l'évaluation automatique est seulement utilisée pour confirmer l'évaluation manuelle d'un examinateur ; si l'écart entre les deux évaluations est jugé trop grand, un se-

cond examinateur est mobilisé. Dans l'approche contributive parallèle, deux scores holistiques, un manuel et un automatique, sont combinés pour déterminer le score final. Dans l'approche contributive divergente, enfin, les évaluations humaines et automatiques portent sur des aspects différents et complémentaires de la production, typiquement de bas niveau pour la machine et de plus haut niveau pour l'humain.

Parmi les tests certificatifs qui intègrent des systèmes de prédiction de scores pour l'évaluation de la production de l'oral en anglais, on trouve le TOEFL iBT avec son système SpeechRater, intégré selon l'approche contributive parallèle (Evanini & Zechner, 2019) ; les Versant English Tests (complètement automatisés, Pearson Education, 2022), le Pearson Test of English (Pearson PTE, 2024), le Duolingo English Test (complètement automatisé également, Cardwell et al., 2024), le Cambridge Assessment English Linguaskill General Speaking Test avec son système Custom Automated Speech Engine (Xu et al., 2021). Le bénéfice de l'utilisation de tels systèmes pour les organismes certificatifs est grand : s'il est coûteux à concevoir, il est vite rentabilisé par les économies en termes de ressources humaines nécessaires pour évaluer les productions des candidats, et permet de réduire drastiquement le temps de délivrance des résultats qui se compte en semaines pour les évaluations humaines (Isaacs, 2018). Par ailleurs, la systématicité de l'évaluation est souvent mise en avant comme garante d'une évaluation plus équitable.

Si ces systèmes se révèlent performants pour prédire un score global à partir d'une production orale, ils sont toutefois peu exploitables en contexte diagnostique, étant donné que les paramètres sur lesquels ils se basent sont majoritairement de bas niveau. Si le score du candidat est influencé par le débit de parole ou la fréquence des pauses, ce n'est pas pour autant que ce dernier doit parler plus vite ou faire moins de pauses. Ces phénomènes sont une conséquence de difficultés en amont, mais pas nécessairement un problème en soi. En contexte formatif, les systèmes d'évaluation ont donc dû adopter d'autres approches pour pouvoir donner un feedback pédagogique à l'utilisateur.

1.2.2 Évaluation formative

Un grand nombre d'applications d'apprentissage des langues proposent aujourd'hui des fonctionnalités d'évaluation de la prononciation. Les plus en vogue en 2022 étaient Duolingo[13], Memrise[14], Babbel[15], Busuu[16] ou Rosetta Stone[17]. D'autres applica-

[13] DuoLingo Inc. (2022). https://www.duolingo.com/
[14] Memrise (2022). https://www.memrise.com/
[15] Babbel (2022). https://uk.babbel.com/
[16] Busuu Online S.L. (2022). https://www.busuu.com/
[17] Rosetta Stone Inc. (2022, v5.0.37). https://www.rosettastone.com/

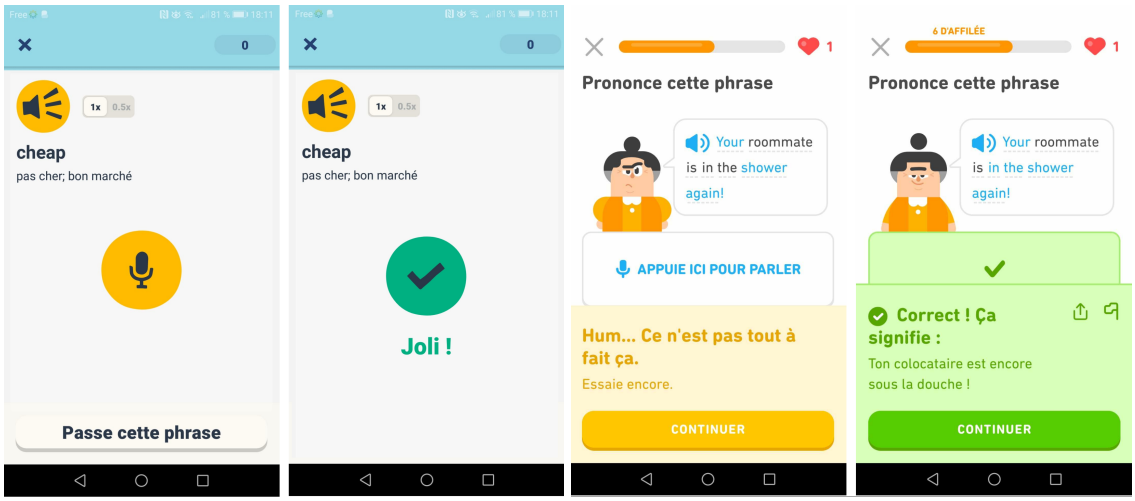


FIG. 1.1 : Captures d'écran de Memrise (gauche) et Duolingo (droite) en novembre 2022.

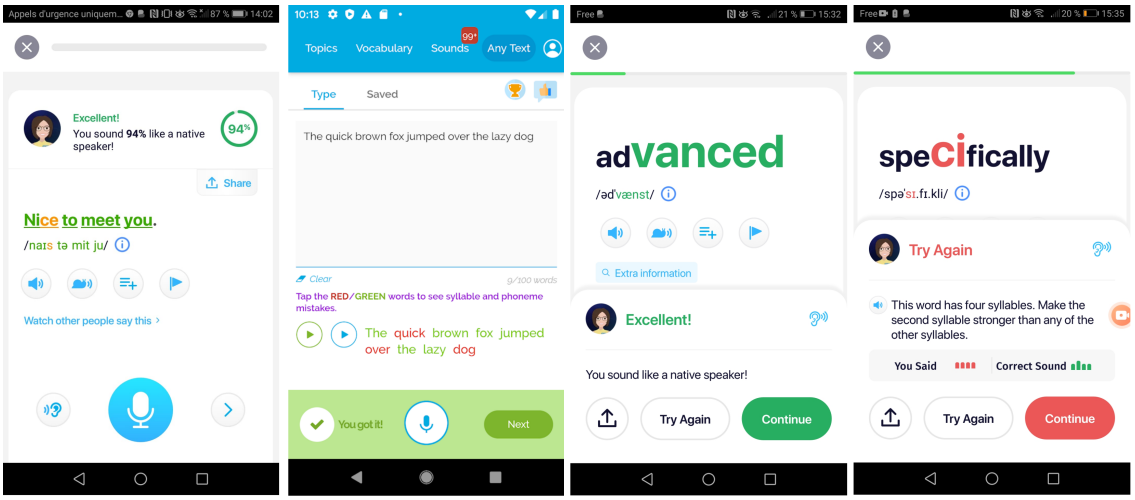
tions sont dédiées à la prononciation de l'anglais, comme ELSA[18] ou IELTS Speaking Practice[19] par exemple. Cette sous-section présente la manière dont ces applications évaluent la prononciation, le type d'activités qu'elles proposent et les feedbacks qu'elles donnent aux utilisateurs. Nous nous basons ici sur une description plus complète des applications proposée dans un chapitre d'ouvrage (Coulange, 2023).

En 2022, l'activité de production orale la plus courante dans les applications d'apprentissage des langues consistait à lire à voix haute ou à répéter un mot ou un énoncé en appuyant sur un bouton d'enregistrement. La production est analysée en temps réel et un feedback immédiat est affiché. Toutes les applications mentionnées dans le paragraphe précédent proposent ce type d'activité, accompagnée d'un modèle audio de ce qui doit être prononcé et de la transcription orthographique affichée par défaut, sauf pour Rosetta Stone qui utilise parfois des images sans texte ni audio pour éliciter la parole. Certaines applications affichent également une traduction dans la langue de l'utilisateur, par défaut, comme Memrise ou Babbel, ou sur demande et mot à mot comme Duolingo. L'audio peut être accompagné d'une vidéo ou d'une image fournissant des indices contextuels. ELSA propose également une transcription phonétique, une fonctionnalité qui semble absente des autres applications. ELSA et Memrise permettent aussi de ralentir la vitesse de lecture.

Dans la majorité des cas, les feedbacks donnés à l'utilisateur sont binaires (correct ou incorrect). Dans Duolingo, par exemple, un écran vert s'affiche avec des félicitations lorsque la production est validée, et dans le cas contraire un écran orange indique à l'apprenant qu'il peut mieux faire, sans toutefois donner de conseils pour améliorer sa

[18] Elsa Speak (2022). https://elsaspeak.com/
[19] SpeechAce LLC (2022). https://www.speechace.com/

prononciation (*cf.* figure 1.1). D'autres applications affichent un pourcentage de réussite indiquant dans quelle mesure les mots ont été reconnus par le système. ELSA et IELTS Speaking Practice vont un peu plus loin en affichant les mots ou les lettres en couleurs : vert pour les mots ou phonèmes reconnus correctement, orange quand ce n'est pas tout à fait bon, et rouge quand le mot ou le phonème est incorrect ou manquant (*cf.* figure 1.2a). En cliquant sur un mot, l'apprenant peut voir les phonèmes attendus et ceux qui ont été identifiés par le système, accompagnés de conseils explicites pour prononcer chaque phonèmes. ELSA propose également une activité dédiée à l'accent lexical. Un mot isolé apparaît à l'écran avec la syllabe à accentuer écrite en gros caractères. Après l'enregistrement, cette syllabe est colorée en vert ou en rouge selon la syllabe accentuée par l'utilisateur. Une représentation visuelle des syllabes est également affichée sous forme de barres plus ou moins hautes pour symboliser la position de l'accent primaire (*cf.* figure 1.2b).



(a) ELSA (gauche) et IELTS Sp. Pr. (droite) *(b) Accent lexical sur ELSA*

FIG. 1.2 : Captures d'écran de novembre 2022

En septembre 2022, ELSA a déployé une nouvelle fonctionnalité premium appelée Speech Analyzer, permettant aux étudiants d'enregistrer une production orale libre et d'obtenir un score global de production orale, ainsi que des scores détaillés pour la prononciation, l'intonation, la fluidité, la grammaire et le vocabulaire. ELSA fournit également des prédictions de scores pour IELTS, TOEFL, Pearson ainsi que le niveau CECRL. Comme pour la lecture à voix haute, Speech Analyzer identifie les erreurs segmentales et propose des conseils pour les corriger. L'outil calcule également des scores d'intonation, de débit de parole et de pauses, exprimés sous forme de pourcentages (*cf.* figure 1.3).

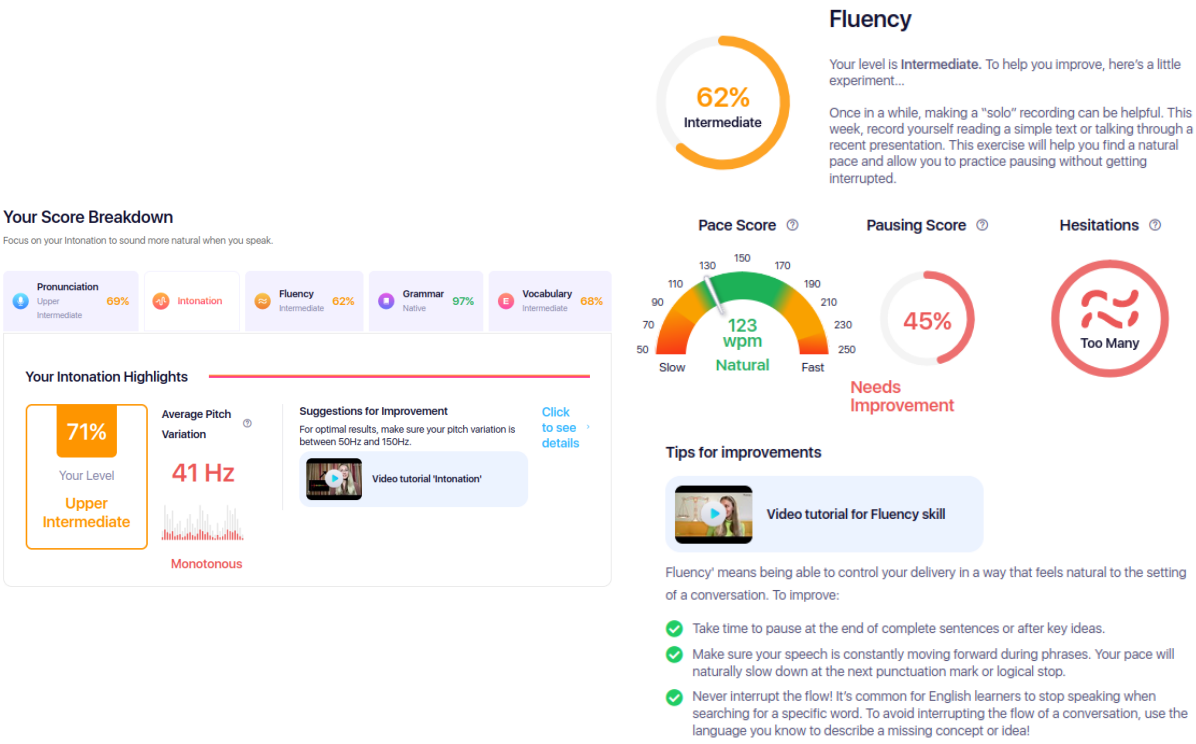


FIG. 1.3 : Captures d'écran de Speech Analyzer (ELSA) en novembre 2022.

À l'exception de l'exercice de détection de l'accent lexical d'ELSA, tous les systèmes mentionnés ci-dessus calculent un score de type *Goodness Of Pronunciation* (GOP) (Witt, 1999), basé sur le niveau de confiance d'un système de reconnaissance vocale. La réponse est considérée comme correcte lorsque le score de confiance dépasse un certain seuil. Dans IELTS Speaking Practice ou ELSA, la reconnaissance phonémique non contrainte permet à l'apprenant de voir quels phonèmes ont été reconnus par le système, bien qu'il soit limité aux phonèmes de l'anglais et au nombre de phonèmes du mot ou de l'énoncé de référence. Aucune information n'a toutefois été trouvée sur le fonctionnement des outils de détection de la syllabe accentuée pour ELSA.

Du côté de la recherche publique, on trouve un grand nombre d'études proposant des systèmes pour évaluer la prononciation. Elles aussi proposent dans leur écrasante majorité le calcule d'un score de type GOP, sur la base de la reconnaissance des phonèmes. Ces scores se déclinent toutefois dans une grande diversité : certaines études proposent d'adapter le système de reconnaissance à la parole L2, en augmentant un lexique phonétisé avec des erreurs phonologiques typiques (*Extended Recognition Network*, Bada et al., 2020 ; Lee et Glass, 2015), ou en adaptant les modèles acoustiques à partir d'enregistrements de la langue maternelle des apprenants (Goronzy et al., 2004 ;

Tan, 2008), ou directement avec de la parole L2 (Duan et al., 2017 ; W. Li et al., 2016). D'autres encore proposent de comparer la reconnaissance d'un système entraîné sur de la parole native avec celle d'un système adapté pour la parole L2, le score est alors basé sur la différence des deux sorties et permet de se passer du texte de référence (*Reference free Error Rate*, Fu et al., 2020 ; Naijo et al., 2021). D'autres systèmes s'emploient à comparer des mesures acoustiques (durées, débit, intonation, traits phonétiques etc.) directement avec un modèle, qu'il s'agisse d'un enregistrement du même énoncé par un locuteur natif (Arias et al., 2010 ; Ding et al., 2020), ou un modèle appris sur un ensemble d'enregistrements pour permettre plus de variabilité (Truong et al., 2018 ; WANG et al., 2015). À part Fu et al. (2020), toutes les études mentionnées ici se concentrent sur l'évaluation de mots ou de phrases lues.

Dans la majorité des systèmes d'évaluation formative, la prononciation est mesurée en termes de distance par rapport à un modèle natif. ELSA va même jusqu'à afficher un pourcentage dans ses feedbacks : "*You sound 94% like a native speaker!*" (*cf.* figure 1.2a). Ces systèmes reposent souvent sur la reconnaissance automatique de la parole pour générer leurs scores, qu'il s'agisse de taux de reconnaissance de mots ou de phonèmes. Certains intègrent des dimensions prosodiques, telles que le débit de parole ou la fréquence des pauses, mais les innovations dans ce domaine restent limitées.

Or, comme le souligne Isaacs (2018), "*the element of accent reduction that the software is targeting may be incompatible with helping learners become intelligible*" (p. 20). En effet, non seulement l'objectif de « parler comme un natif » est ambitieux et souvent irréaliste, mais il n'améliore pas nécessairement l'intelligibilité d'un locuteur (Derwing & Munro, 2015). Certains éléments évalués par ces systèmes automatiques sont secondaires pour la communication. Il semble donc plus pertinent d'encourager les apprenants à se concentrer sur les phénomènes qui entravent ou perturbent, voire rendent impossible, la compréhension mutuelle (Isaacs, 2018).

Conclusion

Les exigences des tests certificatifs accordent une place centrale à l'intelligibilité et la compréhensibilité du locuteur. Le rythme, la fluidité, les pauses, l'intonation ou l'accent sont régulièrement cités comme éléments clés pour estimer le niveau de compétence du locuteur.

Les systèmes d'évaluation automatique adoptés par ces tests offrent des solutions techniques pour estimer un niveau de production orale. En s'appuyant sur des mesures de bas niveau, comme le débit de parole, la fréquence des pauses ou le nombre de mots, ces systèmes parviennent à prédire des scores globaux avec une précision comparable à

celle des évaluateurs humains. Cependant, leur incapacité à juger des compétences de haut niveau, comme la cohérence ou la précision de la réponse, limite leur usage à une utilisation complémentaire à l'évaluation humaine. Par ailleurs, les mesures effectuées ne sont pas directement exploitables dans un contexte formatif car elles ne ciblent pas spécifiquement les phénomènes qui perturbent la communication.

Dans le cadre de l'évaluation formative, les systèmes automatiques se concentrent plutôt sur des scores basés sur la reconnaissance automatique de la parole, de type *Goodness of Pronunciation*, qui mesure une proximité à un modèle natif. L'évaluation porte majoritairement sur de la parole lue ou répétée, délaissant la parole spontanée, pourtant essentielle au développement de la compétence communicative.

Un décalage important persiste entre les objectifs pédagogiques actuels et les approches adoptées par les systèmes automatiques. Si l'on attend du locuteur qu'il soit intelligible et facilement compris par l'auditeur, les systèmes d'évaluation formative restent encore beaucoup centrés sur la comparaison avec un modèle natif. Cette divergence souligne la nécessité de proposer des outils formatifs mieux alignés sur les objectifs pédagogiques, afin qu'ils puissent apporter un complément efficace aux enseignements classiques.

Chapitre 2

Intelligibilité & compréhensibilité

Dans le chapitre précédent, l'intelligibilité et la compréhensibilité ont été identifiées comme deux notions centrales dans l'évaluation de la prononciation en contexte certificatif. Bien qu'omniprésentes dans les descripteurs de niveaux de la production orale, ces notions restent souvent implicites, laissant une large part à l'interprétation des évaluateurs.

Ce chapitre a pour objectif de préciser ces concepts en proposant des définitions claires et en explorant des outils et méthodologies permettant de les mesurer. En nous appuyant sur une revue approfondie de la littérature, nous mettrons en lumière les principaux facteurs linguistiques et cognitifs qui influencent ces évaluations.

Dans un premier temps, nous établirons des définitions opérationnelles d'intelligibilité et de compréhensibilité. Nous examinerons ensuite les approches méthodologiques pour leur mesure, avant d'analyser les interactions entre caractéristiques du locuteur et attentes de l'auditeur. Ce parcours vise à identifier des phénomènes linguistiques récurrents susceptibles de guider le développement d'outils d'évaluation automatique de la prononciation en L2.

2.1 Définitions

> "*Accent is about difference,*
> *comprehensibility is about the listener effort,*
> *and intelligibility is the end result*"
>
> (Derwing & Munro, 2009, p. 480)

Si l'apprentissage de la prononciation s'est historiquement focalisé sur l'imitation d'un modèle natif (Piccardo, 2016), les recherches récentes mettent davantage l'accent sur la capacité à se faire comprendre et à communiquer de manière claire (Conseil de l'Europe, 2018 ; Isaacs et al., 2018 ; Walker et al., 2021). Ce changement de perspective se reflète également dans les grilles d'évaluation des principaux tests certificatifs d'anglais, comme nous l'avons observé dans le chapitre précédent. Deux notions clés émergent dans ce contexte : l'intelligibilité et la compréhensibilité.

Cependant, si le changement de paradigme est bien établi, la définition précise de ces termes reste sujette à débat. Selon les disciplines, les auteurs et les périodes, ces notions ont parfois des significations distinctes, sont parfois utilisées de manière interchangeable, voire considérées comme synonymes. Nous ne ferons pas ici une analyse approfondie des débats terminologiques (Frost, 2023 ; Levis, 2018 ; Pommée et al., 2022), car ils ne font pas l'objet de cette thèse. Nous nous contenterons d'adopter les définitions largement reconnues dans le domaine de l'acquisition des langues étrangères, autour desquelles un consensus s'est formé depuis une trentaine d'années.

L'intelligibilité est définie par Munro et Derwing (1995) comme "*the extent to which a speaker's message is actually understood*" (p. 76), c'est-à-dire ce qui est effectivement compris du message par l'auditeur. La compréhensibilité, quant à elle, fait référence à l'effort perçu par l'auditeur pour comprendre ce message ("*perceived ease of understanding*"). Derwing et Munro (1997) la définissent comme "*judgments on a rating scale of how difficult or easy an utterance is to understand*" (p. 2).

À ces deux notions s'oppose celle d'« accent étranger » *(foreign accent)*, défini par Alazard (2013) comme « l'écart de prononciation commis par les apprenants vis-à-vis de la norme de prononciation attendue et partagée par les natifs d'une langue donnée » (p. 35), ou plus généralement la notion d'« accent » *(accent, accentedness)* qui renvoie à la perception d'une distance en termes de prononciation vis-à-vis de la norme attendue par l'auditeur.

Autrement dit, l'accent représente une perception de différence, tandis que l'intelligibilité et la compréhensibilité se concentrent sur la compréhension du message.

C'est là que réside le changement de paradigme observé ces dernières années dans le domaine de l'enseignement des langues étrangères. Ce n'est plus la différence avec ce que l'on considère comme modèle qui importe, mais l'efficacité de la communication : ce que l'interlocuteur comprend, et l'effort qu'il doit faire pour comprendre. Par ailleurs, comme l'indiquent Derwing et Munro (2015), "*Oral communication is at minimum a two-person enterprise, in which speaker and listener have equal responsibility for ensuring a successful outcome.*" (p. 388). La compréhension est donc le plus souvent une co-construction par le locuteur et l'auditeur, qui interagissent de manière à mener à bien la communication.

2.2 Évaluation

Si l'intelligibilité et la compréhensibilité sont deux notions étroitement liées, elles restent pourtant conceptuellement et méthodologiquement distinctes. La première concerne le résultat de la communication, et peut donc en théorie être évaluée en termes de précision dans la mesure où l'on dispose de l'énoncé de référence. La seconde, en revanche, est plutôt considérée comme un ressenti, une expérience subjective de l'auditeur qui ne peut donc être évaluée autrement que par le témoignage de celui-ci. Dans cette section, nous présentons différentes méthodes employées pour évaluer l'intelligibilité et la compréhensibilité d'un énoncé.

2.2.1 Évaluation de l'intelligibilité

L'intelligibilité peut être mesurée au niveau local ou global. Au niveau local, il s'agit typiquement de la reconnaissance de mots isolés. C'est, par exemple, la méthode employée par Field (2005), qui cherche à savoir si des mots isolés restent reconnaissables lorsque la position de l'accent lexical est modifiée. Cette approche a le mérite, selon l'auteur, de limiter l'influence du contexte pour la reconnaissance des mots, et ainsi mieux appréhender l'impact de la variable étudiée. Elle reste toutefois une méthode assez artificielle pour évaluer l'intelligibilité d'un locuteur, puisque la parole est justement décontextualisée. Ou et al. (2012) comparent les mesures d'intelligibilité de mots isolés ou en contexte, et obtiennent en effet des scores d'intelligibilité bien meilleurs lorsque les mots sont en contexte que lorsqu'ils sont isolés (12 % d'erreur de reconnaissance contre 43 %).

Au niveau global, la méthode la plus courante est de faire transcrire un énoncé entier, puis de compter le nombre de mots correctement reconnus. Cette méthode a l'avantage d'être simple à mettre en place et présente une bonne fiabilité inter-évaluateur (Derwing & Munro, 2015), mais tous les mots sont considérés au même

niveau. Or, certains mots sont plus importants que d'autres pour la compréhension. Une méthode alternative consiste alors à donner à l'auditeur un énoncé pré-transcrit, et le laisser compléter seulement certains mots cibles en fonction de ce qu'il comprend. La contrepartie ici, c'est que la pré-transcription peut fournir des indices à l'auditeur, rendant l'énoncé potentiellement plus intelligible qu'il ne le serait sans transcription. Pour éviter de recourir à l'écrit, Bernstein (2003) et Minematsu et al. (2011) proposent de « transcrire à l'oral » des énoncés courts en les restituant immédiatement et mot pour mot après les avoir écoutés. Le niveau d'intelligibilité de l'énoncé est alors estimé à partir du nombre de mots correctement restitués par l'auditeur. L'inconvénient de cette méthode est que les énoncés doivent rester assez courts pour être retenus en mémoire par l'auditeur.

Ces méthodes de transcription restent une évaluation de bas niveau : elles permettent difficilement d'estimer le niveau de compréhension réelle du message par l'auditeur. Pour combler ce manque, il est courant de poser des questions de compréhension, ou de demander à l'auditeur de résumer les propos du locuteur, mais l'interprétation des résultats est moins évidente que pour les tâches de transcription, et plus difficilement quantifiable. Selon Derwing et Munro (2015), si chacune de ces méthodes permet d'évaluer l'intelligibilité du locuteur, aucune ne permet de couvrir tous les aspects de la compréhension, et il est souvent nécessaire de les combiner pour estimer l'intelligibilité du locuteur de manière plus juste. Par ailleurs, une limite commune à toutes ces méthodes est qu'il est nécessaire de connaître l'énoncé de référence, c'est-à-dire ce que le locuteur a l'intention de transmettre, quelle que soit la façon dont il le fait, ce qui rend l'évaluation difficile en parole spontanée.

2.2.2 Évaluation de la compréhensibilité

La compréhensibilité du locuteur est, quant à elle, généralement exprimée par un jugement global, souvent au moyen d'une échelle de Likert, typiquement de 1 (très facile à comprendre) à 9 (très difficile à comprendre, Thomson, 2017). Ces jugements holistiques scalaires montrent en général une bonne fiabilité inter-évaluateur : les auditeurs tendent à s'accorder sur le niveau de compréhensibilité des énoncés, bien que les raisons données pour expliquer ces jugements puissent varier selon les évaluateurs (Trofimovich et al., 2024).

Mais sur quoi repose notre jugement lorsqu'on évalue la compréhensibilité à partir d'un jugement global porté en fin d'énoncé ? Dans quelle mesure ce jugement n'est-il pas biaisé par une première impression formée au début de l'écoute ou, au contraire, par les dernières secondes ? Reflète-t-il véritablement notre perception tout au long de l'énoncé, ou est-il une approximation influencée par des éléments spécifiques ?

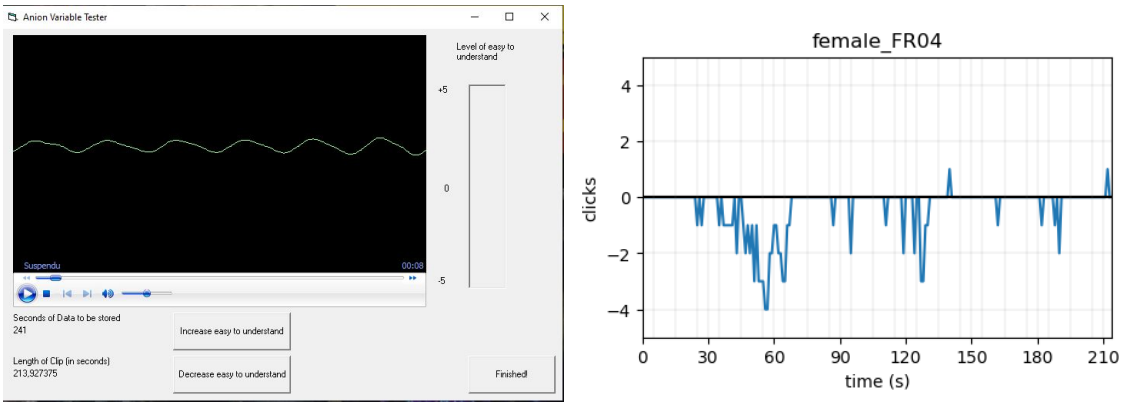


FIG. 2.1 : Interface d'évaluation dynamique de la compréhensibilité, logiciel Idiodynamic (MacIntyre, 2012) utilisé ici par Frost et al. (2024), et visualisation des résultats à droite

Afin de mieux comprendre comment se forme et évolue le jugement au cours de la conversation, Nagle et al. (2019) proposent d'évaluer la compréhensibilité non plus après l'écoute mais de manière dynamique et continue pendant celle-ci. Ils ont demandé à des auditeurs hispanophones natifs d'évaluer la compréhensibilité de trois locuteurs d'espagnol L2. Les stimuli sont des enregistrements d'environ trois minutes, dans lesquels trois locuteurs anglophones natifs s'expriment de manière spontanée au sujet de leur matière préférée à l'université et d'un souvenir d'enfance marquant. Pendant l'écoute, les évaluateurs utilisent des boutons « + » et « - » pour ajuster en temps réel un score de compréhensibilité sur une échelle allant de +5 à -5 (*cf.* figure 2.1). Les résultats ont révélé une grande variabilité dans les stratégies adoptées par les auditeurs. Sur les 24 participants, 18 ont montré une activité limitée, n'ajustant que rarement leur évaluation et préférant souvent attendre la fin de l'énoncé pour donner leur jugement. En revanche, six évaluateurs ont fréquemment ajusté le score en temps réel, mais avec des comportements contrastés : certains ne modifiaient la note qu'à ±1, tandis que d'autres n'exploitaient que la partie positive de l'échelle. Ces résultats suggèrent que les jugements de compréhensibilité sont fortement influencés par les préférences et stratégies des évaluateurs, et que leur comparaison est difficile d'un évaluateur à l'autre. Toutefois, ils mettent également en évidence l'impact tangible de phénomènes micro-structuraux au niveau de la parole du locuteur sur la perception de la compréhensibilité au cours de l'écoute.

2.2.3 (Auto-)évaluation par les participants

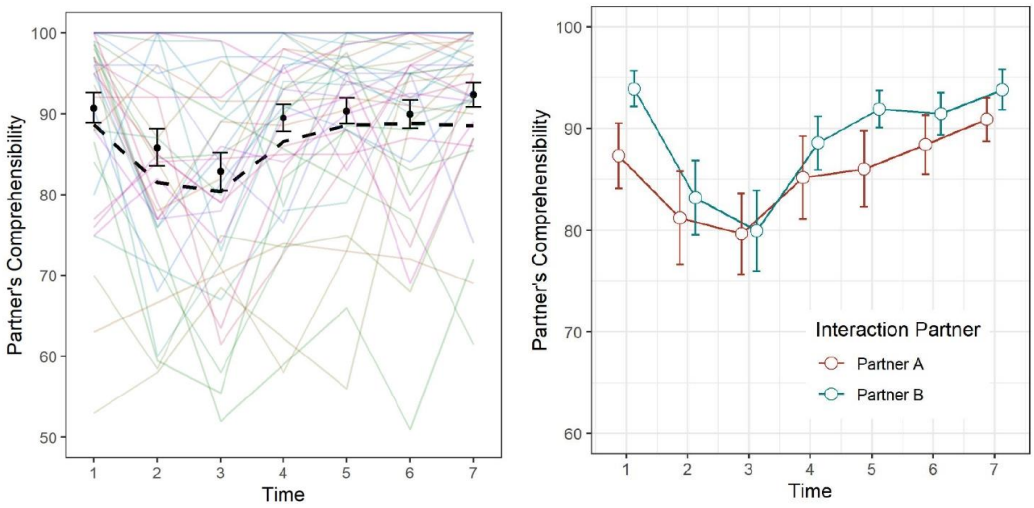
Une critique souvent faite à ces protocoles d'évaluation de la compréhensibilité, qu'ils soient statiques ou dynamiques, est qu'ils font généralement intervenir des évaluateurs extérieurs à l'interaction. Or, comme défini dans la section précédente, la

compréhensibilité est, au même titre que l'intelligibilité, une co-construction issue de l'interaction entre le locuteur et l'auditeur. Elle n'a de sens que pour l'interlocuteur à qui est destiné l'énoncé et qui est engagé dans la conversation. Trofimovich et al. (2020) proposent donc d'examiner l'évolution de la compréhensibilité mutuelle entre des locuteurs L2 au cours d'une interaction. Les chercheurs ont formé 20 paires d'étudiants universitaires anglophones non-natifs de langues maternelles variées. Ces paires ont participé à trois tâches[1] collaboratives et interactives pendant 17 minutes, évaluant leur propre compréhensibilité et celle de leur partenaire à intervalles réguliers de 2 à 3 minutes à l'aide d'un curseur de 0 à 100. Sept évaluations par locuteur ont ainsi été recueillies.

Les chercheurs ont analysé les données à l'aide de modèles mixtes, en incluant des variables telles que les scores IELTS des locuteurs, la diversité lexicale de l'énoncé, ou encore le moment où intervient chaque évaluation. Les commentaires des participants recueillis lors d'entretiens ont également été analysés thématiquement pour éclairer les tendances observées dans les évaluations de compréhensibilité. Les résultats de l'étude montrent que les évaluations suivent une courbe en forme de U, la compréhensibilité étant initialement perçue comme élevée, puis diminuant en raison de la complexité de la tâche, avant d'augmenter à nouveau pour atteindre des niveaux élevés à la fin de l'interaction (*cf.* figure 2.2a). Cette évolution dynamique est indépendante des compétences des locuteurs en matière de production lexicale, et du niveau d'expression orale ou de compréhension orale. De plus, les évaluations des participants de chaque groupe ont tendance à converger au fil du temps, ce qui suggère un alignement entre les interlocuteurs (*cf.* figure 2.2b). L'analyse des entretiens révèle que les changements dans la perception de la compréhensibilité sont souvent attribués à une diminution de l'anxiété, à une augmentation de la confiance, à une meilleure collaboration et à une meilleure connaissance du partenaire.

L'auto-évaluation de la compréhensibilité, c'est-à-dire le jugement que le locuteur porte sur sa propre compréhensibilité, a fait l'objet de peu d'études (Nagle et al., 2022). Trofimovich et al. (2016) observent une corrélation très faible ($r = 0{,}18$) entre les auto-évaluations de 134 locuteurs L2 et les évaluations de trois évaluateurs experts. La plupart des locuteurs ont tendance soit à sous-estimer, soit au contraire à surestimer leur niveau de compréhensibilité. Les auteurs remarquent par ailleurs que les locuteurs qui ont tendance à se surestimer sont ceux dont la prononciation est jugée moins bonne par les évaluateurs, autant sur le plan segmental que prosodique. Isbell et Lee (2022) répliquent cette étude et observent une corrélation un peu plus forte

[1] La première tâche durait 3 min et consistait en une activité brise-glace, où les participants devaient trouver trois caractéristiques qu'ils avaient en commun. La deuxième tâche durait 7 min et consistait à raconter une histoire à partir d'un jeu de cartes réparties entre les participants. Dans la dernière activité, les participants devaient proposer des solutions à des difficultés rencontrées par des étudiants internationaux, pendant 7 min également.



(a) Trajectoire de la compréhensibilité du partenaire estimée par le modèle (ligne pointillée) et trajectoires individuelles observées (lignes continues). Les points pleins indiquent la moyenne du groupe et les barres d'erreur englobent l'intervalle de confiance à 95 % (p. 18)

(b) Moyenne de la compréhensibilité pour les deux locuteurs de chaque paire au cours des sept épisodes d'évaluation. Les barres verticales englobent les intervalles de confiance à 95 % autour des valeurs moyennes. Les désignations des locuteurs A et B sont aléatoires au sein de chaque paire (p. 20)

Fig. 2.2 : *Mesures de compréhensibilité du partenaire observées par Trofimovich et al. (2020), les unités de l'axe temporel représentent chaque temps d'évaluation (avec 1, 4 et 7 correspondant respectivement à la fin des trois activités)*

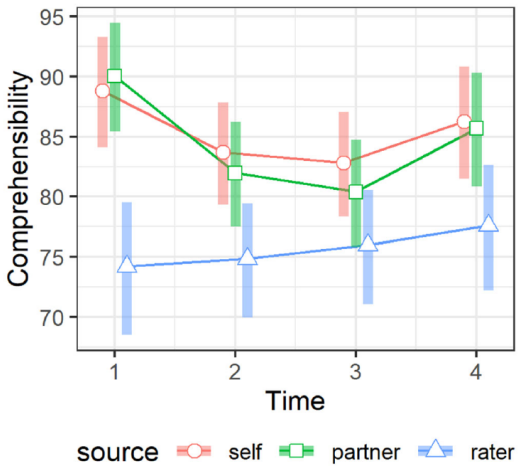


Fig. 2.3 : *Niveau de compréhensibilité estimé par le modèle, des points de vue du locuteur (rouge), de l'interlocuteur (vert) et de l'évaluateur externe (bleu). Les quatre points en abscisse représentent les quatre temps d'évaluation pour les tâches 2 et 3 superposées (une au début, une à la fin et deux au milieu). Les barres verticales représentent les intervalles de confiance à 95 % (Nagle et al., 2022, p. 9)*

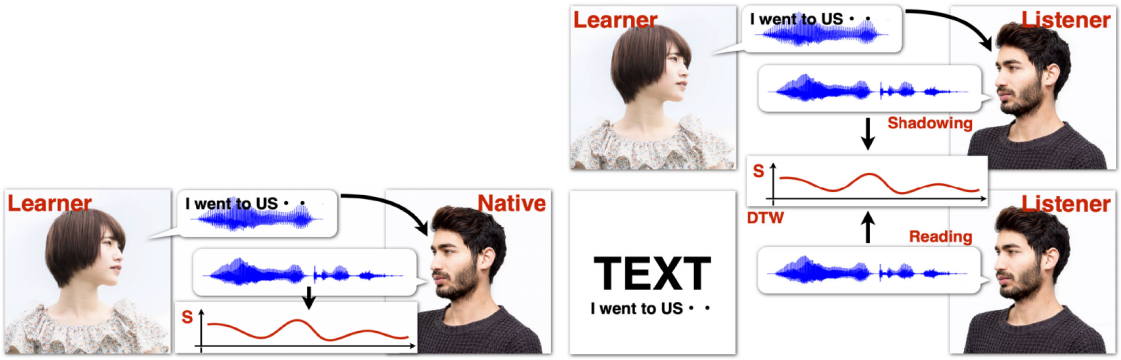
($r = 0{,}54$) entre l'auto-évaluation de 198 locuteurs de coréen L2 et celles d'auditeurs coréanophones natifs. Cette fois-ci, les locuteurs qui ont tendance à se surestimer sont ceux qui ont un niveau plus avancé, et une meilleure estime d'eux-mêmes en termes de prononciation. La justesse de l'auto-évaluation de la compréhensibilité semble aussi pouvoir s'améliorer avec l'expérience : en adoptant une approche longitudinale, Saito et al. (2020) constatent que l'auto-évaluation de leurs participants, 106 locuteurs japonophones de l'anglais L2, tend à s'aligner avec celle de cinq évaluateurs experts après 6 mois de formation sur la production orale et l'importance de la compréhensibilité. Le même constat est fait par Tsunemoto et al. (2022) avec des locuteurs japonophones du français L2, après 15 semaines de travail comportant de fréquentes évaluations entre pairs.

Dans leur article "*Comprehensible to Whom? Examining Rater, Speaker, and Interlocutor Perspectives on Comprehensibility in an Interactive Context*", Nagle et al. (2022) explorent les différences de jugement selon qu'ils sont réalisés par le locuteur lui-même, l'interlocuteur ou un évaluateur externe. Ils font évaluer par 20 étudiants en linguistique les enregistrements audio des conversations de Trofimovich et al. (2020). Comme pour l'expérience initiale, l'évaluation est effectuée sur sept temps, de manière à capturer la dynamique du jugement de compréhensibilité. Les auteurs observent que le niveau de compréhensibilité du point de vue de l'auditeur externe est systématiquement plus bas et ne semble pas aligné avec ceux des points de vue du locuteur et de l'interlocuteur (*cf.* figure 2.3).

2.2.4 Approche par *shadowing*

Une autre approche originale est proposée par une équipe de recherche de l'université de Tōkyō. Ils proposent de mesurer la compréhensibilité non plus par un jugement explicite et subjectif d'un auditeur, mais en analysant la fluidité avec laquelle celui-ci est capable de répéter, en temps réel, l'énoncé du locuteur. Cette répétition en temps réel, ou *shadowing*, consiste à imiter un modèle par répétition la plus simultanée possible. Bien qu'habituellement utilisée par les apprenants pour imiter un modèle natif dans le cadre d'exercices de prononciation, cette méthode est ici détournée : c'est l'auditeur natif qui doit répéter ce qu'il comprend, sans chercher à imiter l'accent du locuteur (*cf.* figure 2.4a). Inoue et al. (2018) postulent que les disfluences et les décalages temporels observés lors du shadowing reflètent les difficultés de traitement de l'auditeur et permettent de mesurer objectivement ces difficultés, sans avoir recours à un jugement explicite. Pour tester cette hypothèse, ils ont demandé à 27 auditeurs natifs du japonais d'effectuer un shadowing d'énoncés lus par six apprenants vietnamophones. Les auteurs analysent alors l'alignement temporel d'une part, et qualitatif d'autre part, des phonèmes entre la version des shadowers natifs et l'enregistrement initial des apprenants. Les scores obtenus sont alors comparés avec les jugements globaux de compréhensibilité et de fluidité effectués par les mêmes auditeurs. Les résultats montrent une corrélation de respectivement $r = -0{,}58$ et $r = -0{,}68$ entre le délai d'alignement temporel et les scores de compréhensibilité d'une part, et de fluidité d'autre part. Lin et al. (2019, 2020) reproduisent l'expérience, mais au lieu de comparer directement le shadow de l'auditeur avec l'énoncé du locuteur, ils le comparent à une lecture « naturelle » de l'énoncé par l'auditeur natif (*cf.* figure 2.4b). Cette démarche permet de comparer deux enregistrements du même auditeur en se concentrant sur les disfluences de répétition, et ainsi mieux appréhender les mécanismes qui sous-tendent la compréhension.

En réalité, ces chercheurs proposent un changement de point de vue. On ne mesure plus la compréhensibilité d'un locuteur, mais plutôt la compétence de compréhension des auditeurs, ou ce qu'ils appellent la « disfluence d'écoute » *(listening disfluency, LD)*. Et cette LD est estimée à partir des disfluences observées pendant le *shadowing*. Dernièrement, la même équipe de chercheurs du laboratoire de l'université de Tōkyō a proposé un nouveau type de diagramme permettant de visualiser à la fois la compréhensibilité relative d'un locuteur, et sa capacité à comprendre les autres locuteurs d'un groupe (Tomita et al., 2024). Ils l'appellent le diagramme de *communicabilité (communicability chart)*, dont un exemple est donné figure 2.5a. Le locuteur se situe au centre du diagramme, ici il s'agit du locuteur n°6 du groupe A. Le demi-cercle rouge permet de visualiser la compréhensibilité du locuteur A6 du point de vue des autres locuteurs du groupe, ou plus exactement leur LD vis-à-vis de A6. Le demi-cercle bleu permet quant à lui de visualiser la compréhensibilité des autres



(a) Comparaison directe entre le shadow de l'auditeur et l'énoncé de l'apprenant dans Inoue et al. (2018)

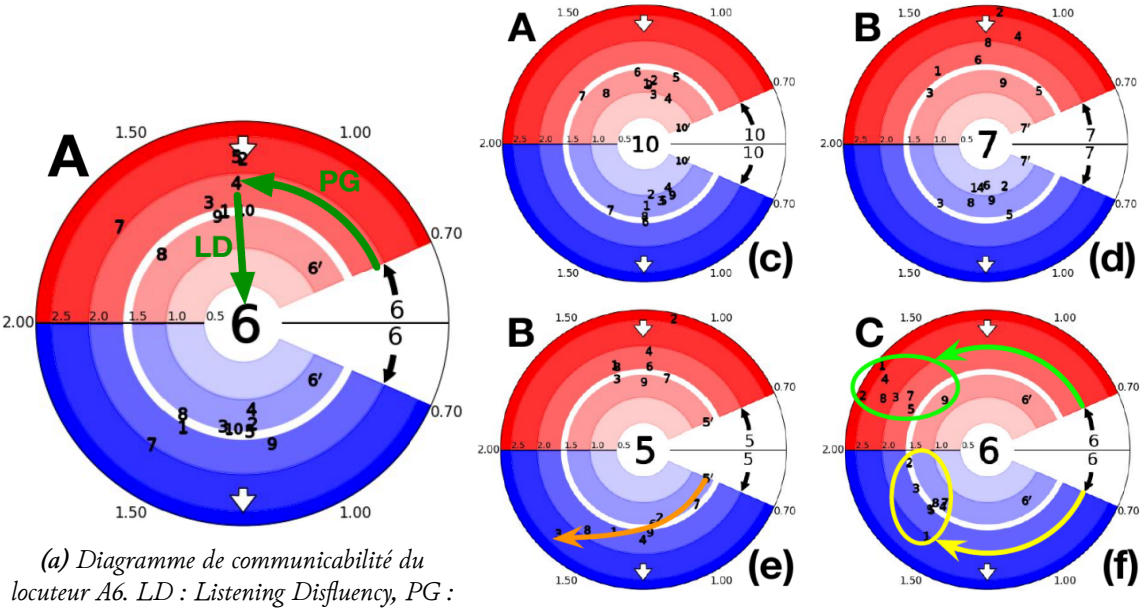
(b) Comparaison entre le shadow de l'auditeur et sa lecture du même énoncé dans Lin et al. (2019)

FIG. 2.4 : Shadowing d'un apprenant par un auditeur natif (Illustrations issues de Lin et al., 2020, p. 1-2)

locuteurs vis-à-vis de A6, ou plus exactement sa LD vis-à-vis de chacun des autres locuteurs. Les chiffres noirs sur le diagramme représentent les différents locuteurs du groupe. Plus ils sont situés proches du centre, plus A6 est compréhensible de leur point de vue (côté rouge), ou plus ils sont compréhensibles du point de vue de A6 (côté bleu). La ligne blanche indique un seuil en dessous duquel les auteurs considèrent la LD comme satisfaisante. Enfin, plus les locuteurs sont situés sur la gauche du cercle, plus leur prononciation est jugée différente de A6[2]. L'étude analyse le profil de 28 locuteurs de 11 langues maternelles différentes, tous anglophones L2. La figure 2.5b présente quatre locuteurs aux profils différents : le locuteur A10 est facilement compris par tout le monde et comprend bien tout le monde (profil idéal) ; B7 comprend aussi tous les autres locuteurs de son groupe mais reste lui-même difficile à comprendre par la plupart d'entre eux ; B5 comprend les locuteurs dont la prononciation est proche de la sienne, mais plus leur prononciation diffère, plus il a des difficultés ; C6, enfin, présente le cas typique d'un locuteur dont la prononciation est éloignée de celle de l'ensemble des autres locuteurs, et reste globalement difficile à comprendre, lui-même comprenant difficilement les autres locuteurs de son groupe.[3]

[2] La prononciation des locuteurs est caractérisée par une mesure de *Goodness of Pronunciation (GOP)*, basée sur l'identification des phonèmes prononcés (Inoue et al., 2018 ; Tomita et al., 2024).

[3] Une présentation des travaux de l'équipe autour du concept de *listening disfluencies* est disponible à l'adresse suivante : https://sites.google.com/g.ecc.u-tokyo.ac.jp/listening-disfluency/



(a) Diagramme de communicabilité du locuteur A6. LD : Listening Disfluency, PG : Pronunciation Gap

(b) Diagrammes de locuteurs aux profils différents

FIG. 2.5 : Diagrammes de communicabilité proposés par Tomita et al. (2024, p. 4027)

2.3 Facteurs d'impact

L'identification des facteurs influençant les jugements d'intelligibilité et de compréhensibilité a fait l'objet de nombreuses recherches ces 30 dernières années. Les études ont révélé des résultats variés, parfois contradictoires, en raison de la complexité intrinsèque du processus de compréhension. En effet, celle-ci apparaît influencée par une multitude de paramètres, dépendant non seulement des caractéristiques linguistiques et discursives du locuteur, mais aussi du profil de l'auditeur et du contexte de communication. Cette diversité reflète le caractère multifactoriel des jugements d'intelligibilité et de compréhensibilité, lesquels sont en partie façonnés par les biais cognitifs des évaluateurs et les contraintes de la tâche d'évaluation.

Deux approches principales ont été utilisées pour étudier ces facteurs :

- Une approche qualitative, basée sur des entretiens avec les évaluateurs, permettant de recueillir des données introspectives sur les éléments perçus comme déterminants dans leurs jugements.
- Une approche quantitative, plus répandue, reposant sur l'analyse de corrélations entre des phénomènes linguistiques spécifiques et les jugements globaux des auditeurs.

Les études proposant une approche qualitative, comme celles d'Isaacs et Thomson (2013), Nagle et al. (2019) ou Frost et al. (2024), invitent les évaluateurs à expliciter les raisons de leurs jugements à travers des entretiens d'auto-confrontation *(stimulated recall)*. Cette méthode permet d'identifier les éléments jugés influents par les auditeurs, et d'obtenir des explications parfois très détaillées sur les processus qui ont amené les participants à formuler leur jugement. En contrepartie, les commentaires des évaluateurs reflètent avant tout ce dont ils ont conscience, ce dont ils se souviennent ou parviennent à expliciter, et ils peuvent être influencés par les représentations qu'ils se font de la parole des apprenants ou des locuteurs de telle ou telle langue.

L'approche quantitative repose quant à elle sur des analyses systématiques. Elle consiste généralement à mesurer la corrélation entre les jugements subjectifs des évaluateurs et certains phénomènes, linguistiques ou non, ciblés par les auteurs. Une méta-analyse réalisée par Saito (2021) synthétise les résultats de 37 études portant sur la perception de compréhensibilité et d'accent en anglais L2. Elle fait état de nombreux facteurs influençant ces jugements, mais met également en lumière des divergences importantes entre les conclusions des différentes études. Outre la complexité de la notion de compréhension, ces contradictions reflètent aussi les différences méthodologiques et contextuelles entre les études, rendant la tâche de recensement compliquée.

Sur la base d'une revue approfondie des recherches dans le domaine de l'acquisition L2, nous proposons une synthèse structurée des différents facteurs identifiés dans la littérature comme influençant les jugements d'intelligibilité et de compréhensibilité, et en les illustrant par les résultats observés dans différentes études.

2.3.1 Facteurs liés au locuteur

Commençons par les facteurs intrinsèques au locuteur. Ces facteurs concernent évidemment une dimension linguistique, mais également méta-linguistique, moins souvent considérée.

Facteurs linguistiques

Fluidité de la parole Quand Nagle et al. (2019) demandent à leurs participants d'expliciter leurs jugements pendant l'entretien d'auto-confrontation qui suit l'évaluation dynamique de la compréhensibilité, la fluidité des locuteurs ressort des témoignages comme une des causes premières d'augmentation du score. Une parole perçue comme

fluide semble donc associée à une meilleure compréhensibilité. Toutefois, les évaluateurs restent assez vagues sur les éléments linguistiques qui sous-tendent cette perception de fluidité.

Dans les études qui ont adopté une approche plus quantitative, plusieurs paramètres sont considérés comme relevant de la fluidité. Dans une méta-analyse pour examiner la relation entre la fluidité de l'énoncé *(utterance fluency)* et la fluidité perçue *(perceived fluency)*, Suzuki et al. (2021) recensent les mesures utilisées dans plusieurs études qui traitent du sujet. La mesure la plus commune est le débit de parole, généralement représenté par le nombre de mots ou de syllabes par unité de temps, suivi du débit d'articulation (débit de parole en excluant les pauses), la longueur moyenne des énoncés entre pauses, la fréquence des pauses, leur durée moyenne, ou encore le nombre de répétitions, de faux départs, ou d'auto-corrections, que les auteurs regroupent dans un « taux de disfluence » *(disfluency rate)*. À partir d'une modélisation à effets aléatoires, les auteurs obtiennent une estimation globale de la corrélation entre chaque paramètre et le jugement subjectif de fluidité. Les paramètres qui apparaissent les plus corrélés avec le jugement de fluidité sont le débit de parole ($r = 0{,}76$) et la longueur moyenne des énoncés entre pauses ($r = 0{,}72$). Viennent ensuite le débit d'articulation ($r = 0{,}62$), la fréquence des pauses ($r = -0{,}59$), leur durée moyenne ($r = -0{,}46$), et, dans une moindre mesure seulement, le taux de disfluences ($r = -0{,}20$). Les auteurs ont également examiné l'influence de certains facteurs modérateurs sur la relation entre la fluidité de l'énoncé et la fluidité perçue. Ils ont constaté que la force de la corrélation variait en fonction de facteurs tels que la L1 des locuteurs, le type de tâche de production, la durée des extraits de parole et le profil des auditeurs (natifs ou non-natifs).

En ce qui concerne la perception de compréhensibilité, le paysage est similaire : un énoncé a tendance à être perçu comme plus compréhensible quand son débit est plus rapide (Huensch & Nagle, 2021 ; Munro & Derwing, 2001 ; Saito et al., 2015), que les énoncés entre pauses sont plus longs (Suzuki & Kormos, 2023), qu'il y a généralement moins de pauses et que celles-ci sont plus courtes (Suzuki & Kormos, 2020). Par ailleurs, les pauses situées à l'intérieur des unités syntaxiques ont tendance à nuire davantage à la compréhensibilité que les pauses situées entre celles-ci (Suzuki & Kormos, 2020).

Précision phonologique La précision phonologique joue un rôle central dans la compréhensibilité. Elle concerne l'articulation des consonnes et des voyelles, mais également leur accentuation. Les erreurs phonologiques peuvent réduire la compréhensibilité, voire l'intelligibilité du message, notamment lorsqu'elles touchent des contrastes phonémiques dits à « haut rendement fonctionnel » ("*high functional load phenemic contrasts*" , Catford, 1987). Ces contrastes font référence aux paires minimales de phonèmes qui permettent de distinguer une grande quantité de mots. Ainsi, en anglais,

les contrastes /i - a/, /i - ɪ/, /k - h/, /p - b/ ont un rendement plus important que les contrastes /ɪ - e/, /aː - ɜː/, /b - v/ ou /f - θ/. Les erreurs phonologiques impliquant des contrastes à haut rendement se sont révélées effectivement plus délétères sur la compréhensibilité (Isaacs & Trofimovich, 2012 ; Munro & Derwing, 2006), indiquant que toutes les erreurs n'ont pas la même importance. Les erreurs affectant les voyelles ont des répercussions particulièrement importantes, notamment sur la perception de l'accent lexical, dans le cas de l'anglais. Field (2005) montre qu'une simple modification de la fréquence fondamentale d'une voyelle peut altérer la reconnaissance des mots, et ce chez les auditeurs natifs comme non-natifs. Tajima et al. (1997) observent que la durée de la voyelle joue également un rôle important pour la compréhension. Hahn (2004) constate que les auditeurs se souviennent mieux des informations contenues dans un message quand celui-ci présente une accentuation lexicale correcte. La précision de l'accent lexical s'avère être, en outre, un facteur important quel que soit le niveau de compétence en langue du locuteur, au moins chez les locuteurs francophones de l'anglais (Isaacs & Trofimovich, 2012).

Ressources linguistiques La variété et la précision du vocabulaire, ou encore l'exactitude grammaticale de l'énoncé, ont également une influence sur la perception de la compréhensibilité (Crowther et al., 2017 ; Trofimovich et al., 2024). Plus spécifiquement, l'utilisation du lexique semble avoir un impact plus important chez les locuteurs de niveau débutant, tandis que la compréhensibilité est plus influencée par la précision grammaticale chez les niveaux avancés (Isaacs & Trofimovich, 2012).

Structuration du discours La manière dont le locuteur organise son discours, utilise des connecteurs logiques et maintient la cohésion générale contribue également à la compréhensibilité du message, et a également un impact plus important pour les locuteurs de niveaux avancés (Isaacs & Trofimovich, 2012).

Facteurs paralinguistiques

Les indices visuels, tels que les gestes, les expressions faciales et le regard du locuteur, participent également à sa compréhensibilité. Tsunemoto et al. (2023) montrent par exemple que l'accès à des informations visuelles réduit la sévérité des jugements des auditeurs : les locuteurs sont jugés plus compréhensibles lorsque les auditeurs peuvent voir leur visage, et plus encore si leurs gestes sont visibles.

2.3.2 Facteurs liés à l'auditeur

Bien que la recherche sur la production orale en L2 se concentre souvent sur les caractéristiques des productions d'apprenants, on ne peut pas parler de compréhen-

sibilité sans tenir compte des destinataires du message. Plusieurs études mettent en évidence des facteurs côté auditeur, que nous pouvons synthétiser à travers les points suivants :

Facteurs linguistiques

L'impact de la familiarité de l'auditeur avec la langue cible et la langue source du locuteur suscite des conclusions divergentes dans la littérature. D'un côté, la maîtrise de la langue cible, incluant une connaissance approfondie des variations phonétiques, des structures syntaxiques et du vocabulaire, peut faciliter le décodage et l'interprétation du message malgré la présence d'erreurs ou d'un accent marqué (Derwing & Munro, 2015 ; Trofimovich et al., 2024). De l'autre, une familiarité avec la langue source du locuteur peut également jouer un rôle favorable dans la compréhension (Gass & Varonis, 1984). Par exemple, Minematsu et al. (2003) rapportent que des évaluateurs japonophones, jugeant des locuteurs japonophones de l'anglais, considèrent intelligibles certains énoncés que des anglophones natifs jugent non intelligibles. Cependant, ce bénéfice lié à la connaissance de la langue source ou de la langue cible n'est pas systématique.

Ainsi, Munro et al. (2006) ont étudié les jugements d'intelligibilité, de compréhensibilité et de perception de l'accent en anglais par des locuteurs et des auditeurs de différentes origines linguistiques. Ils concluent que, de manière générale, les jugements sur ces trois dimensions sont remarquablement cohérents, indépendamment de la langue maternelle des auditeurs ou de leur familiarité avec la langue du locuteur ou la langue cible. Cependant, ils notent une exception : les auditeurs japonophones montrent une meilleure compréhension des locuteurs japonophones que les anglophones natifs, alors que ce phénomène n'est pas observé pour les autres couples de langues. Dans une autre étude, Hayes-Harb et al. (2008) observent que l'avantage lié à la connaissance de la langue source, dans cette étude, le mandarin, se manifeste principalement chez les auditeurs et locuteurs ayant un niveau limité dans la langue cible.

Au-delà de la familiarité avec une langue spécifique, l'exposition régulière à divers accents non natifs semble améliorer la capacité des auditeurs à comprendre la parole en L2 (Munro et al., 2012). Cette exposition favorise le développement de stratégies d'écoute et la reconnaissance des caractéristiques phonétiques partagées par différents accents. En ce sens, Kennedy et Trofimovich (2008) montrent que les auditeurs habitués aux accents étrangers jugent la compréhensibilité de manière moins sévère que ceux qui y sont peu exposés.

La formation linguistique et l'expérience d'enseignement des langues sont également souvent citées comme facteurs pouvant influencer le jugement des auditeurs

(Isaacs & Thomson, 2013, 2020). On parle généralement d'« auditeurs experts » *(expert listeners)*. D'après Saito (2021), ce type d'auditeurs a tendance à s'appuyer davantage sur des informations phonologiques (précision segmentale en particulier) que les auditeurs dits « naïfs », qui font plus souvent référence à la fluidité de parole et à des jugements basés sur des intuitions.

Implication de l'auditeur dans l'interaction

Nous avons déjà brièvement présenté l'étude de Nagle et al. (2022), qui s'intéresse à la perception de la compréhensibilité tantôt par le locuteur lui-même, tantôt par son interlocuteur ou encore par un auditeur externe qui ne participe pas à la conversation. Les auteurs constataient que si les jugements du locuteur et de l'interlocuteur étaient bien alignés, celui de l'auditeur extérieur était systématiquement plus bas que celui des participants à la conversation.

Comme mentionné dans la première section de ce chapitre, la compréhensibilité est co-construite par le locuteur et l'auditeur, il s'agit d'un processus dynamique où l'implication de l'auditeur joue donc un rôle évident. Le fait que l'auditeur soit impliqué dans l'interaction, ou au contraire qu'il soit seulement auditeur passif d'un enregistrement de cette interaction, est donc susceptible d'influencer le locuteur, et indirectement d'impacter sa compréhensibilité. En effet, un auditeur impliqué dans l'interaction fournit des indices verbaux et non verbaux sur sa compréhension. Des signaux comme des haussements de sourcils, des clignements ou des gestes peuvent signaler des difficultés de compréhension (Trofimovich et al., 2024). Face à ces indices, le locuteur peut adapter son discours, en clarifiant, en ralentissant ou en répétant (Saito et al., 2022). Cette adaptation contribue *in fine* à une meilleure perception de la compréhensibilité par l'auditeur.

Engagement cognitif de l'auditeur Par ailleurs, un auditeur impliqué est plus susceptible de déployer des efforts cognitifs pour comprendre le message. L'attention accrue aux détails de la parole et l'utilisation de stratégies de compensation pour combler les lacunes de compréhension peuvent faciliter le processus de compréhension et, par conséquent, améliorer la perception de la compréhensibilité (Trofimovich et al., 2024).

Influence des facteurs socio-affectifs La perception de la compréhensibilité est également influencée par des facteurs sociaux et affectifs. Une plus grande implication de l'auditeur peut conduire à une perception plus positive de l'interlocuteur, influençant ainsi l'évaluation de sa compréhensibilité. Par exemple, une perception de collaboration et de faible anxiété chez l'interlocuteur peut mener à une évaluation plus favorable de sa compréhensibilité (Nagle et al., 2022 ; Trofimovich et al., 2024).

Impact des indices visuels L'accès aux indices visuels du locuteur, comme ses expressions faciales, ses gestes et ses postures, influence la perception de la compréhensibilité. Les auditeurs engagés dans une interaction en face à face bénéficient de ces indices, contrairement aux auditeurs qui écoutent des enregistrements audio (Nagle et al., 2022). *cf.* facteurs paralinguistiques de la sous-section précédente.

Facteurs individuels

Des facteurs individuels tels que l'attention, la mémoire de travail et les capacités de traitement auditif peuvent également jouer un rôle dans la perception et l'évaluation de la compréhensibilité (Derwing & Munro, 2015 ; Saito et al., 2022). L'âge des auditeurs est également cité comme facteur d'influence sur la compréhension. Munro et al. (2012) observent par exemple que les auditeurs plus jeunes (8 à 10 ans) ont tendance à moins bien comprendre la parole L2 que les auditeurs plus âgés (14 à 16 ans et adultes), suggérant l'importance d'une certaine maturité cognitive. La motivation de l'auditeur et sa volonté de communiquer *(willingness to communicate)* est aussi susceptible d'impacter le jugement (Nagle et al., 2022).

Un autre facteur individuel parfois mis en avant est celui de la sensibilité musicale. Il semblerait en effet que les auditeurs ayant une sensibilité musicale accrue soient plus sensibles aux variations prosodiques et aux erreurs phonétiques. Isaacs et Trofimovich (2011) constatent en l'occurrence que les auditeurs musiciens ont tendance à être plus sévères que les non-musiciens dans l'évaluation de la compréhensibilité, de la fluidité et de l'accent, mais n'observent un effet significatif que pour ce dernier.

2.3.3 Facteurs liés au contexte

Enfin, certains facteurs ne sont liés ni au locuteur, ni à l'auditeur, mais peuvent avoir un impact observable sur le jugement de compréhensibilité.

Le type de tâche de production orale peut influencer la performance linguistique du locuteur en L2, affectant de fait l'intelligibilité et la compréhensibilité, bien que le contraste soit plus marqué pour la perception de l'accent (Crowther et al., 2017). La familiarité de l'auditeur avec le sujet ayant tendance à faciliter la compréhension (Gass & Varonis, 1984), il est également réaliste de penser qu'elle peut influencer sa perception de compréhensibilité, de même que la présence d'indices contextuels, tels que des images ou des informations préalables sur le sujet.

Le contexte social dans lequel la communication a lieu et les attentes de l'auditeur peuvent également influencer la perception de la compréhensibilité (Crowther et al., 2017 ; Isaacs & Trofimovich, 2012 ; Nagle et al., 2022). Par exemple, un auditeur

peut être plus indulgent envers un locuteur non natif dans un contexte informel qu'en situation d'évaluation formelle.

D'autres facteurs contextuels comme la qualité de la transmission de la parole (téléphone, qualité d'enregistrement, niveau sonore, bruits ambiants etc.) et l'environnement dans lequel se situe l'auditeur (conditions sonores etc.) peuvent également impacter le jugement de l'auditeur.

Conclusion

L'objectif premier d'un locuteur L2 est donc d'être compris par son auditeur, et que cette compréhension soit la moins difficile possible pour celui-ci. L'intelligibilité et la compréhensibilité sont intrinsèquement interactionnelles. Elles dépendent non seulement des caractéristiques du locuteur (sa prononciation, la cohérence du discours, etc.), mais aussi des caractéristiques de l'auditeur, de ses attentes et de son investissement dans la conversation. En ce sens, elles ne peuvent être évaluées indépendamment de ce dernier : sans auditeur, il n'y a ni intelligibilité ni compréhensibilité. Évaluer la production orale d'un locuteur reviendrait donc, idéalement, à évaluer la compréhension de l'auditeur qui l'écoute.

En outre, la compréhension est un processus dynamique et continu. Elle varie selon les conditions de l'interaction et évolue au fil de la conversation. Son caractère personnel et dynamique implique qu'elle ne possède pas de valeur absolue. La compréhensibilité est estimée par l'auditeur, et est par conséquent sujette à de nombreux facteurs de variation lorsque celui-ci sur-estime ou sous-estime l'effort qu'il pense devoir faire pour comprendre. Zielinski (2006) parle à juste titre d'"*intelligibility cocktail*" Comme le résument Derwing et Munro (2015) : "*With all this potential variability, one has to ask whether teaching pronunciation is a viable enterprise*" (p. 388). Heureusement, malgré les nombreux facteurs pouvant influencer leurs jugements, il s'avère que les auditeurs évaluent la compréhensibilité de manière consistante, quelle que soit leur langue maternelle, qu'ils soient « experts » ou « naïfs », et qu'ils participent ou non directement à l'interaction. S'il faut retenir une chose, c'est que les éléments linguistiques de la parole du locuteur n'ont pas un « coût » fixe en termes d'intelligibilité et de compréhensibilité : « c'est une histoire de contexte, de perception, d'attente de la part de l'auditeur » (Didelot et al., 2019, p. 10). De plus, si les causes de difficulté de compréhension sont variées, certains éléments sont particulièrement saillants : la fluidité de la parole et la prononciation des voyelles, notamment leur accentuation, qui définit le rythme de la parole.

Facteurs liés au locuteur	Facteurs liés à l'auditeur	Facteurs liés au contexte
• Fluidité de parole – débit de parole et d'articulation – longueur des énoncés entre pauses – fréquence des pauses – durée moyenne des pauses – distribution syntaxique des pauses • Précision phonologique – articulation des consonnes et voyelles à haut rendement fonctionnel – accent lexical • Ressources linguistiques – variété et précision du vocabulaire – précision grammaticale • Structuration du discours • Gestes et expressions faciales	• Facteurs linguistiques – familiarité avec la langue cible – familiarité avec la langue source – familiarité avec la parole L2 – formation linguistique – expérience d'enseignement • Implication dans la conversation – signaux de compréhension – engagement cognitif – accès aux indices visuels du locuteur – facteurs socio-affectifs • Facteurs individuels – attention – mémoire de travail – capacités de traitement auditif – âge – motivation – volonté de communiquer – sensibilité musicale	• Type de tâche de production orale • Familiarité de l'auditeur avec le sujet • Présence d'indices contextuels • Contexte social de la communication • Qualité de transmission de la parole • Environnement de l'auditeur

Tab. 2.1 : *Principaux facteurs influençant la compréhensibilité du locuteur*

Chapitre 3

Rythme & fluence

Nous avons vu dans le chapitre précédent que de nombreux facteurs, côté locuteur comme auditeur, impactent le degré d'effort requis pour comprendre le message. Parmi les facteurs côté locuteur, la fluidité et le rythme de la parole sont deux éléments qui reviennent régulièrement dans les grilles d'évaluation de la production orale (*cf.* chapitre 1). Dans ce chapitre, nous proposons d'approfondir ces deux notions, et de présenter en détail deux phénomènes linguistiques qui y sont étroitement liés : les pauses et l'accent lexical.

3.1 Définitions

Avant de rentrer dans le vif du sujet, il nous paraît important de faire un point terminologique sur ces deux termes qui reviennent souvent dans la littérature, mais pour lesquels les définitions varient parfois selon les auteurs.

Commençons par le terme le plus fréquent – la fluidité. Si le terme *fluency* en anglais fait souvent référence au niveau global de compétence en langue étrangère, nous nous intéressons ici à sa définition restreinte, plus commune dans le domaine de l'enseignement/apprentissage des langues étrangères, et qui concerne spécifiquement la production de parole (*speech fluency* ou *oral fluency*). Cette fluidité, ou *fluence*, est souvent interprétée comme le niveau d'automatisation et de contrôle du locuteur sur les processus cognitifs impliqués dans la planification et la production de la parole (Thomson, 2015). Segalowitz (2010) distingue trois types de fluences : la fluence cognitive *(cognitive fluency)*, la fluence de phrase *(utterance fluency)* et la fluence perçue *(perceived fluency)*. La première correspond à la fluidité des processus cognitifs en amont de la production, la seconde correspond à la fluidité de la parole produite,

la troisième enfin à la perception de fluidité du point de vue de l'auditeur. Lickley (2015) reprend les mêmes catégories mais appelle les deux premières respectivement fluence de planification *(planning fluency)* et fluence de surface *(surface fluency)*. Les trois catégories sont étroitement liées, mais une disfluence dans l'une n'entraîne pas nécessairement une disfluence dans les autres. La plus importante pour la réussite de la communication est la troisième, mais l'évaluer de manière systématique n'est possible que sur la deuxième, tandis qu'y remédier n'est envisageable qu'en agissant sur la première.

C'est le plus souvent la fluence perçue qui est évaluée, de manière intuitive et holistique par des auditeurs, mais certains tentent également de mesurer directement la fluence de surface, à partir d'analyses acoustiques sur le signal de parole. Les critères les plus souvent utilisés dans ce cas sont le débit de parole ou d'articulation (nombre de syllabes par seconde ou par minute, avec ou sans pauses), le ratio de phonation (durée de parole sans pause divisée par durée totale), le nombre moyen de mots ou de syllabes par segment entre pauses, le nombre de pauses par seconde ou par minute ou encore leur durée moyenne (Thomson, 2015). On constate que la présence ou l'absence de pauses et leur durée semblent être fortement liées à la notion de fluence. En effet, selon Derwing et Munro (2015), la fluence se caractérise principalement par la proportion de pauses ou d'autres marqueurs de disfluence, tels que les faux départs ou les répétitions. Ainsi, on a souvent tendance à considérer les pauses et autres interruptions du flux de parole comme des disfluences – ou « dysfluences » en contexte pathologique (Kernou, 2022) –, mais nous allons voir que les pauses sont loin d'être nécessairement problématiques.

Le deuxième terme qui nous intéresse est celui de rythme. En linguistique, la notion de rythme de la parole *(speech rhythm)* fait référence à la façon dont se succèdent des éléments forts et des éléments faibles le long d'un axe temporel. Gibbon et Gut (2001) le définissent par exemple comme la récurrence de patterns temporels perceptibles de valeurs plus ou moins marquées d'un paramètre à travers le temps. Di Cristo et Hirst (1997) le définissent plus simplement comme l'organisation temporelle des proéminences. La notion de rythme est souvent associée à l'accentuation, notamment dans le cas de langues comme l'anglais, historiquement considérée comme prototype des langues dites « isoaccentuelles » *(stress-timed languages)*. En effet, les premières études sur le rythme de la parole proposent l'hypothèse selon laquelle les langues peuvent être classées selon des caractéristiques de durée de syllabes (James, 1929 ; Pike, 1945). On y oppose alors l'anglais et le français – le premier étant caractérisé par une succession régulière de syllabes accentuées séparées par un nombre variable de syllabes plus courtes, et le second caractérisé par une succession régulière de syllabes de durées équivalentes (Abercrombie, 1967). La plupart des langues sont alors associées soit à l'une, soit à l'autre catégorie. Cette théorie, dite de l'isochronie accentuelle ou syllabique, a toutefois largement été débattue et remise en question par la

suite, et des variantes moins dichotomiques ont été proposées, comme le continuum de rythmicité de Dauer (1987) ou l'approche scalaire de Bertinetto (1989). D'autres approches de caractérisation du rythme voient également le jour, comme la distinction de Wenk et Wioland (1982) entre langues à accent en tête de groupe rythmique, dites capochrones *(leader-timed languages)*, et celles dont l'accent se situe plutôt en fin de groupe, dites codachrones *(trailer-timed languages)*, toujours représentées par l'anglais et le français. Toutes ces approches considèrent l'accent comme un facteur clé dans la caractérisation du rythme.

3.2 Les pauses

On appelle communément « pauses » les interruptions ponctuelles du flux de parole du locuteur. Ces interruptions sont le résultat complexe d'un compromis entre des contraintes physiologiques, linguistiques et culturelles, et n'ont pas toutes le même impact sur l'auditeur. Contrairement à la ponctuation dans un texte écrit, les pauses n'interviennent pas nécessairement pour structurer l'énoncé ; et leur position – bien que contrainte – est plus variable et semble dépendre de nombreux facteurs. Les interruptions du flux de parole peuvent être acoustiques (silences), ou linguistiques (allongements, interjections, mots de remplissage etc.), elles peuvent être physiquement présentes (pauses objectives) ou parfois seulement perçues par l'auditeur (pauses subjectives), et peuvent plus ou moins l'aider ou le perturber dans la compréhension du message.

Nous tenterons dans un premier temps de lister les différents types de pauses recensés, ainsi que leurs rôles. Nous décrirons ensuite les caractéristiques physiques de ces pauses ainsi que les contraintes syntaxiques auxquelles elles sont soumises, avant de nous intéresser à leur impact sur la perception de fluence et de compréhensibilité.

3.2.1 Types et rôles des pauses

Il existe probablement autant de typologies de pauses que d'auteurs ayant écrit à leur sujet. Certains les catégorisent selon leurs fonctions, d'autres selon leurs caractéristiques physiques, d'autres encore selon leur impact sur l'auditeur. Di Cristo (2013) identifie six types de pauses : les pauses respiratoires, les pauses structurales, les pauses pragmatiques, les pauses d'hésitation, les pauses aléatoires et les pauses phonostylistiques. Si les pauses respiratoires sont a priori issues de contraintes de bas niveau, elles ont toutefois tendance à éviter de perturber la cohérence grammaticale et sémantique du discours – une pause respiratoire ne peut donc pas survenir n'importe où dans l'énoncé et sera contrainte par sa structure syntaxique. On peut regrouper

les autres types de pauses en deux catégories : les pauses volontaires et les pauses involontaires. Les pauses structurales, qui ont pour objectif de délimiter les groupes syntaxiques de l'énoncé, et les pauses pragmatiques, qui ont un rôle plutôt rhétorique, sont en principe plutôt volontaires et planifiées par le locuteur. Les pauses d'hésitation, engendrées par la recherche lexicale ou la planification du discours, et les pauses aléatoires, causées par des troubles du langage, sont a priori plutôt involontaires et viendront potentiellement perturber la compréhension du discours. Enfin, les pauses phonostylistiques caractérisent le style de parole (i.e. discours politique) ou celui du locuteur ; elles sont plus ou moins volontaires, et peuvent être plus ou moins perturbantes. Candea (2000) propose une classification binaire plutôt tournée vers l'impact sur l'auditeur : elle oppose les pauses structurantes, à fonction de segmentation de la parole, aux pauses non-structurantes, à fonction d'hésitation. Dodane et Hirsch (2018) considèrent quant à eux les pauses en contexte de conversation, et distinguent d'abord les pauses inter-tours, pour la gestion du dialogue, et les pauses intra-tours, comprenant des pauses tantôt dues à des mécanismes physiologiques (déglutition, respiration), tantôt à l'organisation structurelle du discours (délimitation des unités de sens, mise en relief d'informations), ou enfin à la planification de l'énoncé (recherche lexicale, élaboration mentale). En outre, Grosjean et Deschamps (1975) suggèrent qu'une même pause peut porter plusieurs fonctions différentes en même temps, comme profiter d'une frontière syntaxique ou d'une hésitation pour respirer ou pour reformuler, il est donc important de ne pas lui attribuer un type exclusif.

Les pauses inter-tours interviennent comme leur nom l'indique entre les tours de parole des locuteurs. Elles sont souvent rapidement écartées des analyses, soit parce que les corpus analysés sont des monologues, soit parce qu'on ne les considère pas comme relevant de la fluence du locuteur. Or, il arrive que des pauses intra-tours soient utilisées par l'interlocuteur comme une opportunité de prendre la parole, et il s'avère que leur utilisation est contrainte par de nombreux facteurs linguistiques et culturels. Selon Fox et al. (1996), en anglais, les auditeurs sont capables de prédire avec précision quand un énoncé en construction va se terminer. Ils peuvent ainsi planifier leur énoncé et prendre leur tour de parole précisément à un moment de fin possible *(possible completion point)* sans laisser de pause entre les deux tours de parole. D'après Fox et al. (1996) et Sacks (1992), du point de vue d'un anglophone natif, les pauses sont souvent considérées comme un moment de malaise qui perturbe la conversation, et qui incite l'interlocuteur à prendre la parole. De ce fait, de nombreuses stratégies existent pour éviter de se faire prendre la parole, généralement en remplissant tout moment de silence possible ("*eh*" , "*yeah*" , "*well*" , "*you know*" etc., Sacks, 1992). Suivant ce raisonnement, une pause vide en anglais peut avoir tendance à être considérée comme un manque de contrôle sur la conversation par le locuteur.

Il en va autrement en japonais. Selon Shigemitsu (2007), la syntaxe du japonais permet difficilement de prédire la fin de l'énoncé ; c'est une pause un peu plus longue

que les autres à la fin de celui-ci qui indique à l'interlocuteur qu'il peut prendre la parole. En outre, les locuteurs japonophones ont tendance à séparer par de brèves pauses des segments de mots assez courts entre lesquels les interlocuteurs ont la possibilité de réagir. Ces courtes réactions de l'interlocuteur, qui ne provoquent pas de changement de tour de parole, sont appelées *backchannel*, ou 相槌 *aizuchi* en japonais. Elles permettent d'indiquer au locuteur qu'il est compris ou au moins écouté. Ce phénomène est particulièrement fréquent en japonais, mais existe aussi dans une moindre mesure en anglais et en français (White, 1989). Le recourt au backchannel est donc simplifié (ou encouragé) par la présence de pauses entre petits segments de parole (2,36 mots en moyenne en japonais selon Maynard, 1989). Ces pauses permettent ainsi au locuteur de s'assurer que l'interlocuteur comprend le message, car celui-ci aura tendance à ne pas demander explicitement de clarification lorsqu'il ne comprend pas. Elles ont donc, à l'opposé de l'anglais, un rôle d'encouragement du locuteur à poursuivre son discours.

Par ailleurs, la dynamique des tours de parole en japonais est fortement influencée par la relation sociale entre les locuteurs : l'un des participants de la conversation détient généralement le contrôle de la dynamique de conversation – il a le *speakership* – et la prise de parole inattendue de l'un des autres participants peut provoquer un malaise. On parle aussi de *pauses de politesse*, qui sont attendues de la part de certains locuteurs vis-à-vis de certains autres, et interprétées de manière différente en fonction du statut conversationnel du locuteur (Shigemitsu, 2007). Une prise de parole en japonais est donc à la fois régie par la syntaxe, mais également par les contraintes sociales entre les locuteurs.

Shigemitsu (2007) s'intéresse à l'effet que peut avoir l'utilisation de stratégies pausales culturellement différentes dans une conversation entre des locuteurs de langue maternelle différente. Elle analyse 4 conversations spontanées d'une trentaine de minutes en anglais et en japonais, entre 2 ou 4 locuteurs qui ne se connaissent pas. Dans chacune d'elles, la moitié des participants sont de langue maternelle japonaise, l'autre moitié de langue maternelle anglaise. Chaque conversation est suivie d'un entretien individuel avec les locuteurs, pour leur demander ce qu'ils ont ressenti pendant la conversation et s'ils se sont sentis à l'aise ou non. Seules les pauses silencieuses (interruption de phonation) sont considérées dans cette étude. Shigemitsu observe que l'utilisation de stratégies pausales japonaises en anglais, ou anglaises en japonais, peut considérablement impacter la réussite de la conversation. Dans les conversations en anglais qualifiées de moins réussies par les participants, elle observe que les pauses sont rares et très courtes, empêchant les participants japonophones d'y placer une réaction, ou trop courtes pour qu'ils la considèrent comme un moment de prise de parole potentielle. Les participants anglophones ont eu tendance à remplir chaque moment de silence, jusqu'à ceux des locuteurs japonophones, qui l'ont souvent interprété comme une coupure de parole. Par ailleurs, si les anglophones considéraient

important que tout le monde parle autant, certains locuteurs japonais se satisfaisaient de participer sans pour autant prendre la parole, et sans sentir de gêne vis-à-vis de cela. À l'inverse, les locuteurs anglophones ont perçu les participants japonais comme peu coopératifs et parfois impolis par leur manque de conversation et d'initiative de prise de parole, résultant pour certains en un sentiment de culpabilité de ne pas leur laisser le temps de parler. L'utilisation adéquate des pauses est donc clé pour mener à bien une conversation.

Les pauses peuvent ainsi avoir des causes et des objectifs variés. Par ailleurs, une pause peut avoir un objectif précis souhaité par le locuteur, comme vérifier que l'interlocuteur comprend, mais être interprétée différemment par ce dernier, comme un manque d'intérêt dans la conversation ou une invitation à prendre la parole.

3.2.2 Caractéristiques physiques

Plusieurs phénomènes dans la parole du locuteur peuvent être interprétés par l'auditeur comme des pauses. Le premier et le plus évident est l'interruption de la phonation, ou l'arrêt temporaire de production de parole. On parle dans ce cas de « pause silencieuse », et ce sont elles qui sont le plus largement analysées dans la littérature. La plupart des études s'accordent à fixer un seuil minimum de durée à partir duquel considérer une interruption de phonation comme une pause, mais la valeur de ce seuil est très variable d'une étude à l'autre, comme le montre le tableau 3.1. La revue d'une quarantaine d'études analysant les phénomènes de pauses dans la parole non pathologique nous montre que ce seuil varie entre 0 ms et 3 s, avec un grand nombre d'entre elles le fixant entre 100 ms et 300 ms. L'ensemble des études présentées ici traitent des phénomènes d'hésitation, de distribution des pauses ou d'évaluation de la fluence en parole native ou L2, dans différentes langues.

de Jong et Bosker (2013) constatent qu'un seuil de 250 ms à 300 ms obtient la meilleure corrélation avec le niveau de compétence en langue des locuteurs non-natifs en néerlandais (déterminé par un test de vocabulaire), amenant de nombreuses études à fixer un seuil à 250 ms par la suite. Une autre justification souvent donnée pour ne pas considérer les silences inférieurs à 200 ms est le fait que les pauses plus courtes sont moins à même de refléter les difficultés linguistiques de construction du discours, mais semblent plutôt liées à des contraintes coarticulatoires de bas niveau, qui ne sont généralement pas le sujet d'intérêt de ces études. En effet, les études qui considèrent des silences très courts ajoutent souvent un délai supplémentaire devant les consonnes occlusives (50 ms pour Fauth et Trouvain, 2018 ; Smiljanić et Bradlow, 2005), ou complètent la détection automatique des silences par une annotation manuelle (Matzinger et al., 2020).

Seuil minimum	Sources
Pas de seuil	Fauth et Trouvain, 2018 ; Maclay et Osgood, 1959 ; Wilkes et Kennedy, 1969
1 ms	Matzinger et al., 2020
5 ms	Owoicho et al., 2024 ; Smiljanić et Bradlow, 2005
20 ms	Cucchiarini et al., 2000 ; Kirsner et al., 2005
60 ms	Campione et Véronis, 2002
80 ms	Levin et al., 1967
100 ms	Butcher, 1981 ; Kang et Johnson, 2018 ; Lounsbury, 1954 ; Trouvain, 2004
200 ms	Candea, 2000 ; Cucchiarini et al., 2002 ; Fletcher, 1987 ; Goldman-Eisler, 1968 ; Grosjean, 1980 ; Kahng, 2014 ; Lennon, 1990 ; Zellner, 1994
250 ms	de Jong, 2016 ; de Jong et Bosker, 2013 ; Grosjean et Deschamps, 1975 ; Kahng, 2018 ; Kallio et al., 2022 ; Shea et Leonard, 2019 ; Suzuki et al., 2021 ; Witton-Davies, 2018
300 ms	Grosjean et Deschamps, 1972 ; Lacheret-Dujour et Victorri, 2002
400 ms	Tavakoli, 2010
1 s.	Lay et Paivio, 1969 ; Levin et Silverman, 1965
2 s.	Siegman et Feldstein, 1979
3 s.	Siegman et Pope, 1966

Tab. 3.1 : *Seuils de durée minimum de pause utilisés dans la littérature (L1/L2)*

Campione et Véronis (2002) mettent en garde sur le fait que le choix du seuil minimal de durée peut largement impacter les conclusions des analyses qui suivent. Ils observent notamment que la durée moyenne des pauses est plus courte en parole spontanée qu'en parole lue si on ne définit aucun seuil, mais qu'elle est plus longue si on ne considère que les pauses supérieures à 200 ms, et qu'elle est égale si on ajoute un seuil maximum à 2 s. Il devient ainsi pratiquement impossible de comparer les résultats obtenus par différentes études, si celles-ci choisissent des seuils différents. À travers une analyse de corpus de lecture de textes en anglais, français, allemand, italien et espagnol, Compione et Véronis observent que la distribution des durées de pauses suit une distribution logarithmique multimodale et non une loi arithmétique normale comme il est couramment admis jusqu'alors. Ils identifient deux gaussiennes autour de 150 ms et 500 ms, quelle que soit la langue. Ces deux gaussiennes sont observées dans des études ultérieures et semblent relativement stables. Kirsner et al. (2005) vont jusqu'à faire l'hypothèse que la première catégorie (pauses courtes, 50 ms à 70 ms) est due aux processus d'articulation, tandis que la deuxième (pauses longues, 500 ms à 700 ms) l'est plutôt à la structuration du discours. Dans un corpus similaire en anglais, français, italien, espagnol, roumain et néerlandais, Demol et al. (2007) identifie également ces deux gaussiennes et constatent qu'elles ne sont liées ni à la langue ni au débit de parole. Enfin, Goldman et al. (2010) analysent un corpus de 40 min de français de différentes situations de communication (lecture, narration conversationnelle, journal télévisé et conférence scientifique), et constatent que le nombre de gaussiennes fluctue entre 1 et 3 en fonction des situations, mais étant majoritairement bimodal.

Du côté de la parole spontanée, Campione et Véronis (2002) observent toujours deux gaussiennes (autour de 80 et 430 ms), accompagnées d'une troisième autour de 1500 ms. Leur corpus est constitué d'entretiens d'une quinzaine de minutes avec 10 locuteurs, issus du Corpus Français Oral de Référence. Les auteurs en viennent à proposer la catégorisation des durées de pauses suivante : pauses brèves ($<$200 ms), pauses moyennes (entre 200 ms et 1 s) et pauses longues ($>$1 s). Cette catégorisation sera souvent citée par la suite, mais semble peu utilisée dans les faits – la plupart des auteurs préférant fixer un seuil unique.

Grosman et al. (2018) identifient également 2 gaussiennes dans la distribution des durées de pauses du corpus LOCAS-F[1], toutefois, ils remarquent que cette bimodalité ne se retrouve pas nécessairement dans toutes les situations de parole et pour tous les locuteurs : le journal radiophonique et les conférences scientifiques semblent relativement standardisés avec une distribution bimodale similaire pour tous les locu-

[1] *LOuvain Corpus of Annotated Speech-French* (L. Martin et al., 2014). Durée : 3 h38 min, 76 locuteurs belges, français et suisses, en situation de monologue, dialogue, ou multilogue. 14 situations de communication différentes comprenant conférences scientifiques, débats et discours politiques et académiques, interactions formelles et informelles, interviews, journaux radiophoniques, lectures radiophoniques.

teurs ; celles-ci sont plus hétérogènes en discours politique et présentent une bimodalité pour 3 locuteurs sur 5, tandis que les récits conversationnels et les homélies sont plutôt unimodaux. Quant à la durée médiane des pauses, elle varie de 289 à 518 ms selon les situations mais également largement à l'intérieur de celles-ci. Les auteurs conseillent de considérer la distribution des durées de pause en fonction des situations de parole, voire en fonction des locuteurs. Ils préconisent de ne pas définir de seuil de durée fixe, et seulement exclure les mesures aberrantes. Ainsi, certaines études choisissent de ne pas fixer de seuil mais plutôt de se fier à la perception d'annotateurs humains, amenant parfois tout de même à des pauses inférieures à 100 ms, comme Fauth et Trouvain (2018).

Zellner (1994) montre que le seuil de perception des pauses varie en fonction du segment précédent, et Duez (1993) constate que certaines pauses sont perçues même sans interruption de phonation – elle les appelle « pauses subjectives ». Enfin, d'autres études encore font état d'un seuil relatif au débit de parole du locuteur, variant entre 180 et 250 ms chez Duez (1982, 1991) (calculé à partir de la durée moyenne des occlusives intervocaliques), de 98 à 490 ms chez Kirsner et al. (2003) (calculé à partir de la distribution des durées de pauses par locuteur), ou de 138 à 384 ms chez de Jong et Bosker (2013) (calculé à partir du débit d'articulation de chaque enregistrement).

Les pauses ne se limitent toutefois pas aux phénomènes d'interruption de phonation. On parle de « pauses pleines » lorsqu'il n'y a pas d'interruption de phonation (allongements, « heu » et autres interjections. Certains auteurs, comme Fauth et Trouvain (2018), considèrent même les faux-départs, les répétitions ou les reformulations comme pauses pleines : tout ce qui, en somme, interrompt le flux du discours.

3.2.3 Mesures automatiques des pauses

Shea et Leonard (2019) font une revue approfondie des mesures relatives aux pauses utilisées pour l'évaluation de la parole L2. La plupart des études mesurent des fréquences générales de pauses : nombre de pauses par minute, par mot, par syllabe, par proposition ou par énoncé (généralement défini comme une proposition principale avec ses relatives), ou encore par tour de parole. La durée des pauses est généralement considérée à travers des ratios de durée totale de pause par rapport au temps de parole, ou à l'inverse, la durée de phonation par rapport au temps de parole. Les mesures qui prennent en compte la position des pauses sont plus rares, et considèrent généralement celles-ci vis-à-vis de la frontière des propositions (*mid-clause vs. end-of-clause*, Kahng, 2018 ; Suzuki et Kormos, 2020), ou plus largement de l'énoncé (de Jong, 2016). Il peut s'agir de fréquence de pause par type (par exemple, inter- ou intra-proposition, Kallio et al., 2022), d'une durée moyenne, ou encore du nombre d'énoncés suivis d'une pause par exemple.

3.2.4 Pauses et localisation syntaxique

De nombreuses études montrent que la fréquence et la durée des pauses sont corrélées avec la position de celles-ci dans l'énoncé, et en particulier avec le type de frontière syntaxique où elles se trouvent. Lorsque la position d'une pause est inattendue, on parle souvent de pause non-structurante, de pause agrammaticale ou encore de pause disfluente (Fauth & Trouvain, 2018).

Tauberer (2008) utilise les informations de catégories grammaticales des mots et la structure syntaxique de l'énoncé pour prédire la position et la durée des pauses en anglais spontané dans le corpus de conversations téléphoniques Switchboard. Il observe que les pauses ont tendance à apparaître autour des conjonctions, des compléments, ou avant les pronoms ou les sujets. En revanche, elles sont beaucoup plus rares après les sujets, entre les verbes et les syntagmes prépositionnels, ou entre les prépositions et les syntagmes nominaux. Tauberer teste différentes combinaisons entre 12 paramètres[2] pour obtenir la meilleure prédiction. D'après ses résultats, l'analyse structurale par constituants a un plus grand pouvoir prédictif que l'analyse lexicale seule, mais la simple information de durée du constituant précédent combinée au nombre de mots du constituant suivant prédit à peu près aussi bien la position et la durée de la pause que l'ensemble des paramètres combinés (F-score de 78,2% contre 78,5% avec tous les paramètres).

Cao et Chen (2019) s'intéressent quant à eux aux caractéristiques de la parole de locuteurs qu'ils appellent "*successful speakers*", pour identifier les critères qui participent à la compréhensibilité de la parole. Ils analysent des enregistrements de discours politiques, de Ted Talks, ou des vidéos à succès sur les réseaux sociaux chez 15 locuteurs anglophones natifs et non natifs. Le premier critère rapporté par les auteurs est le placement des pauses dans le discours. Ils constatent que les pauses sont souvent placées avant les conjonctions de subordination (exemple : *« we must never forget // **that** those heroes // **who** fought against evil // also fought for // the nations // **that** they loved »*, p. 2050), et plus généralement à la frontière syntaxique entre deux propositions (*« if it is not available in your area // you can also use ham instead »*, p. 2050), et ce sans différence perceptible entre les locuteurs natifs et non natifs.

Dans une analyse de la position des pauses et des marqueurs d'hésitation dans des récits produits par des élèves de 4ème en classe de français, Candea (2000) catégorise les pauses en « structurantes » (lorsqu'elles sont non immédiatement précédées par un marqueur d'hésitation) et « non-structurantes » (lorsqu'elles sont immédiatement

[2] Catégorie du mot précédent, du mot suivant, et combinaison des deux ; catégorie du constituant le plus grand se terminant, se commençant, et combinaison des deux ; nombre de mots et durée du constituant le plus grand se terminant, et commençant ; profondeur syntaxique ; et temps de fin du mot précédent calculé depuis le début de l'énoncé et relatif sa longueur totale.

précédées par un marqueur d'hésitation). Selon sa définition, elle note que 82,5% des pauses sont structurantes. Parmi elles, 78% sont placées en fin d'énoncé ou de proposition syntaxique, tandis que 19% seulement se trouvent en fin de syntagme (qu'elle appelle constituant syntaxique), et 3% à l'intérieur d'un syntagme. Dans un corpus plus long et diversifié en situations de parole (LOCAS-F), Grosman et al. (2018) font des observations similaires : 78% des pauses sont structurantes (selon la même définition que Candea). Toutes pauses confondues, 36% d'entre elles sont en fin de proposition (qu'ils appellent unité de rection, constituée d'un verbe accompagné de ses dépendants), 11% entre ce qu'ils appellent séquences syntaxiques, ou unités syntaxiques intermédiaires, et 9% à l'intérieur des groupes accentuables, leurs unités syntaxiques minimales, qui correspondent à la combinaison d'un mot lexical et des mots grammaticaux qui en dépendent, soit une unité légèrement plus petite que le syntagme. Les 44% restants se situent entre des groupes accentuables. D'après leurs observations, la majorité des pauses surviennent entre les unités syntaxiques, et rarement à l'intérieur des groupes accentuables (ci-après GA). Les auteurs observent également que la parole spontanée est caractérisée par plus de pauses intra-GA, mais aussi plus de pauses entre les unités syntaxiques maximales[3]. Les pauses inter-séquence syntaxique semblent quant à elles plus fréquentes en parole préparée.

En ce qui concerne la durée des pauses, Candea (2000) observe que les pauses sont significativement plus longues en fin d'énoncé, qu'en fin de proposition, et qu'en fin de syntagme. C'est également ce que constatent Grosman et al. (2018) : plus la frontière syntaxique est importante, plus la pause est longue, quelle que soit la situation de parole. Ajoutons que la durée des pauses peut également être influencée par la longueur des constituants la précédant ou la suivant, pour des raisons physiologiques ou de planification du discours (Krivokapić, 2007).

Observe-t-on une différence de distribution des pauses selon les langues ? Grosjean et Deschamps (1975) comparent la distribution des pauses en français et en anglais dans des interviews radiophoniques. Ils fixent un seuil de durée minimale de pause à 250 ms, et considèrent plusieurs positions de pauses possibles : en fin de proposition (qu'ils appellent « phrases », combinant un syntagme nominal (SN) et un syntagme verbal (SV), éventuellement accompagné de compléments), entre ou à l'intérieur des syntagmes. D'après leurs observations, les locuteurs français ont tendance à faire plus de pauses en fin de proposition (60%) que les locuteurs anglais (55 %, $p < 0{,}05$), mais surtout moins de pauses à l'intérieur d'un syntagme SN ou SV (16 % contre 26 %, $p < 0{,}001$). La différence se joue surtout au niveau du SV, où les anglophones font 14 % plus de pauses que les francophones, tandis qu'ils en font 5 % moins à l'intérieur du SN. De plus, les anglophones semblent répartir les pauses plus librement à l'inté-

[3] Les auteurs expliquent cette observation par le fait que les propositions sont plus courtes en parole spontanée.

rieur du SV, avec une préférence devant le complément prépositionnel (45%), alors que les francophones les placent majoritairement entre le verbe et son objet (70%). Il semble donc y avoir des différences de tendance dans la distribution des pauses en français et en anglais, du moins en parole radiophonique, dans les années 70.

Qu'en est-il maintenant pour les locuteurs non-natifs ? Dans une analyse de distribution des pauses en parole lue en français, Fauth et Trouvain (2018) observent que les lecteurs non-natifs ont tendance à faire plus de pauses à l'intérieur des énoncés que les lecteurs natifs, et les débutants en font davantage que les apprenants de niveau avancé. Le premier groupe est constitué de 20 lecteurs germanophones lisant à haute voix un texte en français des Trois Petits Cochons issu du corpus IFCASL (Trouvain et al., 2016). Dix d'entre eux ont un niveau A2-B1, les dix autres un niveau B2-C1. Dix autres locuteurs francophones natifs sont également enregistrés pour comparaison. Les auteurs constatent par ailleurs que les lecteurs non-natifs font plus de pauses et des pauses plus longues en général, et plus encore pour les lecteurs débutants, mais sans toutefois observer de différence significative entre les niveaux.

de Jong (2016) observe également que les locuteurs non-natifs anglophones et turcophones ont tendance à faire plus de pauses à l'intérieur des énoncés[4] que les locuteurs natifs en néerlandais. Elle observe aussi une corrélation avec le niveau du locuteur : plus celui-ci a un niveau élevé, moins il fait de pauses intra-énoncé. En outre, la fréquence des pauses entre les énoncés semble indépendante de la langue maternelle du locuteur et de son niveau de compétence.

La position des pauses est donc fortement contrainte par la syntaxe de l'énoncé. De manière générale, elles ont tendance à être situées aux frontières de haut niveau syntaxique, typiquement entre les propositions, mais plus rarement à l'intérieur de groupes syntaxiques plus petits comme les syntagmes verbaux ou nominaux. Par ailleurs, moins la frontière syntaxique est élevée, moins la pause éventuelle est longue. Du côté des locuteurs L2, et notamment chez les apprenants de niveau débutant, il apparaît que les pauses en frontières de bas niveau sont plus nombreuses et plus longues. Mais toutes les pauses sont-elles perçues de la même manière par l'auditeur ?

3.2.5 Perception des pauses

Les pauses sont perçues différemment selon leur position et leur nature. Duez (1985) montre par exemple que les pauses en français sont mieux perçues lorsqu'elles sont situées entre deux propositions, qu'à l'intérieur de l'une d'elles. Cette observation est également confirmée par Collard (2009) et Lickley (1995). Candea (2000) et

[4] de Jong (2016) les appelle des « unités de paroles » *(speech units)*, constituée d'une proposition indépendante et de ses subordonnées éventuelles.

Duez (1995) remarquent par ailleurs que les pauses qu'ils catégorisent comme « non-structurantes » (immédiatement précédées d'une hésitation) n'occasionnent presque jamais un changement de tour de parole : elles ne sont pas perçues comme des indices de coupe par les auditeurs. Mieux encore, lorsque J. Martin et Strange (1968) demandent à 129 étudiants anglophones natifs de répéter un énoncé spontané avec ses hésitations, ou de le transcrire avec ses hésitations, ils constatent que les hésitations intra-constituants sont systématiquement déplacées en frontière de constituant. Simon et Christodoulides (2016) proposent une expérience intéressante où ils demandent à des auditeurs naïfs d'annoter en temps réel des échantillons de parole francophone de genres variés, en signalant chaque fois qu'ils perçoivent la fin d'un groupe de mots. Les résultats montrent que la simple complétude syntaxique provoque la perception d'une frontière même sans autre indice acoustique. La syntaxe semble donc jouer un rôle important sur la perception et la tolérance des pauses.

Par ailleurs, si Bard et Lickley (1997) observent que les auditeurs peinent à se souvenir des éléments disfluents dans la parole au profit du contenu du message, ils peuvent aussi avoir tendance à mieux retenir les informations lorsqu'elles sont précédées d'une hésitation (Fox Tree, 2001). Corley et al. (2007) et MacGregor (2008) constatent également que la présence d'une pause (pleine ou silencieuse) à l'intérieur d'un énoncé augmente la probabilité que le locuteur se souvienne du mot qui suit. Lundholm Fors (2015) fait le même constat, et ajoute que les pauses inférieures à 500 ms semblent avoir un meilleur impact que les pauses plus longues.

Les pauses peuvent ainsi avoir un effet positif sur l'auditeur, en structurant l'énoncé ou en augmentant ponctuellement son niveau d'attention et en facilitant la mémorisation du message. De manière générale, les pauses situées en frontière de groupes syntaxiques semblent mieux perçues et acceptées que celles survenant à l'intérieur des groupes.

3.2.6 Pauses et évaluation de la fluence

La majorité des études recourent à une fréquence globale ou une durée moyenne de pauses en général (Kahng, 2018 ; Saito et al., 2022). Ces deux paramètres apparaissent effectivement très corrélés avec le niveau global d'un apprenant, ces derniers ayant tendance à faire plus de pauses et des pauses plus longues quand leur niveau de compétence en langue est moins élevé. Toutefois, comme nous l'avons vu dans les sections précédentes, les pauses ne sont pas nécessairement un problème ; au contraire, lorsqu'elles sont bien placées, les pauses permettent une meilleure compréhensibilité (Cao & Chen, 2019 ; Isaacs et al., 2018). La question reste de savoir quelles pauses sont susceptibles d'être problématiques, et lesquelles le sont moins.

De récentes études se sont penchées sur la relation entre la distribution syntaxique des pauses et la perception de fluence. Kahng (2018), par exemple, recrute une cohorte de 46 évaluateurs et leur fait évaluer 80 extraits de paroles au moyen d'une échelle de Likert à neuf points (1=très disfluent, 9=très fluent). Les évaluateurs sont tous de langue maternelle anglaise et étudiants dans une université aux États-Unis ; les locuteurs sont de langue maternelle coréenne ($n = 37$, 74 extraits) et anglaise ($n = 3$, 6 extraits). Les extraits font environ 20 s et sont issus d'un enregistrement plus long dans lequel le locuteur répond à deux questions, sur sa spécialité à l'université et ses loisirs. En parallèle, tous les silences de plus de 250 ms ont été annotés et catégorisés en fonction de leur position dans l'énoncé : entre ou à l'intérieur des propositions ; le ratio pauses/minute, leur durée moyenne et le débit d'articulation par extrait sont également calculés. La durée moyenne, la fréquence des pauses inter- et intra-proposition sont log-transformées pour approximer une distribution normale. Au moyen d'une régression multiple par étapes, Kahng constate que la fréquence des pauses intra-proposition est le paramètre le plus corrélé avec le jugement de fluence, expliquant à lui seul plus de 54 % de sa variance. Combiné avec la fréquence des pauses inter-proposition, seuls 6 % supplémentaires de la variance sont expliqués, et ni la fréquence, ni la durée moyenne des pauses en général ne sont capable d'améliorer significativement ce modèle[5]. La distribution syntaxique des pauses semble donc jouer un rôle important dans la perception de la fluence.

Dans une seconde expérimentation, Kahng tente de vérifier l'impact des pauses sur le jugement de fluence en modifiant artificiellement une sélection de 24 extraits (L1=anglais) et 24 extraits (L1=coréen). Il propose 3 conditions : condition 1) les pauses sont supprimées (réduites à 150 ms) ; condition 2) à partir de ces extraits sans pauses, 5 pauses inter-proposition de 600 ms sont insérées ; condition 3), à partir de ces extraits sans pauses, 5 pauses intra-proposition de 600 ms sont insérées. Kahng fait alors évaluer les extraits ainsi modifiés selon le même protocole, à 92 locuteurs natifs de l'anglais, en veillant à ce qu'ils n'écoutent pas deux fois le même enregistrement original. En comparant les jugements par condition, il observe que les extraits avec pauses ajoutées sont jugés significativement moins fluents que les extraits sans pauses ($p < 0{,}001$), et que les extraits avec pauses intra-proposition sont jugés significativement moins fluents que les extraits avec pauses inter-proposition ($p = 0{,}048$). Ces observations sont confirmées par d'autres études par la suite. Suzuki et Kormos (2020) font évaluer par 10 locuteurs anglophones natifs des enregistrements produits par 40 locuteurs japonophones de niveau A2 à C1. Il s'agit cette fois de parole argumentative, les locuteurs doivent donner leur avis sur un sujet d'actualité. L'évaluation est faite sur deux dimensions : la compréhensibilité et la fluence du locuteur, toujours

[5]Notons que cela ne signifie pas que la fréquence et la durée moyenne des pauses n'expliquent pas une partie de la variance des jugements de fluence. Kahng note que la fréquence seule explique 31 % de la variance, et la fréquence et la durée moyenne en expliquent 43 %.

via une échelle de 9 points. Parmi de nombreux paramètres couvrant la complexité et la précision de la réponse, la fluence, la prononciation ou la cohérence du discours, les auteurs observent que le débit de parole est le plus corrélé avec le jugement de compréhensibilité, tandis que le jugement de fluence est en particulier influencé par la fréquence de pauses inter-proposition.

Dans une autre étude, Kallio et al. (2022) vont plus loin en étudiant l'impact de la position des pauses au niveau du syntagme dans des extraits de 200 locuteurs non-natifs du finnois. Ils classent les pauses en 5 catégories : inter- et intra-proposition, inter- et intra-syntagme, et intra-mot (pauses intervenant à l'intérieur d'un mot non terminé). Une évaluation de la perception de fluence sur une échelle de 4 points et une évaluation du niveau global sur une échelle de 7 points est effectuée par deux évaluateurs parmi une cohorte de 16 évaluateurs certifiés par l'Agence Nationale de l'Éducation Finnoise. Comme dans les études précédentes, des modèles de régression multiples sont utilisés pour déterminer quels paramètres influencent le plus l'évaluation parmi. Les auteurs observent que la fréquence des pauses intra- et inter-syntagme sont les indicateurs les plus corrélés avec le jugement de niveau global ($t = -4{,}96$ et $t = -4{,}33$, $p < 0{,}001$) et la perception de fluence ($t = -6{,}93$ et $t = -5{,}33$, $p < 0{,}001$). La fréquence des pauses intra-mot est également un indicateur fort, suivi par celle des pauses inter-proposition ($t = 2{,}23$, $p < 0{,}05$ pour la fluence, mais non significatif pour le niveau global).

Dans cette section, nous avons tenté de mettre en évidence la complexité et la diversité des phénomènes de pauses dans la parole. Celles-ci peuvent avoir des causes et des objectifs variés : certaines jouent un rôle structurant en délimitant les groupes syntaxiques, en fournissant au locuteur un temps de planification et à l'auditeur un temps de traitement de l'énoncé. Certaines pauses ont un objectif stylistique, permettant au locuteur de moduler son discours ou d'insister sur un élément ; d'autres encore sont causées par la recherche lexicale ou la planification du discours. En outre, une même pause peut avoir plusieurs causes et plusieurs objectifs, et peut être interprétée de manières différentes selon la culture de l'auditeur et la situation de communication.

La position des pauses apparaît étroitement liée à la syntaxe de l'énoncé. Chez les locuteurs L1, et particulièrement en parole préparée, la majorité des pauses est observée au niveau de frontières syntaxiques de haut niveau, et sont au contraire plus rares à l'intérieur de constituants syntaxiques plus petits comme les syntagmes. Par ailleurs, plus la frontière syntaxique est basse, plus la pause éventuelle qui y survient est courte.

Dans le cas de la parole L2, et notamment chez les débutants, on observe plus de pauses de bas niveau syntaxique. Celles-ci sont moins structurantes et n'ont pas, a

priori, d'effet facilitateur pour la compréhension. On peut supposer que la plupart de ces pauses sont dues à des difficultés de production de la part du locuteur, pour qui la recherche lexicale ou la planification du discours est d'autant plus difficile qu'il ne maîtrise pas la langue.

3.3 L'accent lexical

L'accent, au sens de *stress* en anglais, fait référence au degré de force utilisé pour produire une syllabe (Crystal, 2008). C'est un phénomène relatif, c'est à dire qu'une syllabe pourra être plus ou moins accentuée qu'une autre, mais sa valeur absolue n'a pas réellement d'intérêt. On distingue généralement trois catégories fonctionnelles : l'accent de mot, l'accent de phrase et l'accent contrastif (Frost, 2023). Nous nous concentrerons ici sur la première catégorie. Toutes les langues n'ont pas d'accent de mot *(word stress, word-level stress)* ; parmi celles qui en ont un, certaines ont un accent à position fixe (*fixed stress languages*, comme le finnois, le polonais ou le français, où il se place systématiquement sur la première, la pénultième ou la dernière syllabe, respectivement), d'autres ont un accent à position variable comme l'anglais, l'allemand ou l'espagnol (Cutler & Jesse, 2021). Lorsque la position de l'accent de mot est variable, on parle d'accent lexical *(lexical stress)* car il joue un rôle pour l'accès lexical : certains mots ne se distinguent que par sa position, comme ***differ*** et *de**fer*** en anglais, ou *a**un*** et *a**ún*** en espagnol.

Au delà de son intérêt sémantique, l'accent lexical – et plus généralement l'accent de mot – joue un rôle important pour aider l'auditeur à segmenter le flux de parole (Cutler, 2015).

Au niveau acoustico-phonétique, l'accentuation peut se caractériser par des variations au niveau suprasegmental (fréquence fondamentale (f_0), intensité et durée), mais également au niveau segmental par la qualité vocalique. Toutefois, ces différentes dimensions ne sont pas nécessairement exploitées dans toutes les langues, et leur poids respectif peut varier. Ainsi, en espagnol, seuls les 3 niveaux suprasegmentaux sont en jeu, là où en thaï, la f_0 est réservée pour le ton et ne participe pas à l'accent de mot (Cutler & Jesse, 2021). En anglais ou en allemand, en revanche, les 4 niveaux de variation peuvent être exploités pour marquer la syllabe accentuée (Cutler, 2015). Les 3 niveaux suprasegmentaux apparaissent très corrélés entre eux, et si de nombreuses études ont analysé un seul niveau à la fois, ou encore tenté de hiérarchiser leur importance respective, il semble important de ne pas les dissocier complètement et d'adopter une approche intégrée de l'accentuation (Vaissière, 1983).

Comme pour le phénomène de pauses que nous avons analysé dans la section précédente, il est important de différencier l'accent phonétique (physiquement présent et

mesurable) de l'accent perçu par l'auditeur. Aussi, un mot peut présenter toutes les caractéristiques phonétiques d'une accentuation en finale, et toutefois être perçu par l'auditeur comme accentué en initiale. Cette perception de l'accent est influencée par l'accent phonétique produit par le locuteur, mais aussi par les représentations linguistiques et les attentes de l'auditeur (Cooper et al., 2002 ; Sugahara, 2011 ; van Leyden & van Heuven, 1996). Ajoutons à l'accent phonétique et à l'accent perçu une troisième catégorie que nous appellerons accent théorique, et faisant référence à l'accent tel qu'il est généralement prescrit dans une langue donnée. Afin de simplifier la distinction entre ces termes, nous utiliserons le plus souvent dans cette thèse le terme de « proéminence » (Cruttenden, 1997) pour parler de l'accent phonétique, tandis que les termes d'accent lexical ou d'accent théorique feront référence à l'accent prescrit par le dictionnaire, et plus précisément dans notre cas le *Carnegie Mellon University Pronouncing Dictionary*.[6]

Après avoir présenté les caractéristiques fonctionnelles et acoustiques de l'accent lexical en anglais, puis celles de l'accent en français et en japonais, nous nous intéresserons à l'impact de l'accent lexical sur la compréhensibilité du locuteur, aux difficultés de perception et de production de l'accent en anglais L2, et enfin à quelques outils automatiques existants pour mesurer l'accent lexical.

3.3.1 L'accent lexical en anglais

En anglais, l'accent lexical se manifeste par des modifications à la fois prosodiques et segmentales des voyelles. Les syllabes accentuées sont généralement plus longues en durée, plus fortes en intensité, plus hautes en fréquence fondamentale (f_0), et présentent un mouvement de f_0 plus important, avec une qualité vocalique dite « pleine », comparativement aux syllabes non accentuées qui ont tendance à être « réduites » (Cutler, 2015). Par ailleurs, l'accentuation d'une syllabe affecte les syllabes non accentuées environnantes, les rendant plus courtes, moins fortes, moins hautes, centralisées et relâchées (Tortel, 2021). La voyelle réduite par excellence en anglais est le *schwa*, noté /ə/, mais le phénomène d'accentuation-réduction doit plutôt être considéré comme un continuum, allant de fortement accentué (quand se superposent l'accent lexical et l'accent de phrase par exemple) à complètement réduit, voire supprimé (ex. "*cho*colate" souvent prononcé [ˈt͡ʃɒklət]). On distingue jusqu'à 7 niveaux d'accentuation, mais quatre niveaux sont phonologiquement pertinents : l'accent primaire, l'accent secondaire, la syllabe pleine non accentuée (ou accent tertiaire), et la syllabe réduite (Frost, 2023).

Le rôle principal de l'accent lexical est la segmentation du flux de parole et la désambiguïsation lexicale. En anglais, les mots pleins (noms, verbes, adjectifs, ad-

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

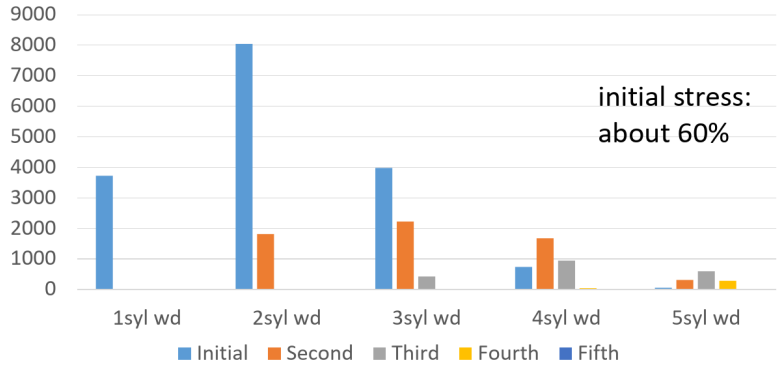


FIG. 3.1 : Distribution de la position de l'accent lexical en anglais dans les mots de 1 à 5 syllabes, à partir de la base de données CELEX (Sugahara, 2020, p. 8)

verbes, etc.) sont généralement accentués, tandis que les mots grammaticaux (prépositions, déterminants, particules, etc.) sont généralement réduits (Tortel, 2021). Par ailleurs, l'accent a tendance à se déplacer selon la catégorie du mot (*person* vs. *personifier*) et aide à distinguer des mots au sein de la même catégorie (*photograph* vs. *photographer*). De manière générale, les noms et adjectifs ont tendance à porter l'accent sur la première syllabe, tandis que les verbes sont plus souvent accentués sur la deuxième syllabe.

Les paires minimales se distinguant seulement par la position de l'accent sont rares en anglais, étant donné que celui-ci s'accompagne souvent de variations segmentales des voyelles (Cutler & Jesse, 2021).

La majorité des mots sont accentués sur la première syllabe. En effet, selon Sugahara (2020), 60 % des lemmes de la base de données CELEX (Baayen et al., 1995) sont accentués sur la première syllabe (*cf.* graphique 3.1). En analysant un corpus de 190 000 mots issus de conversations spontanées en anglais britannique, Cutler et Carter (1987) constatent que 90 % des mots lexicaux commencent par une syllabe accentuée. Ils en concluent que l'auditeur anglophone s'appuie certainement sur les syllabes accentuées pour repérer les frontières de mots dans le flux de parole.

3.3.2 L'accent de mot en français

Le français n'a pas d'accent lexical (Vaissière & Michaud, 2006), mais il n'est pas pour autant dénué d'accentuation. On distingue en général deux types d'accents : l'accent emphatique, qui permet d'attirer l'attention de l'auditeur de manière ponctuelle sur un mot de l'énoncé, et l'accent non-emphatique qui, contrairement au premier, est systématiquement placé en fin de groupe rythmique. Nous nous intéresserons ici seulement à l'accent non-emphatique.

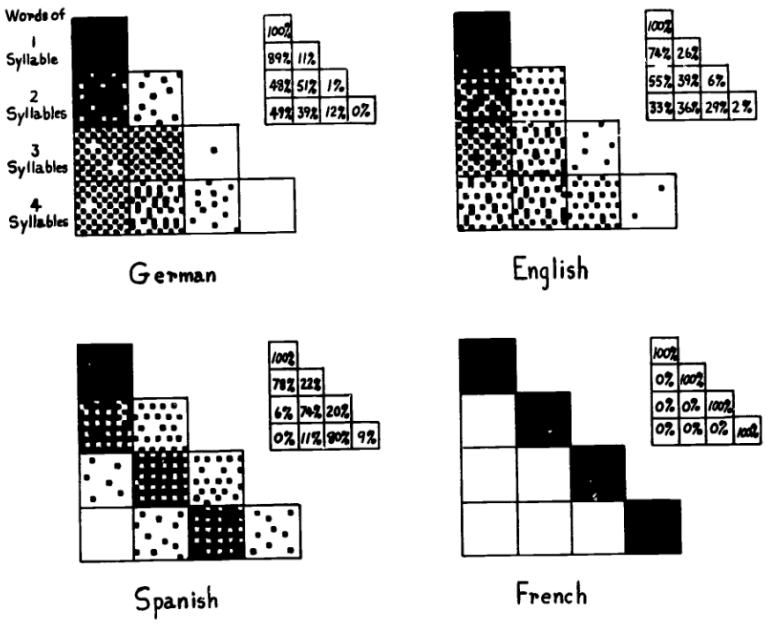


FIG. 3.2 : Comparaison de la position de l'accent primaire dans quatre langues (Delattre, 1963, p. 200)

Le français est traditionnellement décrit comme une langue à accentuation finale, aussi appelée oxytonique, où l'accent (non-emphatique, donc) tombe sur la dernière syllabe d'un groupe de mots (Astesano, 2001). On distingue généralement deux niveaux de groupes rythmiques en français. Le plus grand est appelé groupe de souffle ou unité intonative ; il peut contenir plusieurs groupes plus petits appelés groupes accentuables, composés d'un mot lexical (accentuable) et éventuellement de mots grammaticaux qui en dépendent (généralement non accentués). Selon Di Cristo (1998), la dernière syllabe des groupes accentuables est systématiquement accentuée. Il est parfois fait mention d'un accent rythmique secondaire placé sur la syllabe initiale et permettant de délimiter les unités intonatives (Di Cristo & Hirst, 1993 ; Fónagy, 1980).

Dans une étude comparative, Delattre (1963) analyse la position de l'accent primaire (théorique) dans un corpus de textes de 1500 mots en français et en espagnol, 2400 mots en allemand et 5800 mots en anglais. Il montre que l'anglais et le français se comportent de manière diamétralement opposée, le premier exhibant une grande variabilité, l'autre étant étonnamment stable. La figure 3.2 présente le pourcentage d'accentuation de chaque syllabe pour les mots d'une à quatre syllabes, dans les quatre langues analysées.

L'accentuation du français est avant tout caractérisée par une variation de durée de syllabe (Astesano, 2001 ; Di Cristo, 1998). Cette variation est d'autant plus grande qu'elle se cumule avec l'allongement de fin de groupe, ainsi la dernière syllabe d'un

groupe a tendance à paraître notoirement plus longue que les précédentes (Nord et al., 1990), sans qu'il soit clairement établi quelle proportion de l'allongement final est due à la position en fin de groupe, et quelle proportion est due à l'accentuation (Astesano, 2001).

D'après Vaissière (1991), la fréquence fondamentale sert en français principalement à marquer la frontière des mots, et non pas à accentuer une syllabe. La f_0 a tendance à monter en fin de mot (de manière étalée sur plusieurs syllabes), pour reprendre plus bas au début du mot suivant ; ou bien à tomber si le mot se situe en fin de groupe de souffle.

L'intensité, quant à elle, ne semble pas être un paramètre déterminant de l'accentuation du français. Il apparaît même que la voyelle d'une syllabe finale (donc accentuée) est en moyenne moins intense que les autres syllabes (-0.5 dB en français, contre 4.4 dB en anglais pour la syllabe accentuée, Delattre, 1966).

La fonction première de l'accent non emphatique en français est de structurer l'énoncé en groupes de sens (Astesano, 2001). Il permet à l'auditeur de segmenter le flux de parole et de focaliser son attention sur les informations importantes ou nouvelles. Il peut avoir également des fonctions secondaires expressives, contrastives ou rythmiques.

3.3.3 L'accent lexical en japonais

Selon Sugahara (2020), les deux dialectes principaux du japonais, à savoir le dialecte de Tōkyō et celui du Kansai, ont tous les deux un accent lexical caractérisé seulement par une variation de la f_0. De ce fait, on parle souvent d'accent de hauteur ou *pitch accent*, mais il s'agit bien d'un accent lexicalement contrastif. D'après Shibata et Shibata (1990), 13,6 % des homophones japonais sont distingués exclusivement par la position de l'accent, par exemple 箸 ***ha***si (baguettes de table), 橋 *ha****si*** (pont) ou 端 *hasi* (limite). Toutefois, la position de l'accent varie en fonction du dialecte. On pourra ainsi avoir 心 *ko**kor**o* à Tōkyō et ***ko***kor*o* dans le Kansai (cœur, esprit), ou encore カマキリ ***ka***maki*ri* à Tōkyō et *kama**ki**ri* dans le Kansai (mante religieuse).

Précisons que l'accent est communément rattaché à la more, et non à la voyelle ou à la syllabe. La more est une unité légèrement plus petite que la syllabe, composée dans le cas du japonais d'un noyau vocalique éventuellement précédé d'une consonne et d'un glide, ou elle peut être aussi une consonne nasale seule, un coup de glotte ou un allongement vocalique.

La position de l'accent est régie par un système complexe qui varie selon la région et l'origine du mot, mais sa position la plus courante semble être la more antépénultième ou la pénultième (Kubozono, 2006). La figure 3.3 présente la distribution de

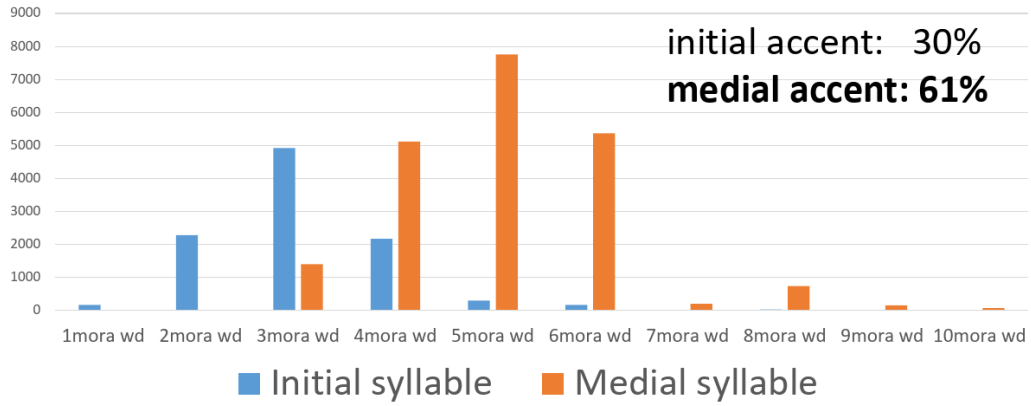


FIG. 3.3 : Distribution de la position de l'accent lexical dans les mots de 1 à 10 mores du Ōsaka-Tōkyō Accent Dictionary (Sugahara, 2020, p. 15)

l'accent lexical dans les mots de 1 à 10 mores dans le Ōsaka-Tōkyō Accent Dictionary (Sugafuji, 1996). On peut voir que la majorité des mots sont accentués sur une more en position médiale (61%). Si la more initiale porte généralement l'accent sur les mots de 1 à 3 mores, Sugahara (2020) indique que le japonais a tendance à avoir des mots de plus de trois mores (comme le montre la distribution), du fait de sa morphologie agglutinante. Par ailleurs, l'accent a tendance à se rapprocher de la frontière du morphème (et donc en position médiale) quand de nouveaux éléments viennent s'y ajouter. Ainsi 京都 *kyooto* (Kyōto) devient 京都市 *kyooto-si* (la ville de Kyōto), mais 京都大学 *kyooto-daigaku* (l'université de Kyōto). Il serait intéressant de connaître la distribution de la position de l'accent sur un corpus de mots courants, mais nous n'avons malheureusement pas pu trouver cette information.

Le japonais est donc une langue à accent lexical. Celui-ci est majoritairement en position médiale, et est caractérisé par une variation de la f_0 sans modification significative des autres dimensions prosodiques ou segmentales.

3.3.4 L'accent lexical en anglais L2

Dans les contextes d'apprentissage d'une langue seconde, les locuteurs/auditeurs non-natifs sont souvent influencés par les règles prosodiques de leur langue maternelle, et cela peut poser plus ou moins de problèmes selon que ces règles ou tendances diffèrent de la langue cible (Cutler, 2015). Par exemple, le locuteur francophone, habitué à un accent fixe sur la syllabe finale et une qualité et une durée stables des voyelles, aura tendance à accentuer la dernière syllabe des mots en anglais et à ne pas réduire les syllabes non accentuées (Tortel & Hirst, 2010). On peut s'attendre par ailleurs à ce que cette accentuation soit plus prononcée en termes de durée de syllabe, que de variation

de f_0 ou d'intensité, comme ces deux derniers paramètres ne semblent pas particulièrement exploités pour accentuer les syllabes en français. De plus, puisque l'accentuation dans leur langue ne joue pas de rôle de désambiguïsation lexicale comme en anglais, les locuteurs francophones ont souvent des difficultés à conscientiser les patterns accentuels de l'anglais, et peuvent avoir du mal à reconnaître leur propre tendance à accentuer les syllabes finales. Dupoux et al. (1997) proposent le terme de « surdité accentuelle » *(stress deafness)* pour décrire cette capacité limitée à percevoir et à être conscient de l'accent, notant que les locuteurs de langues à accent fixe rencontrent plus de difficultés comparativement à ceux des langues à accent lexical. De plus, adopter un rythme différent de celui de sa langue maternelle peut être psychologiquement éprouvant, car celui-ci est ancré depuis l'enfance et fortement associé avec sa personnalité et sa culture (Calbris & Montredon, 1975). En conséquence, un accent mal placé et l'absence de réduction syllabique peuvent significativement entraver la segmentation et la reconnaissance des mots pour les auditeurs (Cutler, 2015). Tortel (2021) souligne que les apprenants francophones de l'anglais devraient prioriser l'amélioration de la position de l'accent lexical, le contraste entre les syllabes accentuées et réduites, éviter l'allongement des syllabes finales non accentuées, et réduire les mots fonctionnels.

Si le japonais a quant à lui un accent lexical, et que les locuteurs japonophones semblent avoir moins de difficultés à percevoir et produire l'accent en anglais, ils restent toutefois influencés par la distribution de l'accent du japonais – plus souvent en position médiale –, et ont tendance à ne pas réduire les voyelles non-accentuées (Sugahara, 2011, 2016).

3.3.5 Accent lexical et compréhensibilité

Isaacs et Trofimovich (2012) constatent que l'accentuation lexicale est le troisième paramètre le plus corrélé avec la compréhensibilité, parmi les 19 paramètres qu'ils analysent (*cf.* section 2.3). Ils calculent un taux d'erreur d'accent lexical à partir du nombre de mots polysyllabiques dont l'accent primaire est mal placé ou absent, divisé par le nombre de mots polysyllabiques. La corrélation entre la proportion d'erreur d'accentuation et le jugement de compréhensibilité est de $-0{,}76$ ($p < 0{,}01$), suivie immédiatement de la proportion de réduction vocalique ($0{,}74$, $p < 0{,}01$). Contrairement aux paramètres de fluence qui apparaissent plus discriminants dans les premiers niveaux du CECRL, l'accentuation lexicale est discriminante pour tous les niveaux de locuteurs.

Saito et al. (2015) reprennent les données de Isaacs et Trofimovich (2012), et proposent à une autre cohorte d'évaluateurs experts et non-experts d'évaluer chaque locuteur en termes de compréhensibilité, puis selon 11 critères linguistiques comme

l'accentuation lexicale, le rythme ou le débit de parole[7]. Il s'agit cette fois de voir sur quelles dimensions les évaluateurs s'appuient explicitement pour juger les productions langagières des locuteurs, et de déterminer les dimensions les plus corrélées avec le jugement global de compréhensibilité. Deux éléments dans leurs résultats sont intéressants à mentionner ici. Tout d'abord, les évaluations subjectives du rythme et de l'accentuation lexicale sont apparues fortement corrélées avec les annotations effectuées par Isaacs et Trofimovich (2012) ($r = 0{,}76$, $p < 0{,}01$ entre l'évaluation du rythme et la proportion de réduction vocalique ; $r = 0{,}70$, $p < 0{,}01$ entre l'évaluation de l'accentuation lexicale et le taux d'erreur d'accent lexical). Ensuite, le rythme s'est révélé être le critère d'évaluation le plus corrélé avec le jugement de compréhensibilité ($r = 0{,}79$) parmi les cinq critères de prononciation. L'accentuation lexicale est quant à elle en quatrième position, après les erreurs segmentales et le débit de parole.

3.3.6 Mesures automatiques de l'accent lexical

Plusieurs études ont proposé des systèmes de classification automatique de l'accent lexical depuis le début des années 2000. La plupart de ces systèmes s'appuient sur des mesures de f_0, d'intensité et de durée de syllabe ou de segments vocaliques (J.-Y. Chen & Wang, 2010 ; L.-Y. Chen & Jang, 2012 ; Deshmukh & Verma, 2009 ; Johnson & Kang, 2015 ; K. Li et al., 2018 ; Tepperman & Narayanan, 2005). et intègrent parfois des informations segmentales, comme les coefficients cepstraux (Ferrer et al., 2015 ; C. Li et al., 2007 ; Shahin et al., 2016).

Shahin et al. (2016) ont par exemple développé un système de catégorisation de l'accent pour l'anglais et l'arabe standard, en entraînant des réseaux de neurones profonds (DNN) et convolutionnel (CNN). Ils entraînent leurs systèmes à catégoriser chaque syllabe en deux ou trois classes, à partir d'un corpus de lecture de mots isolés et de phrases en anglais (TIMIT, Garofolo et al., 1993 et OGI, Shobaki et al., 2000), et un corpus de lecture en arabe standard provenant d'un manuel scolaire. Dans les données d'apprentissage, l'accent en anglais est annoté avec le dictionnaire CMU Pronouncing Dictionary, et chaque syllabe est représentée par un vecteur de 2 451 mesures, composées de 27 coefficients cepstraux sur 30 trames de signal à l'intérieur de la syllabe, de l'intensité et de la f_0 moyennes et de leur variation, ainsi que de la durée de la syllabe et du noyau syllabique. Les auteurs reportent une précision maximale de 92,8 % pour le corpus OGI avec le CNN et une catégorisation binaire de l'accent (accentué ou non accentué). Toutefois, les auteurs n'ont pas testé leur système sur de la parole L2.

[7] 5 critères d'évaluation de l'enregistrement audio (erreurs segmentales, accent lexical, intonation, rythme, débit de parole) et 6 critères d'évaluation de la transcription des enregistrements (précision et richesse du lexique, précision et complexité grammaticale, richesse et cohésion du discours).

Ferrer et al. (2015) présente le classifieur développé pour le logiciel EduSpeak[8], intégrant également une combinaison de paramètres spectraux et prosodiques (durée de syllabe, f_0 et intensité). Ils entraînent leur système à distinguer trois catégories d'accentuation à partir d'un corpus de phrases courtes en anglais, lues par des enfants anglophones natifs. Les données d'apprentissage sont annotées avec un dictionnaire de prononciation (non précisé), et les auteurs comparent trois types de modèles : une mixture de gaussiennes (GMM), des arbres de décision, et un réseau de neurones. Les trois systèmes sont ensuite testés sur des phrases lues par des enfants anglophones et japonophones. Le GMM obtient les meilleurs résultats avec une précision de 88,4 % pour les locuteurs natifs et 80,0 % pour les apprenants.

Dans une étude plus récente, Korzekwa et al. (2021) ont entraîné un réseau de neurones à attention sur des mesures de f_0, d'intensité et de durée de phonèmes. Ils entraînent leur système sur un corpus combinant de mots isolés partiellement générés par synthèse vocale, de manière à ajouter des accentuations non standard. Leur système atteint une précision de 94,8 % sur de la lecture de mots isolés par des locuteurs polonais, et dont l'accentuation a été annotée en parallèle par cinq linguistes anglophones natifs. Les auteurs rapportent également que la moitié des erreurs ne sont pas détectées (rappel de 49,2 %). Enfin, les auteurs ajoutent que leur modèle n'est pas adapté à l'analyse de mots en contexte, car les résultats sont biaisés par les phénomènes de liaison et de coarticulation.

Les systèmes présentés ici obtiennent une précision généralement élevée, malgré un rappel important, lorsque celui-ci est mentionné par les auteurs. Toutefois, ils sont tous entraînés et testés exclusivement sur de la parole lue, voire sur des mots isolés. Nous n'avons trouvé aucune étude sur la parole spontanée, ni même sur de la lecture plus longue que celle de phrases telles que celles du corpus TIMIT. Par ailleurs, peu de détails sont donnés quant à la méthode d'évaluation de ces outils. La plupart des études s'appuient sur un codage manuel de l'accent par des linguistes, mais vantent en parallèle l'intérêt des systèmes automatiques face aux faibles taux d'accord inter-évaluateur. D'après Saito et al. (2022), on ne sait pas encore comment mesurer automatiquement la précision de l'accent lexical.

Enfin, nous n'avons trouvé aucun système open-source, qui aurait pu être réutilisé et adapté pour notre recherche.

Nous avons vu dans cette section que l'accent de mot joue un rôle important dans la compréhension de l'énoncé, en fournissant à l'auditeur des indices de segmentation du flux de parole. L'anglais se caractérise par un accent à position variable, dit lexical, le plus souvent placé sur la syllabe initiale, et formant un contraste prosodique important avec les syllabes non accentuées qui l'entourent. Cette accentuation est produite

[8] https://eduspeaks.online/ (consulté en janvier 2025)

par une variation de hauteur, d'intensité, de durée et de qualité vocalique. Le français, quant à lui, présente un accent à position fixe en fin de groupe rythmique, principalement caractérisé par une variation de durée. Le japonais, enfin, possède un accent lexical majoritairement placé en position médiale, et marqué exclusivement par une variation de hauteur.

En outre, les tendances accentuelles de la L1 influencent souvent la production en L2, et lorsque les tendances sont opposées, comme en anglais et en français, il en résulte souvent un rythme général déstabilisant pour l'auditeur. L'absence d'accent lexical en français ajoute une difficulté supplémentaire pour ses locuteurs lorsqu'ils apprennent l'anglais ou le japonais : l'accent n'est pas conscientisé comme un critère linguistique – on parle parfois de surdité accentuelle *(stress deafness)* (Dupoux et al., 1997). Enfin, si de nombreuses études ont proposé des systèmes de catégorisation automatique de l'accent depuis une vingtaine d'années, aucune ne semble encore avoir proposé de système adapté pour la parole spontanée.

Conclusion

Nous avons vu dans ce chapitre que les patterns de pause et d'accentuation lexicale participent à rendre la parole du locuteur plus ou moins compréhensible. Les pauses, ou plus largement les interruptions du flux de parole, peuvent structurer le message et aider l'auditeur à traiter l'information. Dans certains cas, au contraire, elles peuvent le déstabiliser et perturber la compréhension. Il apparaît que les pauses situées au niveau de frontières syntaxiques importantes ont un impact positif sur la compréhension, tandis que celles situées à l'intérieur des constituants syntaxiques ont tendance à être associées à une parole moins fluente et plus difficilement compréhensible. Le lien entre les pauses et la syntaxe est clair, et il est important de les considérer en contexte, pour ne pas pénaliser les pauses qui structurent l'énoncé.

L'accent lexical en anglais joue un rôle important dans la segmentation du flux de parole, et apparaît également fortement corrélé avec la perception de compréhensibilité. Le fait que les patterns accentuels de la L1 influencent souvent la production en L2 peut rendre la segmentation ou la reconnaissance de certains mots difficile. La différence majeure qui ressort entre l'anglais d'une part, et le français ou le japonais d'autre part, au-delà des tendances de position, c'est que les syllabes de l'anglais présentent la caractéristique d'être « élastiques », tantôt étirées (plus hautes, plus fortes, plus longues), tantôt réduites, aboutissant à un contraste important entre les syllabes. Les syllabes du français et du japonais sont quant à elles beaucoup plus « stables », malgré un allongement fréquent en finale en français, ou une variation lexicale de hauteur en japonais. Aussi, en anglais L2, les locuteurs de ces langues ont tendance à ne pas produire de contraste marqué entre les syllabes, impactant le rythme général de leur parole, et les rendant plus difficilement compréhensibles.

Problématique

Nous avons observé un changement de paradigme dans le domaine de l'évaluation de la prononciation en L2 : ce n'est plus la distance avec une prononciation native qui importe, mais l'efficacité de la communication en termes de compréhension. Ce qui compte, c'est que l'auditeur comprenne le locuteur et qu'il le comprenne avec autant de facilité que possible. On parle d'« intelligibilité » pour qualifier le niveau de compréhension de cet auditeur vis-à-vis de la production du locuteur, et de « compréhensibilité » pour qualifier le degré d'aisance avec lequel il parvient à comprendre.

Les descripteurs de compétences linguistiques internationaux comme le Cadre Européen Commun de Référence, ainsi que les grilles d'évaluation de production orale des principaux tests certificatifs d'anglais, mettent en effet clairement en avant l'importance de la compréhension. Le niveau de l'apprenant est maintenant déterminé en fonction de son intelligibilité du point de vue de l'auditeur. Un seuil particulièrement important est décrit au niveau B2 ou équivalent, où le locuteur devient globalement fluide et facilement compréhensible.

Cependant, les descripteurs manquent de définitions explicites de ce qui constitue une parole fluide et compréhensible, et le jugement reste à la charge de l'interprétation de l'évaluateur. On constate pourtant que leur évaluation est généralement consistante parmi les évaluateurs, qu'ils soient experts ou non, et qu'ils participent ou non directement à l'interaction.

En parallèle, les outils d'évaluation automatique de la prononciation reposent soit sur des paramètres de bas niveau corrélés avec le niveau global de l'apprenant mais toutefois généralement peu pertinents sur le plan pédagogique – comme le débit de parole ou la fréquence des pauses –, soit sur la mesure d'une distance vis-à-vis d'un modèle natif, sans chercher à identifier les éléments qui réduisent la compréhensibilité. De plus, il est important de pouvoir évaluer la performance de l'apprenant en contexte spontané, en interaction, afin d'estimer sa compétence en situation de communication.

Il apparaît donc souhaitable de développer des systèmes automatiques capables d'évaluer la production orale spontanée en se basant sur des phénomènes linguistiques spécifiquement susceptibles d'entraver la compréhension.

Facteurs linguistiques clés

De nombreux facteurs peuvent impacter l'intelligibilité et la compréhensibilité, depuis la précision phonologique de l'énoncé du locuteur, à l'implication de l'auditeur dans la conversation, en passant par sa familiarité avec le sujet abordé. Parmi ces facteurs, plusieurs dépendent avant tout de la performance linguistique du locuteur, bien que leur effet puisse varier selon la compétence de compréhension de l'auditeur. Nous proposons de nous concentrer ici sur les deux éléments suivants :

- **Distribution des pauses** : Les pauses situées au niveau de frontières syntaxiques de haut niveau, typiquement entre propositions, tendent à améliorer la compréhension et reflètent une meilleure planification de l'énoncé. Par exemple, une pause après une conjonction de coordination ou avant une relative permet de structurer l'information de manière plus accessible pour l'auditeur. En revanche, celles situées à l'intérieur d'unités syntaxiques de petite taille sont souvent perçues comme des disfluences qui rendent la compréhension difficile.
- **Rythme et accentuation lexicale** : Le rythme de la parole joue un rôle important dans la compréhensibilité du locuteur. En anglais, il est défini en grande partie par l'accentuation des voyelles, qui, au-delà de jouer un rôle lexical permettant une meilleure reconnaissance des mots, permet aussi une segmentation de la parole à un plus bas niveau : l'accent se trouve fréquemment en tête de mot et en début de groupe prosodique, et les mots grammaticaux ne sont généralement pas accentués. Par ailleurs, les voyelles non accentuées sont souvent réduites, créant un contraste prosodique marqué avec les voyelles accentuées.

Les récentes avancées en traitement automatique de la parole ouvrent de nouvelles perspectives pour l'évaluation de la production orale spontanée des apprenants de langues étrangères. En combinant reconnaissance automatique de la parole et analyses syntaxiques, il devrait être possible de détecter les pauses et d'identifier leur position par rapport aux frontières syntaxiques. Cette approche permettrait de considérer les pauses relativement à la structure syntaxique de l'énoncé et ainsi de valoriser celles qui sont placées à des frontières stratégiques, ou de ne pénaliser que celles qui ne le sont pas.

De même, en exploitant des techniques de détection des noyaux syllabiques et d'analyses acoustiques, il devrait être possible d'identifier les syllabes acoustiquement proéminentes susceptibles d'être perçues comme accentuées par les auditeurs, de comparer leur position avec la position attendue de l'accent lexical selon un dictionnaire de référence, mais aussi de quantifier le contraste acoustique entre les syllabes accentuées et les syllabes réduites.

Questions de recherche

L'objectif de cette thèse est de répondre à trois questions principales :

1. Peut-on concevoir un outil d'annotation automatique de la parole spontanée L2 capable de caractériser la distribution des pauses et les patterns accentuels qui rendent la compréhension difficile ?
2. Observe-t-on des différences significatives entre les locuteurs de niveaux B1 et B2 en termes de distribution syntaxique des pauses et d'accentuation lexicale ?
3. Peut-on mesurer dynamiquement l'impact de ces facteurs sur la perception de compréhensibilité chez les auditeurs ?

Par la réalisation d'un outil de ce type, notre objectif est de proposer un complément d'informations quantifiables pour les évaluateurs, mais aussi un outil de diagnostic pour les enseignants et les apprenants, en exploitant le potentiel des outils actuels de traitement automatique de la parole.

Hypothèses

En appliquant des outils de traitement automatique à un corpus de productions orales spontanées d'apprenants francophones et japonophones de niveaux B1 et B2, plusieurs comparaisons peuvent être envisagées :

- **Distribution des pauses** : Les locuteurs B2, perçus comme plus fluides et compréhensibles, devraient produire une proportion plus élevée de pauses en frontières syntaxiques de haut niveau (ex. : entre propositions), et moins de pauses en frontières de bas niveau (ex. : au sein d'unités syntaxiques réduites).
- **Accentuation lexicale** : Les productions devraient refléter une influence variable des langues sources. Les francophones, issus d'une langue à accent fixe marqué par un allongement de la syllabe finale, devraient présenter une tendance à rallonger systématiquement les syllabes en fin de mot et faire peu varier la f_0 et l'intensité. Pour les locuteurs japonophones, en revanche, nous nous attendons à observer moins de difficultés à positionner correctement l'accent, puisque le japonais possède également un accent lexical, mais toutefois un contraste plus marqué par une variation de la f_0, que par celle de l'intensité ou de la durée.

Enfin, nous nous questionnons sur l'effet concret que peuvent avoir les pauses situées en frontières syntaxiques de bas niveau, et les patterns accentuels non adéquats, sur la perception de difficulté de compréhension chez l'auditeur. Nous faisons l'hypothèse que ces phénomènes ont un impact direct et observable sur la perception de compréhensibilité, et qu'ils peuvent être mesurés à l'aide d'un protocole d'évaluation dynamique semblable à celui utilisé par Nagle et al. (2019). Le jugement de difficulté perçue par les auditeurs devrait avoir tendance à augmenter à la suite de ces phénomènes, et au contraire à diminuer, ou au moins ne pas augmenter, à la suite de pauses en frontières syntaxiques de haut niveau et de patterns accentuels adéquats.

Deuxième partie

Corpus & méthodologie d'analyses

Chapitre 4

Collecte de données de parole

Notre objectif est d'observer les schémas de pauses et d'accentuation lexicale chez des locuteurs francophones et japonophones d'anglais, situés à la frontière entre deux niveaux de compétence : « intelligible mais difficile à comprendre » (niveau B1) et « intelligible et facile à comprendre » (niveau B2). Pour mener cette étude, plusieurs contraintes méthodologiques doivent être respectées. Premièrement, les enregistrements de parole doivent être d'une qualité audio suffisante et d'une durée adéquate pour permettre des analyses automatiques. Deuxièmement, la parole recueillie doit être spontanée, de manière à ce que les pauses reflètent la capacité du locuteur à planifier et structurer son discours. Enfin, les locuteurs doivent appartenir aux niveaux B1 ou B2 afin de comparer les schémas de pauses et d'accentuation entre ces deux groupes de compétence.

L'épreuve d'interaction orale du CLES B2 s'est révélée être une opportunité particulièrement adaptée à la constitution de ce type de corpus. Elle propose en effet une mise en situation originale sous la forme d'un jeu de rôles d'une dizaine de minutes entre deux ou trois candidats, évalués *in situ* par un ou deux examinateurs accrédités.

Avec le soutien de la Direction du CLES, des coordinateurs et des évaluateurs sur le terrain, nous avons obtenu l'autorisation d'enregistrer plusieurs sessions du CLES B2 à Grenoble et à Valence. Ces enregistrements ont permis de constituer un premier corpus de locuteurs francophones, ci-après désigné sous le nom de CLES-FR. En parallèle, afin de disposer d'un corpus comparable pour des locuteurs japonophones, nous avons collaboré avec les universités Dōshisha (Kyōto) et Waseda (Tōkyō) au Japon. Ce corpus sera désigné sous le nom de CLES-JP. Enfin, pour observer les comportements de locuteurs anglophones natifs dans une situation de communication similaire, nous avons constitué un troisième corpus, nommé CLES-EN.

Dans ce chapitre, nous présenterons en détail chacun de ces trois corpus : leurs caractéristiques, leurs points communs et leurs spécificités.

4.1 Corpus CLES-FR

La certification CLES évalue les candidats sur quatre habiletés langagières, à trois niveaux différents : B1, B2 et C1. Au niveau B2, l'expression orale est évaluée dans le cadre d'une épreuve d'interaction orale. Lors de cette épreuve, deux ou trois candidats participent à un jeu de rôles, où il leur est demandé de présenter et défendre un point de vue sur un sujet d'actualité, puis d'aboutir à un compromis en une dizaine de minutes. Le rôle attribué à chaque candidat détermine leur position : en faveur ou contre le sujet traité. Ce sujet est directement lié aux épreuves de compréhension orale et écrite réalisées en amont ; les candidats ont donc déjà été exposés à la thématique.

Avant de commencer la discussion, les candidats disposent de 2 minutes de préparation pour organiser leurs idées et prendre des notes, sans toutefois être autorisés à les lire durant l'échange. La discussion s'achève au bout de 10 minutes, ou lorsque les candidats arrivent à un consensus et que les examinateurs jugent avoir suffisamment de matière pour évaluer les participants.

Le protocole d'évaluation CLES stipule que les candidats peuvent être enregistrés, pour permettre au jury de réécouter certains candidats lors de la délibération, d'utiliser certains enregistrements à des fins de formation d'évaluateurs ou de recherches dans le cadre du CLES. La présence des microphones dans la salle n'a donc pas, a priori, influencé la production des candidats, puisqu'il s'agit d'une situation normale d'épreuve d'interaction orale du CLES. Toutefois, nous sommes conscients que le contexte d'examen impacte largement la production des locuteurs (stress, sentiment de jugement, enjeux), à quoi s'ajoute la présence d'un microphone qui peut être une source de stress supplémentaire.

De nouveaux sujets de certification sont régulièrement édités par le CLES. Les sujets utilisés cette fois-ci concernent l'usage de la cigarette électronique, la généralisation des caméras de surveillance et les tests cliniques sur les animaux. La formulation des sujets varie légèrement lorsque trois candidats participent au débat, de manière à ce que le troisième prenne un rôle de médiateur, comme présenté table 4.1. L'ensemble des énoncés du CLES B2 sur le thème de la cigarette électronique est accessible en ligne[1]. Les autres sujets sont encore utilisés par le CLES au moment de l'écriture de ce manuscrit, et ne peuvent donc pas être rendus publics.

[1] https://www.certification-cles.fr/se-preparer/exemples-de-sujets/exemple-de-sujet-cles-b2-anglais-1219069.kjsp?RH=8204107280166102

Consigne générale	You will participate in a contradictory debate with another candidate. You will be asked to defend a given position. You will use the information from the documents you have been studying to discuss, negotiate, and reach a compromise according to the role that has been assigned to you.
Situation	The journalists you are working with are preparing a documentary film on e-cigarettes. They hesitate on what to insist on. You and your colleague(s) have been asked to suggest the best possible angle for the film. At the end of the discussion, you should reach an agreement based on a compromise between your different perspectives.
Rôle A	You think that the report should concentrate on the concerns about e-cigarettes and the need for regulation.
Rôle B	You think that the report should concentrate on the possible benefits of e-cigarettes.
Rôle C	You think that the report should be objective and integrate all points of views.

Tab. 4.1 : *Exemple de sujet du CLES B2 sur l'utilisation de la e-cigarette (situation 2)*

L'évaluation est réalisée par un ou deux examinateurs accrédités présents dans la salle, selon une grille standardisée comportant huit critères couvrant tant la qualité linguistique que le respect des consignes (cf. grille CLES B2, section 1.1.2 et annexe A.3). Le niveau B2 est attribué uniquement si tous les critères sont validés ; à défaut, le candidat peut valider le niveau B1 si ses compétences sont jugées suffisantes, ou ne rien valider du tout.

Les enregistrements ont été réalisés entre janvier 2020 et janvier 2023, avec des équipements variés (microphones et paramètres d'échantillonnage). Tous les fichiers ont été normalisés pour garantir une qualité homogène (44,1 kHz, stéréo 16-bit PCM, 1411 kb/s).

Données recueillies

Un total de 260 locuteurs ont été enregistrés lors de cinq sessions CLES. Parmi eux, 232 ont participé en binômes, 15 en trinômes, et 13 en monômes. La répartition par genre est parfaitement équilibrée (130 femmes, 130 hommes). Sur l'ensemble, 151 locuteurs (58%) ont obtenu la certification B2, 75 (29%) ont été certifiés B1, et 34 (13%) n'ont validé aucun niveau. Quarante-cinq locuteurs (17%) ont déclaré une langue maternelle autre que le français, incluant notamment l'arabe, le chinois et d'autres langues déclarées par une ou deux personnes chacune.

Pour les besoins de cette thèse, seuls les candidats de langue maternelle française ayant obtenu un niveau B1 ou B2 ont été conservés dans le corpus final, excluant également les monômes. Cela a permis d'obtenir un corpus de 170 locuteurs, désigné par le terme *CLES-FR*. Parmi ces locuteurs, 99 ont été certifiés B2 (58%), 71 B1 (42%). Concernant spécifiquement l'épreuve d'interaction orale, 118 locuteurs (69%) ont obtenu un niveau B2, contre 52 (31%) un niveau B1. La répartition par genre reste équilibrée, avec 89 femmes (52%) et 81 hommes (48%). Enfin, 11 locuteurs (6%) ont participé en trinômes.

4.2 Corpus CLES-JP

Pour obtenir un corpus comparable de locuteurs japonais, une seconde collecte a été organisée au Japon, en reproduisant autant que possible les conditions du CLES B2. Les locuteurs ont été recrutés parmi les étudiants des universités Waseda et Dōshisha. Ils remplissaient les critères suivants : être de langue maternelle japonaise, avoir un niveau d'anglais équivalent à B1 ou supérieur (TOEFL iBT Speaking 16, IELTS Speaking 5.5, Eiken English Proficiency test pre-1 level), et accepter les conditions d'utilisation des enregistrements. Chaque candidat a complété un questionnaire détaillé, incluant des informations sur leurs séjours à l'étranger et leur profil linguistique. Les participants ont été répartis en binômes en veillant à ce que chacun des locuteurs ait un niveau proche de celui de son interlocuteur. Grâce au concours des universités locales, une rétribution financière a pu être octroyée aux participants à hauteur de 1 500 ¥ pour les étudiants de Waseda, et 2 000 ¥ pour ceux de Dōshisha (respectivement 9,4 € et 12,5 € au moment de l'expérimentation).

Il n'a pas été possible d'utiliser les mêmes sujets que pour le corpus CLES-FR car ceux-ci étaient toujours en circulation en France. Aussi, deux sujets similaires ont été conçus pour l'occasion : le premier sur l'utilisation des intelligences artificielles génératives en classe de langue et le second sur le travail en parallèle des études. Les deux sujets sont donnés en annexe B. Créer ces sujets nous a permis d'être plus près des préoccupations actuelles des étudiants et a permis une discussion plus riche, sachant qu'ils n'avaient pas eu l'occasion de travailler ces sujets en amont comme c'est le cas pour les candidats CLES. Par ailleurs, les participants ont bénéficié d'un temps de préparation de durée flexible avant de commencer la discussion, et d'une liste de mots clés écrits qu'ils étaient libres d'utiliser ou non. Pour laisser le temps aux participants de rentrer dans la discussion, la durée de celle-ci n'était pas limitée mais devait durer un minimum de 10 minutes.

L'ensemble des enregistrements a été effectué avec un Zoom Handy Recorder H2n, et échantillonné à 44,1 kHz, stéréo 16-bit PCM, 1411 kb/s.

Données recueillies

Un total de 29 locuteurs de langue maternelle japonaise ont été enregistrés dans le cadre de la constitution du corpus CLES-JP. On compte 17 femmes (59%) et 12 hommes (41%). Leur niveau de compétence en anglais est estimé à partir des résultats obtenus à différentes certifications, sur la base d'une auto-déclaration : 5 d'entre eux sont de niveau équivalent B1 (17%), 15 de niveau équivalent B2 (52%), et 9 de niveau équivalent C1 (31%). Suite à l'absence d'un candidat, l'un des binômes enregistrés est constitué d'un étudiant de niveau C1 et d'un enseignant d'anglais de langue maternelle japonaise qui a dû assurer le remplacement. Les tours de parole de l'enseignant ne sont pas incluses dans le corpus CLES-JP.

4.3 Corpus CLES-EN

Afin d'observer les patterns de pauses et d'accentuation par des locuteurs anglophones natifs dans une situation similaire, nous avons constitué un corpus supplémentaire de locuteurs natifs dans les mêmes conditions que le corpus de locuteurs japonophones. Les participants ont été recrutés à l'université de Dōshisha, parmi un groupe d'étudiants en échange originaires des États-Unis, arrivés au Japon quelques jours avant les enregistrements.

L'ensemble des enregistrements a été effectué avec un Zoom Handy Recorder H2n, et échantillonné à 44,1 kHz, stéréo 16-bit PCM, 1411 kb/s.

Données recueillies

Le corpus CLES-EN est constitué de 14 locuteurs, tous originaires des États-Unis, et inscrits en licence dans une université américaine. Ils ont entre 20 et 22 ans ($M = 20{,}5$), 9 d'entre eux sont des femmes et 5 sont des hommes.

4.4 Comparabilité des corpus

Les corpus CLES-FR et CLES-JP partagent un objectif commun : analyser les patterns de pauses et d'accentuation lexicale dans la parole spontanée de locuteurs de niveaux B1 et B2, dans une situation de communication comparable. Le corpus CLES-EN permettra de comparer, à titre indicatif, les résultats obtenus avec ceux de locuteurs anglophones natifs dans une situation de parole similaire.

Ces trois corpus présentent l'avantage d'être comparables à plusieurs niveaux :

- **Format des interactions** : Chaque corpus repose sur des dialogues argumentatifs d'une dizaine de minutes, en binômes ou trinômes, réalisés dans un contexte formel. Les participants ont eu un temps de préparation pour se familiariser avec le sujet avant de commencer la discussion.
- **Contraintes** : Les conversations sont des jeux de rôles, les locuteurs n'ont pas eu le choix du sujet, du rôle qui leur est attribué, ni du ou des partenaires de discussion. Dans la majorité des cas, ils ne connaissent pas leur interlocuteur.
- **Nature des sujets** : Les sujets abordés dans chaque corpus encouragent une prise de position et une argumentation structurée.
- **Enregistrements standardisés** : Les discussions ont été enregistrées dans des conditions acoustiques similaires et rééchantillonnées selon des paramètres uniformes (44,1 kHz, stéréo 16-bit PCM).
- **Niveaux de compétence** : Les participants sont majoritairement de niveaux B1 et B2, estimés à partir de certifications standardisées ou de critères comparables.

Malgré tout, certaines différences importantes demeurent et nous devrons en tenir compte lors de l'interprétation des résultats :

- **Contexte et enjeux** : Le corpus CLES-FR a été enregistré dans un cadre réel d'examen, avec le stress associé à une évaluation officielle. Les corpus CLES-JP et CLES-EN, en revanche, ont été constitués dans des conditions expérimentales rétribuée financièrement, avec nécessairement moins d'enjeux.
- **Préparation des participants** : Les locuteurs du CLES-FR ont pu travailler en amont le sujet abordé pendant la conversation, à travers les épreuves précédentes du CLES. Ce n'est pas le cas des locuteurs des deux autres corpus, qui ont découvert le sujet de discussion seulement quelques minutes avant le début de la conversation.
- **Taille des corpus** : Le corpus CLES-FR est le plus important avec 170 locuteurs, suivi par le corpus CLES-JP (29 locuteurs) et le corpus CLES-EN (14 locuteurs). Ces différences influenceront nécessairement la robustesse des analyses statistiques qui suivront.

4.5 Publication des corpus

Un corpus public constitué des enregistrements des sessions CLES pour lesquels l'ensemble des participants ont donné leur accord pour la publication des données a pu être mis à disposition sur la plateforme Ortolang[2]. L'accès au corpus est toutefois restreint à un cadre de recherches académiques. Parmi les 260 locuteurs initialement enregistrés, 162 ont donné leur accord (62%), et 138 d'entre eux font partie d'un binôme ou trinôme où tous les locuteurs impliqués ont donné leur accord de diffusion.

Le corpus publié réunit 62 enregistrements de 128 locuteurs ayant passé l'épreuve du CLES B2, totalisant 10 h de parole. Parmi les locuteurs, 119 ont déclaré avoir le français comme langue maternelle (93%). On compte 61 femmes (48%) et 67 hommes (52%). La durée moyenne des enregistrements est de 9 min 35 s (min.: 5 min 12 s, max.: 14 min 30 s). Le résultat obtenu au CLES est B2 pour 62 d'entre eux (48%), 50 ont validé le niveau B1 (39%) et 16 n'ont rien validé (13%).

Les corpus CLES-JP et CLES-EN ont quant à eux été publiés en l'état, sur la même plateforme[3][4].

4.6 Annotations *gold standard*

Un échantillon du corpus CLES-FR a été transcrit semi-automatiquement puis corrigé pour constituer un sous-corpus de référence. Vingt enregistrements ont été sélectionnés aléatoirement avec pour contrainte de présenter 20 candidats certifiés B2 et 20 certifiés B1, et un équilibre homme/femme. Le travail d'annotation a été effectué par un stagiaire au sein du Laboratoire d'Informatique de Grenoble. Une transcription automatique des enregistrements a d'abord été effectuée à l'aide du logiciel Whisper (Radford et al., 2022, modèle `base` multilingue), puis manuellement corrigée et segmentée en locuteurs.

Du côté des corpus CLES-JP et CLES-EN, les corpus ont d'abord été automatiquement segmentés en locuteurs avec la pipeline de diarisation `pyannote.audio` (Bredin, 2023), puis vérifiés manuellement dans leur intégralité.

[2] CLES-FR : https://hdl.handle.net/11403/cles-spontaneous-english
[3] CLES-JP : https://hdl.handle.net/11403/cles-jp_corpus
[4] CLES-EN : https://hdl.handle.net/11403/cles-en_corpus

Conclusion

Un total de 304 locuteurs ont pu être enregistrés pendant les deux collectes de données organisées, totalisant 26 h de parole spontanée. Parmi eux, 260 ont été enregistrés lors d'épreuves de la certification CLES sur les campus de Grenoble et Valence de l'université Grenoble Alpes ; et 44 locuteurs ont été enregistrés lors de simulations de sessions CLES, organisées dans les universités de Waseda et Dōshisha, au Japon.

Trois corpus ont été créés sur la base de ces enregistrements : CLES-FR, qui contient 170 locuteurs francophones B1 et B2 ; CLES-JP, qui contient 29 locuteurs japonophones B1, B2 et C1 ; et CLES-EN qui contient 14 locuteurs anglophones natifs originaires des États-Unis. CLES-FR constitue notre corpus principal en termes de volume et de représentativité des niveaux cibles de notre recherche, B1 et B2. Les corpus CLES-JP et CLES-EN sont de taille plus modeste mais permettront d'avoir un aperçu des résultats obtenus avec des locuteurs japonophones de différents niveaux, ainsi que des locuteurs natifs de l'anglais. Par ailleurs, ils ont été réalisés dans les mêmes conditions (recrutement des locuteurs, sans enjeu d'examen) et sur les mêmes sujets de conversation. Aussi, dans la partie résultats, nous proposons de comparer les niveaux B1 et B2 du corpus CLES-FR d'une part, et les niveaux B1, B2, C1 et locuteurs natifs d'autre part.

Dans le chapitre suivant, nous présentons les outils utilisés pour segmenter, transcrire et annoter les enregistrements de ces trois corpus, ainsi que la méthodologie adoptée et les outils développés pour analyser les patterns de pauses et d'accentuation lexicale.

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[2], 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/

[2] Système d'Évaluation en Langues à visée Formative : https://self.univ-grenoble-alpes.fr/

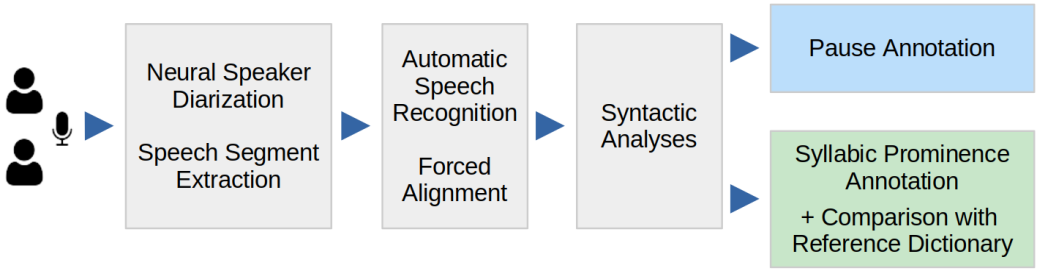


FIG. 5.1 : Architecture générale de PLSPP

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[3].

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

[3]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
```

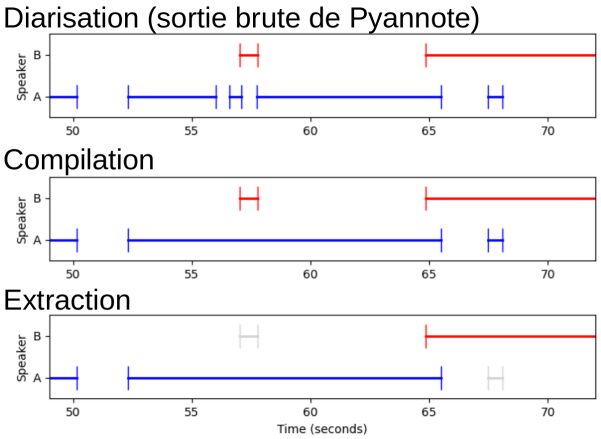


FIG. 5.2 : Exemple de segmentation en locuteurs, avec la sortie brute de Pyannote à gauche, sa visualisation graphique (en haut à droite), la compilation des segments consécutifs (seuil de durée fixé à 1 s*), et les segments extraits (sup. ou égal à* 8 s*)*

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 à 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[4] (v2.29), EML Transcription[5] (v1.19), Amberscript[6] (v1.3), Fraunhofer Speech Recognition[7] (v2.13), Radboud University LST[8] (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[9] et l'autre sur LibriSpeech[10].

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.

[4] Google Speech Cloud (2022) : https://cloud.google.com/speech-to-text/
[5] EML Transcription (2022) : https://www.eml.org/
[6] Amberscript (2022) : https://www.amberscript.com/
[7] Fraunhofer Speech Recognition (2022) : https://www.idmt.fraunhofer.de
[8] LST (2022) : https://webservices.cls.ru.nl/
[9] https://huggingface.co/speechbrain/asr-wav2vec2-commonvoice-en (consulté avril 2022)
[10] 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`[11].

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.

[11] 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

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

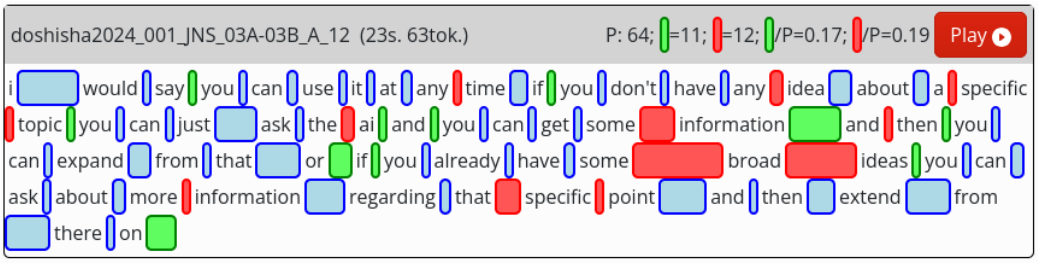


FIG. 5.4 : Transcription d'un segment audio du corpus CLES-JP, affichant chaque intervalle vide générée par l'aligneur automatique. La couleur des intervalles correspond au type de frontière syntaxique identifié (vert pour inter-proprosition, bleu pour inter-syntagme et rouge pour intra-syntagme). La 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[12] 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 :

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

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

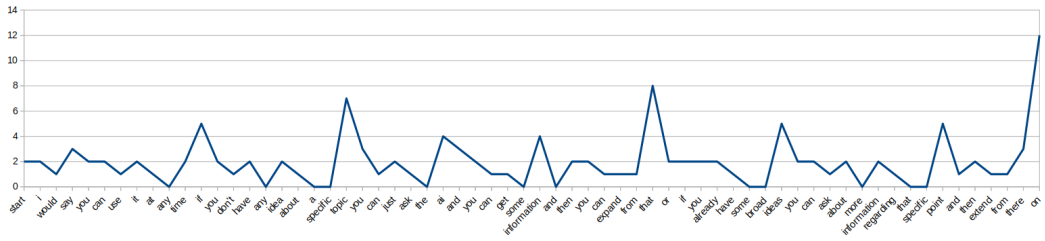


FIG. 5.5 : Nombre de constituants se fermant ou s'ouvrant après chaque mot

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*[13].

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

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

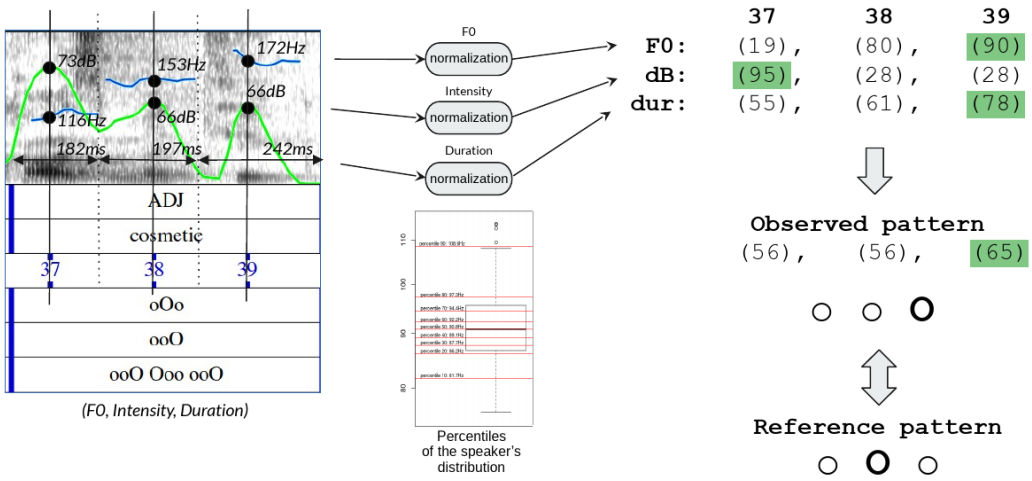


FIG. 5.6 : Extraction des paramètres prosodiques (PLSPP v1). À gauche un aperçu du fichier TextGrid de sortie avec les mesures acoustiques absolues indiquées en surimpression, la courbe bleue indique la f_0, la courbe verte indique l'intensité. À droite sont affichées les mesures normalisées. "Observed 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

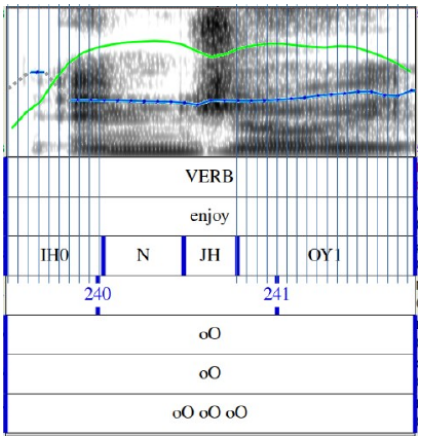


FIG. 5.7 : Extraction des paramètres prosodiques avec PLSPP v2. Les barres bleues ajoutées en surimpression représentent les frames de 10 ms pour le calcul de la f_0 (courbe bleue) et de l'intensité (courbe verte)

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 robuste 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 (Erickson et al., 2025 ; Raso et al., 2024).

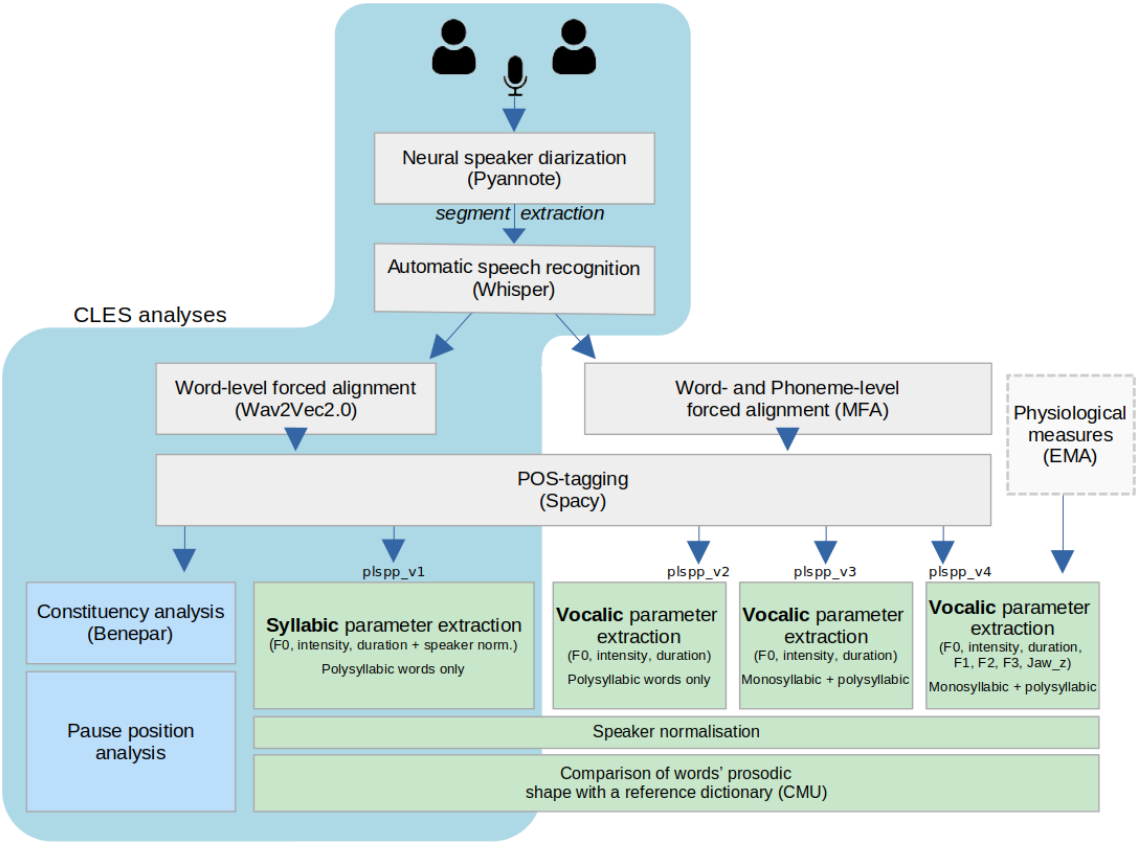


FIG. 5.8 : Architecture détaillée de PLSPP et ses différentes versions

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[14].

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.

[14] 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).

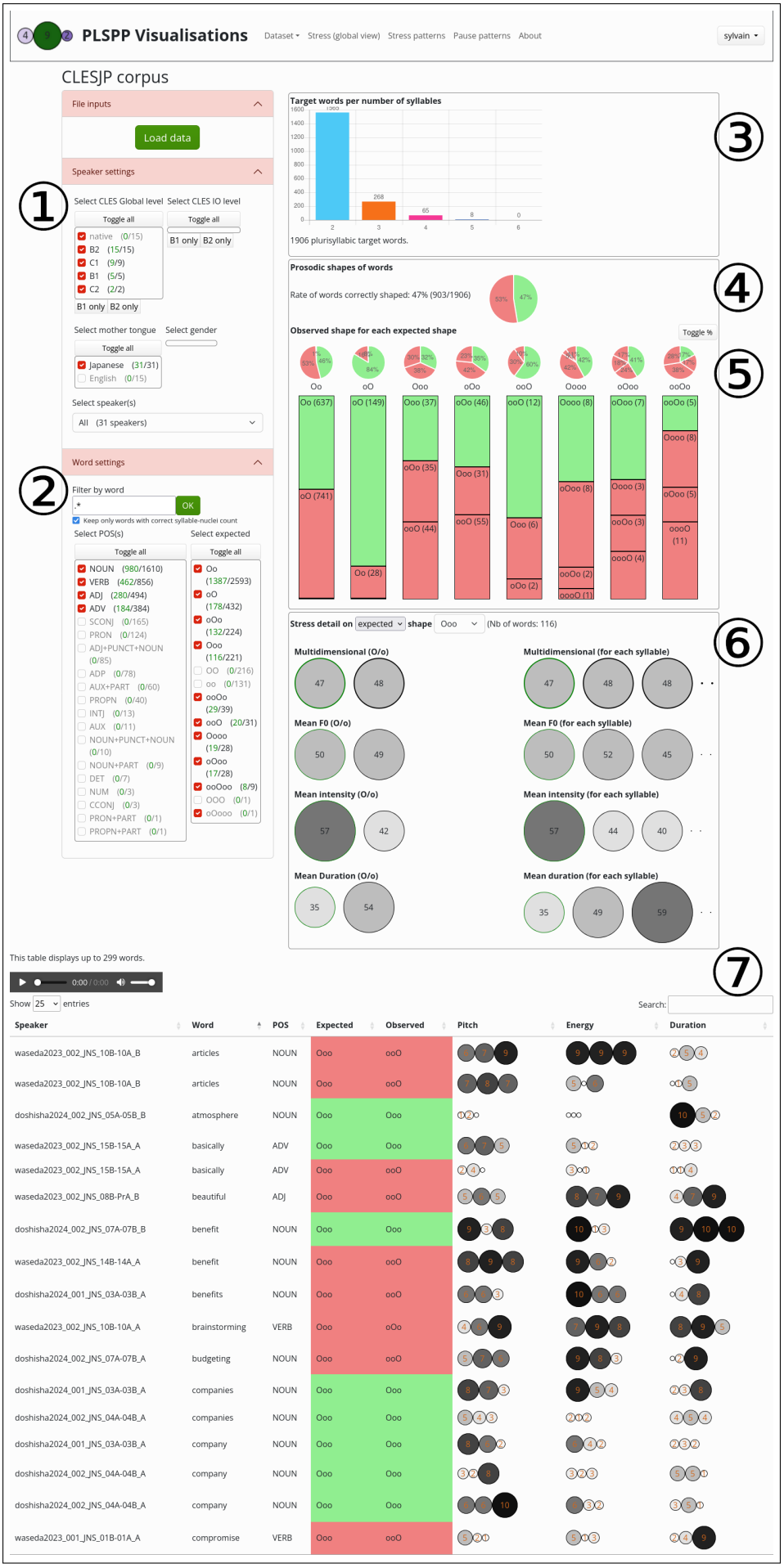


Fig. 5.9 : Aperçu de l'interface de visualisation de PLSPP, vue globale

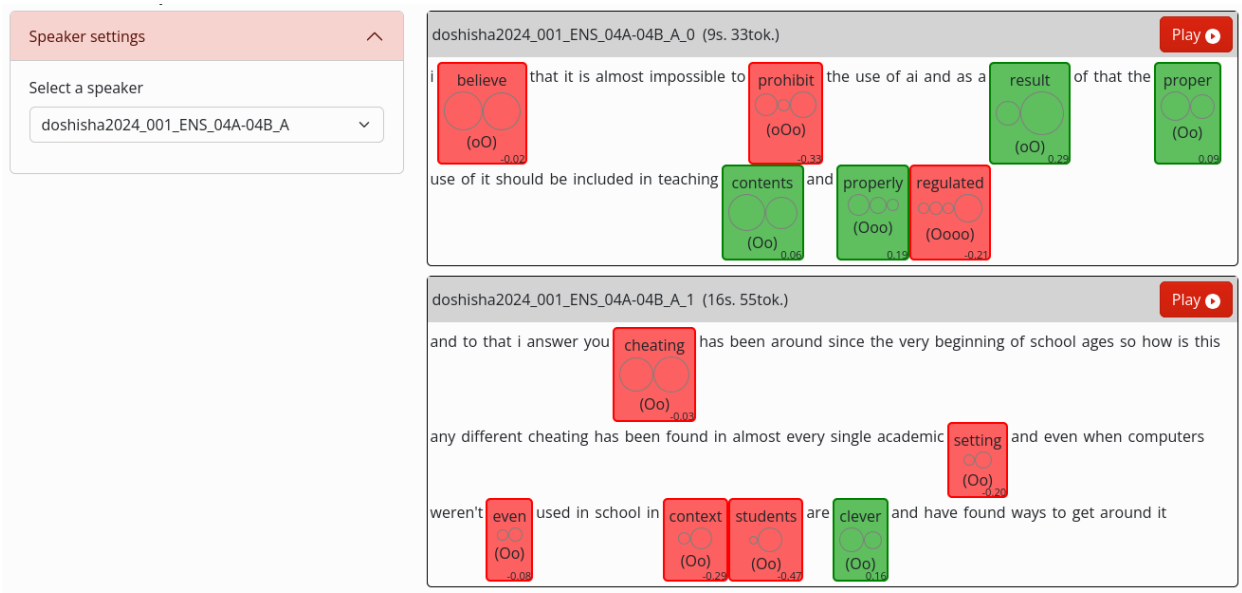


FIG. 5.10 : Aperçu de l'interface de visualitation de PLSPP, page Stress patterns

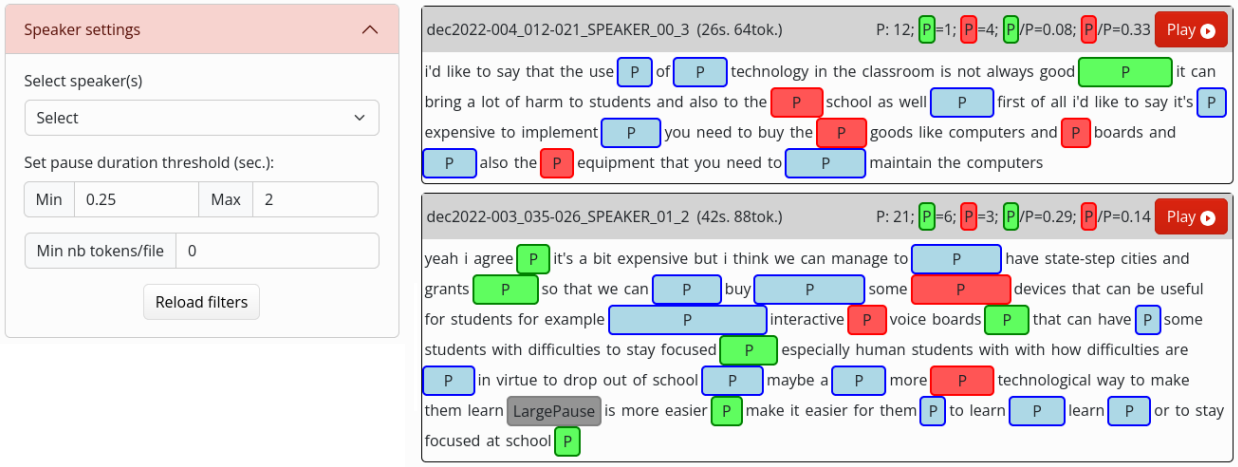


FIG. 5.11 : Aperçu de l'interface de visualitation de PLSPP, page Pause patterns

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 d'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)

Chapitre 6

Mesure de l'impact des pauses et de l'accent lexical sur la compréhensibilité du locuteur

Une fois que nous disposons d'un prototype permettant d'analyser automatiquement la distribution syntaxique des pauses et l'accentuation lexicale, il est possible d'évaluer leur impact sur la compréhensibilité du locuteur. Or, comme nous l'avons vu chapitre 2, cette compréhensibilité est un phénomène perceptif, et ne peut donc être évaluée qu'à travers le jugement d'auditeurs.

Nous proposons un protocole expérimental pour tenter de mesurer l'impact des pauses et de l'accentuation lexicale sur la compréhensibilité en temps réel. Nous nous inspirons du protocole employé par Nagle et al. (2019) pour évaluer dynamiquement la compréhensibilité du locuteur, qui eux-mêmes adaptent une méthode d'évaluation continue utilisée en psychologie cognitive (MacIntyre, 2012). Nous avons simplifié et adapté leur protocole pour permettre une évaluation de plus grande échelle, en *crowdsourcing*, et en privilégiant l'aspect quantitatif de l'approche.

Nous tenterons de répondre aux deux questions suivantes : Q1) Les auditeurs montrent-ils un comportement cohérent dans l'évaluation dynamique de la compréhensibilité malgré les variations inter- et intra-individuelles ? Q2) Une diminution de la compréhensibilité est-elle observable à la suite d'occurrences de pauses intra-syntagme ou de patterns accentuels inappropriés ?

Après avoir décrit notre protocole expérimental, les stimuli audio sélectionnés, et les participants recrutés pour l'expérience, nous présenterons la plateforme d'évaluation développée pour les besoins de l'expérience. Enfin, nous détaillerons les différents traitements effectués sur les données collectées, et les analyses réalisées.

6.1 Adaptation du protocole

Nous avons choisi de partir du protocole expérimental mis au point par Nagle et al. (2019). Celui-ci tente d'évaluer de manière dynamique le jugement de compréhensibilité, afin de pouvoir observer les fluctuations de ce jugement au fur et à mesure de l'écoute. Si Nagle et al. analysent ces fluctuations de manière globale dans une approche exploratoire, sans cibler de phénomène linguistique précis, nous proposons, pour notre part, d'exploiter cette méthode pour observer comment varie le jugement des participants spécifiquement à la suite de certaines pauses ou patterns accentuels. Plus concrètement, nous souhaitons observer si le jugement de compréhensibilité a tendance à diminuer à la suite de pauses de bas niveau (intra-syntagme, a priori disfluentes) et de patterns accentuels inappropriés, comparé au jugement mesuré à la suite de pauses de haut niveau (inter-proposition, a priori structurantes) et de patterns accentuels corrects.

Trois modifications majeures du protocole de Nagle et al. (2019) ont été réalisées. Pour permettre à un plus grand nombre d'évaluateurs de participer, nous avons opté pour une passation en ligne, sur une plateforme d'évaluation dédiée. Nous n'avons pas effectué de captation vidéo ni d'entretiens individuels comme c'est le cas dans le protocole original. L'expérimentation a ainsi été repensée pour permettre une passation en complète autonomie : elle a été simplifiée et raccourcie pour ne pas dépasser une durée globale de 35 minutes. Une rapide explication de la tâche est donnée à l'écrit en début d'expérience, suivie de trois questions pour vérifier le profil du participant, et d'une phase d'entraînement. La consigne reste écrite jusqu'à la fin de l'expérience. Après chaque stimulus, si l'évaluateur a été jugé trop peu actif, une *pop-up* de rappel s'ouvre avant le stimulus suivant.

La tâche d'évaluation elle-même a été simplifiée de manière à n'avoir plus qu'un seul bouton sur la page au lieu de deux, et donc une seule action possible. Il est simplement demandé à l'auditeur de cliquer sur le bouton dès qu'il sent qu'il fait un effort pour comprendre le locuteur, quelle que soit la raison. De plus, pour simplifier la tâche d'évaluation, il n'y a plus d'incrémentation du jugement comme c'est le cas sur le logiciel utilisé par Nagle et al. Ainsi, au lieu de varier entre 5 et -5, le jugement ne peut plus être que -1. Lorsque l'auditeur clique sur *start* au début de chaque stimulus, celui-ci démarre sans possibilité de mettre pause ni de réécouter. Chaque clic est enregistré sous la forme d'un *timestamp* correspondant à la position du curseur de lecture. À la fin de la lecture, il lui est demandé d'évaluer globalement la performance du locuteur en termes de qualité de prononciation, de fluidité, et de facilité de compréhension à l'aide de curseurs libres. Enfin, une question optionnelle est posée incitant l'évaluateur à expliciter les aspects de la prononciation de l'extrait entendu qui l'ont rendu difficile à comprendre, et à suggérer des conseils pour s'améliorer.

6.2 Sélection des stimuli

Afin de mesurer l'impact des différentes catégories de pauses et d'accentuation lexicale, il est nécessaire de présenter des stimuli audio contenant suffisamment d'occurrences de chacune d'elles pour pouvoir observer une tendance significative. Seize segments audio issus des analyses du corpus de locuteurs francophones ont ainsi été sélectionnés pour l'expérimentation. Les critères de sélection sont les suivants : 8 segments de parole présentant une grande proportion de pauses intra-syntagme et de mots au pattern accentuel inapproprié, et 8 segments présentant les conditions inverses. Par ailleurs, nous avons veillé à ce que les proportions B1/B2 et homme/femme soient respectées dans les deux groupes.

Pour pouvoir caractériser chaque segment en termes de fluence et de qualité d'accentuation, nous avons calculé deux scores par segment : la proportion de pauses de type intra-syntagme ($P_{p,WP}$, nombre de pauses intra-syntagme divisé par le nombre de pauses total) et un score de contraste accentuel moyen $\overline{C''}$. Ce score est une variante de $\overline{C}$ et $\overline{C'}$ présentés dans le chapitre précédent (*cf.* section 5.7), où C'' représente le degré de contraste entre la syllabe censée porter l'accent primaire P_s et la moyenne des autres syllabes $\overline{P_u}$, normalisée par la somme des deux valeurs pour obtenir une différence relative comprise entre -1 et 1. La formule est la suivante :

$$C''_w = \frac{P_{s,w} - \overline{P_{u,w}}}{P_{s,w} + \overline{P_{u,w}}} \tag{6.1}$$

Comme pour C, un score positif indique que la proéminence est mesurée sur la syllabe censée être accentuée, et une valeur proche des extrémités représente un contraste prosodique élevé entre la syllabe accentuée et les autres syllabes du mot. Il s'agit en réalité du premier score accentuel que nous avons mis au point. Celui-ci a par la suite évolué légèrement dans le cadre de l'étude de Kimura et al. (2024), présentée dans le chapitre précédent (score entre 0 et 1). Mais nous avons finalement choisi d'utiliser la simple différence de centiles (C) dans la partie résultats.[1]

En projetant les segments sur un plan défini par ces deux dimensions, on peut alors sélectionner les segments situés aux extrémités : P_{WP} élevé et score accentuel bas, et P_{WP} faible et score élevé (*cf.* figure 6.1). On appellera le premier groupe « *low* », et le second « *high* ». Le seuil de durée minimale des pauses est fixé à 250 ms et le nombre minimum de tokens à 60 pour éviter les segments trop courts.

[1] Il s'est avéré en effet que C' et C'' sont moins appropriés pour représenter le degré de contraste, car la différence entre une syllabe de valeur 95 et d'une autre de 85 donnera un score plus petit que la différence entre 15 et 5, tandis que le contraste C correspondant est toujours de 10 points.

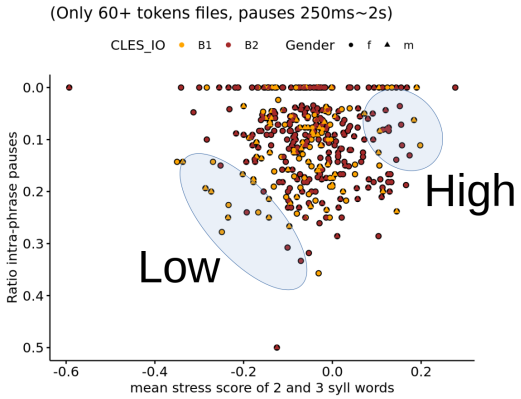


FIG. 6.1 : Choix des stimuli à partir des segments de plus de 60 tokens, projetés en fonction de P_{WP} et du score accentuel moyen $\overline{C''}$

	LOW		HIGH			LOW		HIGH	
	freq	%	freq	%		freq	%	freq	%
BC	59	29,9	73	39,5	**StressO**	1	1,4	22	31,9
BP	99	50,3	98	53	**StressΔ**	35	50	44	63,8
WP	39	19,8	14	7,6	**StressX**	34	48,6	3	4,3
total	**197**		**185**		**total**	**70**		**69**	

TAB. 6.1 : Nombre et proportion de pauses et de mots polysyllabiques de chaque catégorie dans les 16 segments sélectionnés

Les 16 segments sélectionnés sont produits par 15 locuteurs différents, les segments du groupe *low* sont produits par 4 locuteurs B1 et 4 locuteurs B2, ceux du groupe *high* par 3 locuteurs B1 et 5 locuteurs B2. La répartition homme/femme est respectivement de 7 pour 9. La durée des segments s'étend de 26 à 66 secondes (médiane à 38), et le nombre de tokens de 61 à 132 (médiane à 75), sans différence significative entre les groupes *low* et *high*.

Le tableau 6.1 présente le nombre et la proportion de pauses et de mots polysyllabiques de chaque catégorie. Dans le cas de l'accentuation lexicale, les mots sont divisés en trois catégories : *StressO* pour les mots dont le score est élevé ($C'' \geq 0{,}2$), *Stress*Δ pour les mots au contraste peu marqué ($-0{,}2 \leq C'' < 0{,}2$) et *StressX* pour les mots au contraste négatif fort ($C'' < -0{,}2$).

Pour s'assurer que l'annotation des pauses et des patterns accentuels est de qualité acceptable, une vérification manuelle a été effectuée sur la moitié des segments, comprenant 193 pauses et 89 mots polysyllabiques. Les pauses dont l'alignement temporel et la catégorie syntaxique sont corrects totalisent 82,4 %, et les mots polysyllabiques correctement reconnus et alignés au niveau du mot et des syllabes totalisent 82,0 %.

6.3 Sélection des participants

Les participants ont été recrutés sur la plateforme britannique Prolific[2]. Cette plateforme permet de mettre en relation des chercheurs ou des entreprises avec des personnes de profils variés pour participer à des expérimentations en ligne. Les critères de recrutements que nous avons choisis sont les suivants : être de langue maternelle anglaise, vivre en Angleterre au moment de l'expérience, ne pas avoir déclaré de compétences en langue étrangère (critère "*English speaking monolingual*" sur la plateforme) et respecter une balance de genre. Une rétribution financière a été fixée à hauteur de £10.86 de l'heure, soit £5.25 pour une durée prévue de 35 min (6,14 € (12,7 €/h) au moment de l'expérimentation, en février 2024).

Soixante personnes ont participé à l'expérience, 30 femmes, 30 hommes, de 25 à 72 ans (moyenne à 44, écart type de 12). Seuls les participants qui ont cliqué au moins une fois dans toute l'expérience, et n'ont pas concentré plus de 50 % de leurs clics sur un seul segment ont été retenus pour les analyses.[3]

6.4 Développement de *Dynamic Rater*

Une application web appelée *Dynamic Rater*[4] a été développée spécifiquement pour les besoins de cette étude. Cet outil s'inspire largement du logiciel Idiodynamic de MacIntyre (2012), mais propose un protocole d'évaluation plus simple et permet une passation à distance et en autonomie. Notre application se compose de 4 vues principales : une page d'accueil avec la présentation du déroulement de l'expérimentation, une page de questionnaire linguistique, la page d'expérimentation, et la page de fin d'expérimentation. Chaque page est décrite en détail et illustrée en annexe G. Nous nous concentrons ici sur la fonctionnalité d'évaluation dynamique de la compréhensibilité de la page d'expérimentation.

Après une phase d'entraînement, les 16 stimuli audio sont présentés dans un ordre aléatoire aux participants. Lors de l'écoute, ces derniers doivent signaler en cliquant sur le bouton "*I'm struggling*", qu'ils perçoivent une difficulté à comprendre le locuteur, quelque soit la raison (*cf.* figure 6.2). À chaque clic, une barre verticale s'affiche sur la waveform à l'endroit du curseur de lecture, de manière à visualiser l'historique des clics et confirmer au participant que l'action a bien été prise en compte.

[2] https://www.prolific.com/

[3] Trois participants supplémentaires ont été retirés des analyses car leur activité pendant l'expérimentation était trop limitée ou jugée anormale.

[4] Code source : https://gricad-gitlab.univ-grenoble-alpes.fr/lidilem/dynamic-rater

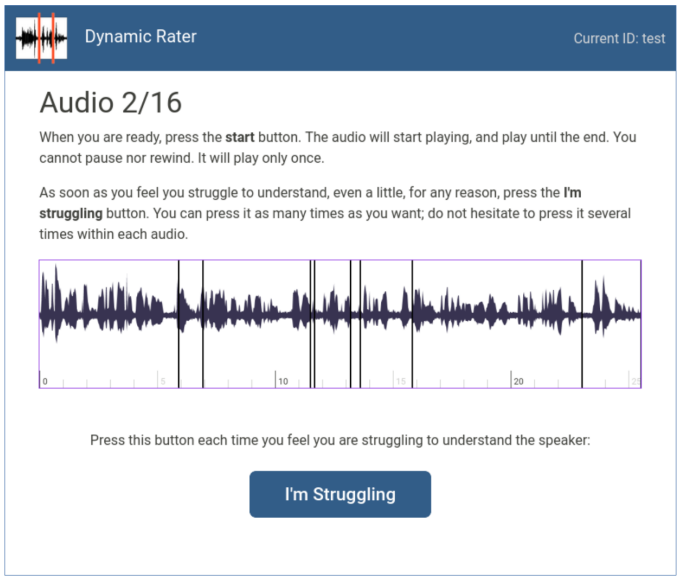


FIG. 6.2 : Aperçu de l'écran d'évaluation dynamique de Dynamic Rater

Le participant peut cliquer autant de fois qu'il veut et à tout moment, mais ne peut pas éditer les clics produits, ni mettre pause ou réécouter l'audio. Une fois la lecture terminée, les trois curseurs d'évaluation globale ainsi que le champ texte libre s'affichent en dessous de la waveform (les clics restent visibles). Il est nécessaire de modifier la valeur du troisième curseur *(Overall easiness to understand)* pour pouvoir valider et passer au stimulus suivant. À chaque validation, la liste des clics et l'évaluation globale sont envoyés au serveur, afin d'enregistrer les résultats au fur et à mesure.

6.5 Traitement des données

En fin d'expérience, nous avons donc 16 segments audio auxquels sont associés pour chaque évaluateur une liste de *timestamps* correspondant aux clics produits, trois scores globaux numériques et un commentaire textuel optionnel. Nous avons utilisé plusieurs mesures pour évaluer la cohérence et la fiabilité des évaluations globales : le coefficient de corrélation intra-classe (*2-way random model* (ICC2k) du package R psych v2.4.1) pour vérifier la consistance des évaluations entre les évaluateurs, et l'alpha de Cronbach pour évaluer la cohérence interne des évaluations sur l'ensemble des participants. Les deux coefficients ont été calculés à partir des scores bruts de chacune des trois dimensions. Les scores sont ensuite standardisés (z-scores) de manière à les rendre comparables.

Nous proposons d'abord d'analyser les résultats de l'évaluation globale des enregistrements. Le nombre limité d'enregistrements permet difficilement d'envisager une régression linéaire pour calculer un coefficient de corrélation. Aussi, nous proposons de diviser les segments en deux groupes pour chaque dimension : ceux qui se situent en dessus et ceux qui se situent en-dessous de la fréquence médiane des pauses intra-syntagme (nombre d'occurrences par token), des pauses inter-proposition et du score accentuel moyen. Les deux distributions sont ensuite comparées à l'aide du test non paramétrique Wilcoxon-Mann-Whitney et de la taille d'effet avec le delta de Cliff. Nous ferons la même chose avec le nombre total de clics normalisés pour voir comment celui-ci évolue globalement en fonction des enregistrements.

Pour l'évaluation dynamique, nous avons calculé la somme des clics par locuteur sur une fenêtre glissante d'une seconde. Afin d'éviter que les « cliqueurs compulsifs », comme les appellent Nagle et al. (2019), ne couvrent les clics des évaluateurs moins actifs, nous proposons d'y soustraire le nombre de clics moyen par minute par locuteur. Nous appellerons ces clics normalisés *m-clics*. Le calcul est effectué de la manière suivante :

$$M_w = \sum_{r=1}^{R} \left(C_{r,w} - \overline{C_r} \right) \tag{6.2}$$

avec M_w le nombre de m-clics dans la fenêtre w, R le nombre d'évaluateurs, $C_{r,w}$ le nombre de clics de l'évaluateur r dans la fenêtre w, et $\overline{C_r}$ la fréquence moyenne de clics de r. Cette normalisation permet de centrer les valeurs autour de 0, et ainsi de considérer les valeurs positives comme anormalement élevées, et les valeurs négatives comme inférieures à la moyenne. Concrètement, cela ne fait que centrer la courbe des patterns de clics sur 0, mais c'est un moyen intéressant de définir un seuil à partir duquel considérer les pics de clics.

La fréquence de m-clics sur les 5 secondes suivant chaque type de pause et de pattern accentuel est ensuite analysée pour déterminer si les clics ont tendance à augmenter, stagner ou diminuer à la suite de l'événement. Le test de rangs non-paramétrique est à nouveau utilisé pour vérifier la significativité de la différence entre la distribution de valeurs à la suite des pauses inter-proposition et intra-syntagme, et des mots en fonction de leur accentuation.

Conclusion

Nous avons adapté le protocole d'évaluation dynamique de la compréhensibilité de Nagle et al. (2019) pour l'utiliser en *crowd-sourcing* et obtenir un plus large

échantillon d'évaluations. Seize segments de parole issus du corpus CLES-FR ont été présentés à 60 auditeurs anglophones natifs. À l'aide d'une interface développée pour l'expérience, les participants ont dû signaler par un clic chaque fois qu'ils percevaient qu'ils faisaient un effort pour comprendre le locuteur. Cette activité de clics, une fois normalisée, représente alors une mesure dynamique de la compréhensibilité du locuteur.

Cette évaluation devrait nous permettre d'observer si les pauses de bas niveau syntaxique, et les patterns accentuels inattendus ont tendance à faire augmenter le niveau de difficulté de compréhension perçu par les auditeurs. Les résultats obtenus sont décrits chapitre 9.

Troisième partie

Résultats & discussion

Chapitre 7

Évaluation du système

Ce chapitre présente les résultats obtenus lors de l'évaluation des différents modules de PLSPP. Nous avons souhaité tester les performances de chaque module de manière isolée, afin de minimiser l'introduction d'erreurs résultant de traitements effectués en amont. Chaque étape d'évaluation a nécessité l'utilisation de données annotées manuellement, permettant une comparaison entre les annotations automatiques produites et une référence.

Pour évaluer la qualité de la segmentation en locuteurs et de la reconnaissance de la parole, nous avons exploité une portion du corpus CLES-FR, dont la transcription et la segmentation en locuteurs ont été vérifiées et corrigées manuellement (*cf.* chapitre 4.6). Nous nous référerons à ce corpus sous le nom de *CLES-gold*. En revanche, l'évaluation de la précision de l'alignement mot-signal a requis un corpus de référence différent, car l'alignement des mots dans le corpus CLES-gold n'a pas été vérifié. À cette fin, nous avons utilisé un corpus de parole spontanée de locuteurs francophones, issu de l'étude de Frost et al. (2024).

L'évaluation de la détection et de l'étiquetage des pauses s'est révélée plus difficile, car les corpus disposant d'une annotation syntaxique des pauses en parole spontanée sont peu nombreux. Nous avons utilisé ici un corpus de dialogues spontanés entre apprenants slovaques de l'anglais, gracieusement fourni par l'université de Nitra en Slovaquie (Mareková & Beňuš, 2024).

Enfin, l'analyse des performances d'annotation de l'accent lexical a été réalisée en la confrontant à plusieurs références : la perception d'auditeurs natifs (Kimura et al., 2024), la conscience phonologique de locuteurs natifs et non natifs (Sugahara et al., 2024), ainsi que les résultats obtenus sur des données de parole native contrôlée (Nakanishi & Coulange, 2024).

7.1 Modules de prétraitements

1.1 Segmentation en locuteurs

Pour évaluer la précision de la diarisation en locuteurs obtenue grâce au premier module de PLSPP, nous avons d'abord calculé le taux d'erreur de diarisation (DER) par binôme sur les fichiers de sortie de Pyannote après avoir analysé les 20 discussions du corpus CLES-gold.

Sur les 20 enregistrements du corpus CLES-gold, totalisant 2 h 58 min 38 s de parole, le DER moyen obtenu est de 19,42 % (médiane : 13,21 %, *cf.* figure et tableau 7.1). Ce DER moyen correspond à peu près au taux d'erreur de 18,9 % obtenu par Pyannote sur le corpus d'interactions en réunions professionnelles AMI-IHM *(Augmented Multi-Party Interaction - Individual Headset Microphone)* (Bredin, 2023).

On remarque que deux discussions obtiennent un DER particulièrement élevé (69 % et 64 %), contrastant avec les 18 autres dont la moyenne est de 14,23 % (médiane 10,78 %, min 4 %, max 30 %). On peut constater que le fichier qui obtient un DER de 69 % présente une proportion importante de parole non détectée (59,3 % du temps d'enregistrement). Il semblerait que la voix des locuteurs n'ait pas été correctement détectée par Pyannote pour une raison que nous ne sommes pas parvenus à identifier. Dans le cas du fichier dont le DER est 64 %, il s'agit cette fois d'une importante confusion entre les locuteurs. Comme indiqué dans le tableau 7.1, 39,9 % du temps d'enregistrement de ce fichier n'est pas annoté avec le bon locuteur, alors que la moyenne de durée de confusion sur les autres fichiers n'est que de 3,1 %. Malgré cela, sur l'ensemble des fichiers, on constate que la durée de parole qui n'est pas attribuée au bon locuteur reste limitée : 5,1 % sur la durée totale d'enregistrement, et 4,9 % en moyenne par enregistrement.

Nous avons ensuite calculé l'indice d'interférence I_L de chaque locuteur à partir des segments de parole extraits par PLSPP (durée supérieure ou égale à 8 s). Cet indice donne la proportion de temps de parole d'un locuteur L correspondant en réalité à la parole de l'interlocuteur. C'est un moyen de quantifier la présence de l'interlocuteur dans les segments de parole attribués au locuteur, et qui sera donc potentiellement à l'origine de mauvaises attributions de patterns de pauses ou d'accents lexicaux.

L'indice d'interférence moyen obtenu est de 2,98 % (médiane 1,57 %, min 0 %, max 29,26 %, *cf.* figure 7.2 et tableau 11.2). Pour la discussion qui a obtenu le DER le plus élevé, l'indice d'interférence pour les deux locuteurs reste assez bas (4,27 % et 0,27 %), mais on voit que la durée de parole extraite est réduite (73 s et 64 s, contre 235 s en moyenne pour les autres locuteurs, *cf.* tableau en annexe E). Les deux locuteurs de l'enregistrement avec un DER de 64 % présentent quant à eux un pourcentage

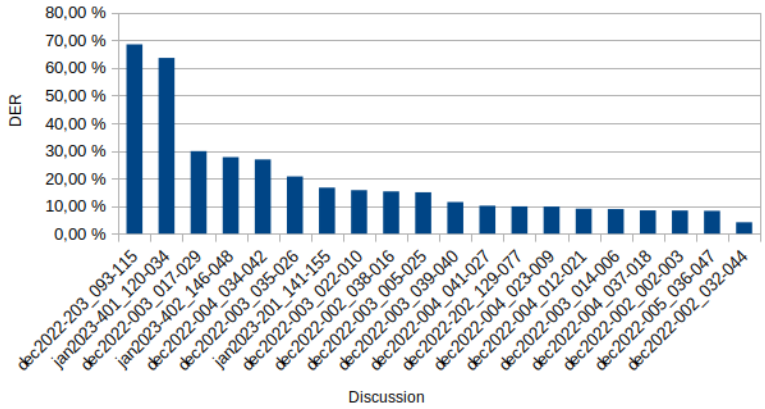


FIG. 7.1 : Taux d'erreurs de diarisation (DER) sur le corpus CLES-gold

	File	DER	Missed Speech		False Alarm		Confusion		Total Duration
1	dec2022-203_093-115	68,54 %	330,98	(59,3 %)	24,31	(4,4 %)	24,33	(4,4 %)	00 :09 :18
2	jan2023-401_120-034	63,62 %	118,62	(19,1 %)	87,94	(14,2 %)	247,92	(39,9 %)	00 :10 :21
3	dec2022-003_017-029	29,87 %	35,15	(8,3 %)	45,92	(10,8 %)	42,69	(10,0 %)	00 :07 :05
4	jan2023-402_146-048	27,70 %	130,13	(22,4 %)	17,02	(2,9 %)	8,7	(1,5 %)	00 :09 :40
5	dec2022-004_034-042	26,85 %	47,69	(9,5 %)	50,6	(10,1 %)	29,32	(5,8 %)	00 :08 :22
6	dec2022-003_035-026	20,72 %	61,94	(10,1 %)	33,41	(5,4 %)	31,33	(5,1 %)	00 :10 :15
7	jan2023-201_141-155	16,66 %	77,73	(13,8 %)	8,71	(1,5 %)	7,36	(1,3 %)	00 :09 :25
8	dec2022-003_022-010	15,75 %	26,3	(8,4 %)	12,04	(3,9 %)	10,03	(3,2 %)	00 :05 :12
9	dec2022-002_038-016	15,30 %	37,91	(6,9 %)	22,81	(4,2 %)	22,72	(4,1 %)	00 :09 :08
10	dec2022-003_005-025	14,97 %	32,21	(6,0 %)	23,83	(4,4 %)	23,83	(4,4 %)	00 :08 :56
11	dec2022-003_039-040	11,45 %	33,68	(4,5 %)	26,27	(3,5 %)	25,44	(3,4 %)	00 :12 :30
12	dec2022-004_041-027	10,11 %	34,29	(8,0 %)	4,83	(1,1 %)	3,65	(0,9 %)	00 :07 :06
13	dec2022-202_129-077	9,87 %	39,25	(7,5 %)	7,14	(1,4 %)	4,14	(0,8 %)	00 :08 :45
14	dec2022-004_023-009	9,80 %	24,41	(4,6 %)	13,56	(2,6 %)	13,56	(2,6 %)	00 :08 :46
15	dec2022-004_012-021	9,03 %	24,94	(4,9 %)	10,88	(2,1 %)	9,98	(2,0 %)	00 :08 :29
16	dec2022-003_014-006	8,91 %	26,22	(4,2 %)	15,28	(2,5 %)	13,28	(2,2 %)	00 :10 :17
17	dec2022-004_037-018	8,43 %	17,31	(4,0 %)	9,91	(2,3 %)	7,91	(1,8 %)	00 :07 :12
18	dec2022-002_002-003	8,38 %	18,93	(3,4 %)	15,65	(2,8 %)	10,51	(1,9 %)	00 :09 :11
19	dec2022-005_036-047	8,26 %	11,94	(2,2 %)	20,84	(3,8 %)	11,12	(2,0 %)	00 :09 :03
20	dec2022-002_032-044	4,09 %	17,56	(3,0 %)	2,94	(0,5 %)	2,94	(0,5 %)	00 :09 :36
			1147,19	**(10,7 %)**	**453,89**	**(4,2 %)**	**550,76**	**(5,1 %)**	**02 :58 :38**

TAB. 7.1 : Taux d'erreurs de diarisation (DER) par discussion sur le corpus CLES-gold. Les durées de parole manquée (Missed Detection), fausse alerte et confusion de locuteur sont données en secondes. Les pourcentages sont calculés par rapport à la durée totale de chaque enregistrement.

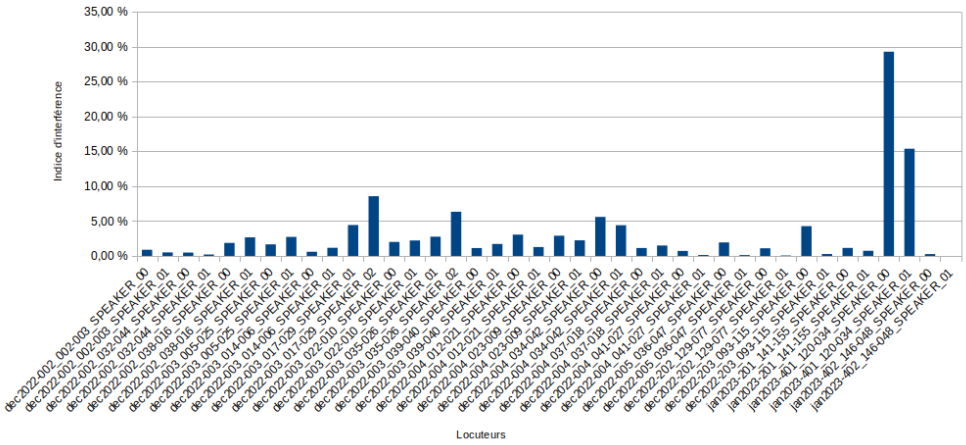


FIG. 7.2 : Indice d'interférence par locuteur (sur l'ensemble des segments de durée supérieure ou égale à 8 s)

de confusion important. Ces deux locuteurs ont un indice d'interférence de respectivement 29,26 % et 15,36 %. À l'écoute de l'enregistrement, on constate effectivement de nombreux chevauchements entre les deux locuteurs, à l'origine de ces confusions d'étiquetage.

Nous considérons que la proportion d'erreur d'identification du locuteur dans les segments de paroles extraits par PLSPP est suffisamment basse pour ne pas affecter sensiblement les résultats de nos analyses.

1.2 Reconnaissance automatique de la parole

Pour évaluer la qualité de la reconnaissance automatique de la parole avec le système Whisper et le modèle *base.en*[1], nous avons calculé le taux d'erreur de mots (WER) pour chaque segment du corpus CLES-gold (n = 349).

Le WER est la métrique de référence pour évaluer les performances des systèmes de reconnaissance de la parole. Elle permet une comparaison directe entre différents modèles ou systèmes analysant un même corpus. Toutefois, il est essentiel de noter que la précision de la reconnaissance est fortement dépendante du type de parole analysée (Evain, 2024). Par exemple, sur le corpus de textes lus LibriSpeech (environ 1000 h d'enregistrements de livres audio en anglais, Panayotov et al., 2015), le modèle *base.en* obtient un taux d'erreur très faible (4,2 %). En revanche, sur le corpus de réunions professionnelles AMI-IHM cité précédemment, le taux d'erreur s'élève à

[1] https://huggingface.co/openai/whisper-base.en

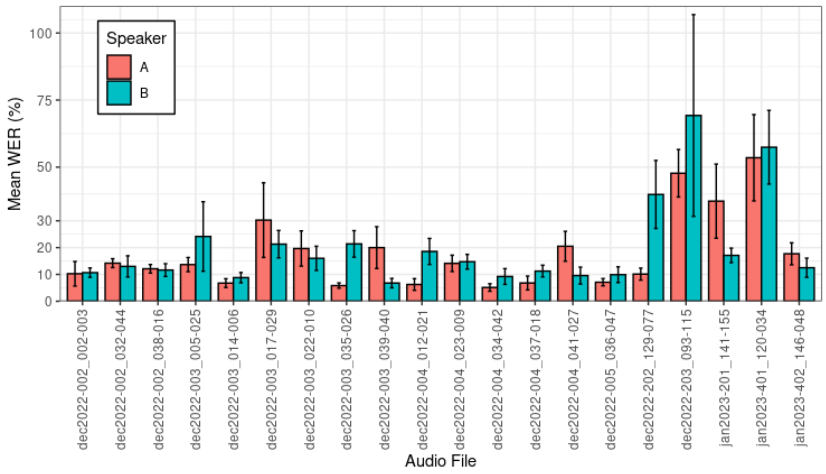


FIG. 7.3 : Taux d'erreur de mots (WER) moyen par locuteur avec barre d'erreur indiquant le degré de variabilité selon les segments du même locuteur (écart type)

20,5 % (Radford et al., 2022, p. 22 pour les WER obtenus par différents modèles de Whisper sur divers corpus). Ainsi, la précision du modèle dépend de la situation de parole, et notamment du degré de spontanéité et de la présence de chevauchements.

Pour notre corpus CLES-gold, le WER moyen de l'ensemble des 349 segments est de 19,0 % (médiane : 11,0 %). Le WER par locuteur, calculé par concaténation des segments, est de 16,76 % (médiane : 13,5 %) avec une amplitude allant de 6 % à 63 %. La figure 7.3 et le tableau en annexe F indiquent le WER moyen par locuteur, ainsi que le nombre de substitutions, de délétions et d'insertions.

Les quatre locuteurs qui obtiennent les WER les plus élevés sont ceux des deux enregistrements au DER élevé, identifiés dans la section précédente. On constate notamment que le locuteur *jan2023-401_120-034_SPEAKER_00* a un très grand nombre d'insertions (135, contre 16 en moyenne pour les autres). En observant de plus près les segments de ce locuteur, on trouve des phénomènes d'hallucination de Whisper, dus à des répétitions successives de mots. Par exemple, pour le segment *jan2023-401_120-034_SPEAKER_00_2* (57 mots, WER : 186 %, $S = 51$, $D = 0$, $I = 55$), la transcription de référence commence par "*you can you can you can write wait wait wait wait wait wait we have to write an article* [...]", mais la transcription automatique ne contient que le mot "*wait*" répété 112 fois. Dans d'autres cas, il s'agit d'une interférence importante de l'interlocuteur, aboutissant à un grand nombre d'insertions et de substitutions. Pour les deux locuteurs de la discussion *dec2022-203_093-115*, il s'agit principalement d'une mauvaise reconnaissance générale de la parole du SPEAKER_00, combinée à la quantité de parole limitée pour ce binôme.

En excluant ces quatre locuteurs, le WER moyen des 36 locuteurs restants est de 12,91 % (médiane : 11,0 %, min : 6 %, max : 28 %). Ces observations mettent en évidence une certaine variabilité du WER selon les locuteurs, mais les résultats restent globalement satisfaisants compte tenu de la nature de la parole analysée, spontanée et L2. Ces taux d'erreurs pourraient toutefois être réduits en utilisant un modèle de langue avec plus de paramètres, comme *medium.en*[2] (769 millions de paramètres contre 74 millions pour *base.en*).

1.3 Alignement mot-signal

Pour évaluer la précision de l'alignement temporel des mots de la transcription orthographique au signal de parole, nous avons mesuré les performances du modèle utilisé pour l'alignement par PLSPP, Wav2Vec2.0, en nous appuyant sur deux enregistrements issus de l'étude de Frost et al. (2024). Ces enregistrements incluent un alignement entièrement vérifié, pouvant ainsi servir de référence. Comme expliqué dans le chapitre 5.1.2, deux métriques ont été calculées :

- **Le score de précision** (P), qui mesure la proportion de durée alignée automatiquement correspondant à l'alignement de référence ;
- **Le score de rappel** (R), qui mesure la proportion de l'alignement de référence correctement identifié par l'aligneur.

Nous souhaitons en particulier obtenir un score de précision élevé, car il est important que l'alignement corresponde précisément à ce qui est dit dans l'enregistrement, même si certaines portions ne sont pas alignées. Le score de rappel, quant à lui, renseigne sur la proportion de la référence qui a été correctement alignée.

La comparaison entre l'alignement de référence (REF) et l'alignement automatique (AUTO) s'effectue au niveau des mots (intervalles pleins), en excluant les intervalles vides. Pour chaque mot aligné automatiquement, nous avons mesuré la durée correspondant au même mot dans REF, représentée par un segment vert dans la figure 7.4.

Sur les 442 s d'enregistrement du corpus de référence, Wav2Vec a aligné 239 s, contre 315 s pour l'alignement de référence. Parmi ces 239 s alignées par Wav2Vec, 218 s correspondent effectivement à REF, ce qui donne un score de précision élevé ($P = 0{,}91$). Cependant, la figure 7.4 montre que les segments alignés par Wav2Vec sont souvent légèrement plus courts que ceux de REF, ce qui reflète une tendance de l'aligneur à réduire la durée des mots par rapport à un alignement manuel. Cette limitation se traduit par un score de rappel plus faible ($R = 0{,}69$).

[2] https://huggingface.co/openai/whisper-medium

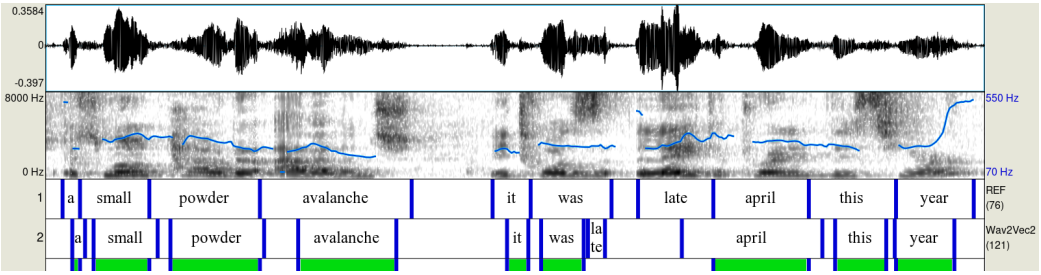


FIG. 7.4 : Visualisation de la correspondance entre l'alignement de référence (REF) et l'alignement automatique de Wav2Vec2.0. Les intervalles verts indiquent les portions d'alignement automatique correctes.

	Word Duration	Correct Duration	P	R
Reference	315	-	-	-
Wav2Vec 2.0	239	218	**0,91**	0,69
WebMaus 3.4	322	270	0,84	**0,86**
MFA 2.0	329	245	0,74	0,78

TAB. 7.2 : Scores de précision et de rappel obtenus par les différents systèmes d'alignement automatique évalués. Word Duration indique la durée totale d'alignement (en secondes).

Nous avons également évalué deux autres systèmes d'alignement sur le même corpus : WebMaus v3.4 et Montreal Forced Aligner v2.0. Ces systèmes présentent des scores de rappel plus élevés, respectivement $R = 0{,}86$ et $R = 0{,}78$, indiquant qu'une plus grande proportion de REF est alignée. Cependant, leurs scores de précision sont inférieurs, $P = 0{,}84$ pour WebMaus et $P = 0{,}74$ pour Montreal Forced Aligner, ce qui signifie qu'une part plus importante des alignements automatiques ne correspond pas à REF (voir tableau 7.2).

7.2 Annotation des pauses

L'évaluation de l'annotation des pauses a consisté en une comparaison entre les annotations automatiques produites par PLSPP et des annotations manuelles réalisées sur un corpus de parole présentant des caractéristiques similaires à celles des données analysées dans le cadre de notre étude sur la parole spontanée.

Ce corpus est composé de 72 dialogues spontanés enregistrés auprès de 24 binômes de locuteurs slovacophones (Mareková & Beňuš, 2024). Dans chaque dialogue, un des participants guide son interlocuteur en décrivant un itinéraire sur une carte, sans que les locuteurs puissent se voir. Les tours de parole sont transcrits et alignés au signal acoustique, et les pauses sont annotées selon plusieurs catégories : pauses

	N	d_{min}	d_{med}	d_{max}
btw	1804 (23,53%)	147	484	2470
mid	1937 (25,26%)	200	442	3548
fp	1416 (18,47%)	67	398	1492
p	2511 (32,75%)	11	319	2760
Total	**7668**			

Tab. 7.3 : *Nombre de pauses annotées manuellement présentes dans les 983 segments de parole extraits du corpus de Mareková et Beňuš (2024), et durée minimum, médiane et maximum (ms)*

inter-propositions (btw) et intra-propositions (mid), ainsi que pauses pleines (fp) et pauses inter-tours (p).

Nous avons analysé les 72 enregistrements avec l'ensemble des modules de PLSPP, incluant la segmentation en locuteurs, l'extraction des segments de parole, la transcription et l'alignement au signal, ainsi que les analyses syntaxiques et l'annotation des pauses. Ces processus ont permis d'extraire et d'analyser 983 segments. L'évaluation des annotations des pauses repose sur ces 983 segments.

Ces segments contiennent 7 668 pauses annotées manuellement[3]. Elles se répartissent comme suit : 24 % sont des pauses inter-propositions, 25 % intra-propositions, 18 % sont pleines, et 33 % inter-tours (*cf.* tableau 7.3). Il convient de noter que chaque pause est associée à une seule de ces catégories : les pauses pleines et inter-tours ne sont donc pas annotées en fonction de leur position syntaxique dans l'énoncé. Cependant, il est possible que deux pauses de types différents se succèdent, par exemple une pause pleine suivie d'une pause intra-propositionnelle.

L'annotation automatique réalisée par PLSPP a quant à elle identifié un total de 8 194 pauses d'une durée comprise entre 180 ms et 2 s[4]. La figure 7.5 illustre la comparaison des annotations. Les tiers 1 à 7 proviennent du corpus de Mareková et Beňuš (2024), tandis que la tier 8 représente la transcription alignée par PLSPP, où les pauses sont signalées en vert.

Pour chaque pause identifiée par PLSPP, nous avons calculé la proportion de chaque type de pause annotée manuellement sur la portion correspondante du signal. Par exemple, dans le cas illustré, la première pause détectée par PLSPP correspond à une pause inter-tour (p) pour 70 %, la deuxième ne correspond à aucune pause annotée manuellement, et la dernière correspond simultanément à deux pauses inter-

[3]Les pauses situées en début et en fin de segments ont été exclues de l'analyse, car elles peuvent avoir été tronquées lors de l'extraction.

[4]Les pauses situées en début et en fin de segments ont été ignorées ici également. Les seuils de durées minimum (180 ms) et maximum (2 s) ont été retenus car ils correspondent aux critères utilisés majoritairement dans les analyses présentées au chapitre suivant.

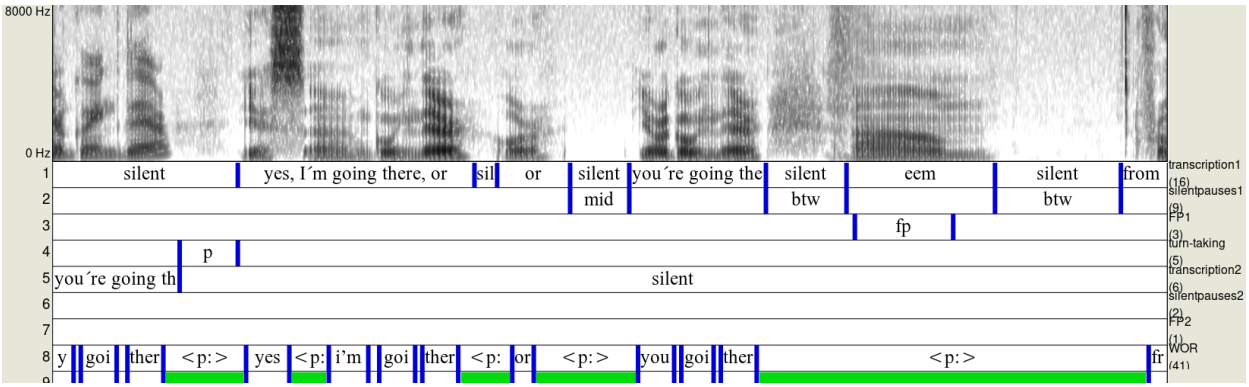


*FIG. 7.5 : Illustration de la comparaison de l'annotation des pauses entre PLSPP et l'annotation manuelle de Mareková et Beňuš (2024). La tier 1 correspond à la transcription manuelle du premier locuteur, la tier 2 aux pauses silencieuses (inter et intra-proposition), la tier 3 aux pauses pleines et la tier 4 aux pauses inter-tour. Les tiers 5, 6 et 7 correspondent aux 1, 2 et 3 pour le second locuteur. La tier 8 correspond à la transcription et l'alignement automatique de PLSPP. Les zones vertes indiquent les pauses identifiées par PLSPP (*180 ms-2 s*).*

	N	btw	mid	fp	p	silent	Ø
BC	2650	**662 (24,98%)**	207 (7,81%)	144 (5,43%)	**914 (34,49%)**	335 (12,64%)	388 (14,64%)
WC	5544	771 (13,91%)	**1149 (20,73%)**	424 (7,65%)	1001 (18,06%)	889 (16,04%)	**1310 (23,63%)**
Total	**8194**						

TAB. 7.4 : Type de pause annotée manuellement identifié pour chaque pause inter-proposition (BC) et intra-proposition (WC) annotée par PLSPP

propositions (btw) pour 53 % et à une pause pleine (fp) pour 25 %. À partir de ces proportions, nous avons déterminé le type de pause majoritaire pour chaque pause PLSPP, afin de faciliter la comparaison entre les annotations automatiques et manuelles.

La majorité des 2 650 pauses inter-propositions (BC) détectées par PLSPP correspondent à des pauses annotées comme inter-tours (p, 34 %), tandis que 25 % correspondent à des pauses inter-propositions (btw) et 8 % à des pauses intra-propositions (mid) (*cf.* tableau 7.4). Environ 15 % d'entre elles ne correspondent à aucune annotation manuelle (Ø), et 13 % correspondent à des intervalles notés "*silent*" sans annotation explicite de pause (comme illustré par la troisième pause PLSPP dans la figure 7.5).

Concernant les 5 544 pauses intra-propositions (WC) détectées par PLSPP, les correspondances avec les annotations manuelles sont les suivantes : 24 % ne correspondent à aucune annotation (Ø), 21 % à des pauses intra-propositions (mid), 18 % à

des pauses inter-tours (p), 16 % à des intervalles “silent” sans annotation explicite, 14 % à des pauses inter-propositions (btw), et 8 % à des pauses pleines (fp).

Ces résultats révèlent des performances mitigées. La faible précision des annotations automatiques pour les pauses inter- et intra-proposition (respectivement 25 % et 21 %) soulève plusieurs interrogations. Cette imprécision pourrait être liée à des erreurs dans l'étiquetage syntaxique, mais également au fait qu'un grand nombre de pauses sont annotées comme inter-tours dans le corpus de référence. Le choix du corpus de comparaison n'était peut-être pas optimal, dans la mesure où la présence de multiples catégories de pauses et d'un grand nombre de courtes réactions de l'interlocuteur, non éliminées malgré l'extraction des segments par PLSPP, pourrait compliquer la comparaison. Une évaluation spécifique de la précision de l'analyse grammaticale par constituant permettrait de compléter la vue d'ensemble pour mieux identifier les limites de l'annotation automatique des pauses.

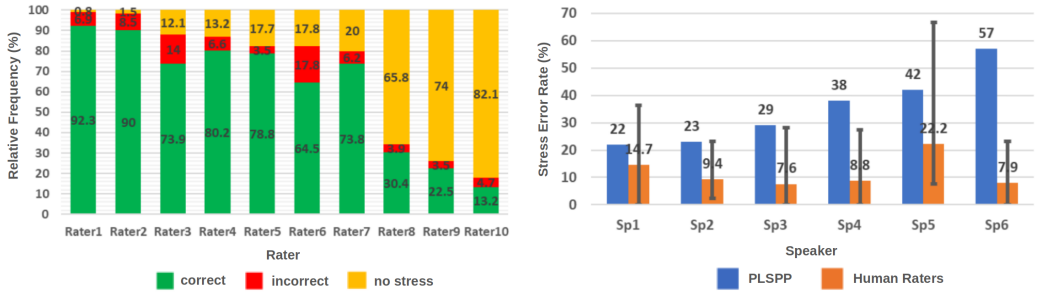
Enfin, le choix de n'inclure que les pauses PLSPP d'une durée comprise entre 180 ms et 2 s repose sur les critères retenus pour l'analyse des annotations des corpus CLES. Toutefois, un ajustement du seuil de durée minimum à 250 ms améliore légèrement les performances : la précision des pauses inter-proposition atteint alors 29 %, tandis que celle des pauses intra-proposition passe à 25 %. Par ailleurs, la proportion de pauses non annotées dans le corpus de référence diminue, atteignant 8 % et 13 % respectivement pour les pauses inter- et intra-proposition. On peut donc penser qu'un certain nombre de pauses courtes n'ont pas été annotées manuellement.

7.3 Annotation de l'accent lexical

3.1 Évaluation perceptive par des auditeurs natifs

Pour évaluer la qualité de l'annotation de l'accentuation lexicale par PLSPP, nous avons comparé les résultats obtenus avec ceux d'une annotation manuelle, réalisée par des locuteurs anglophones natifs. Nous présentons ici les résultats obtenus grâce au travail de recherche d'un étudiant du *Spoken Language Processing Laboratory* de l'université Dōshisha, détaillés plus longuement par Kimura et al. (2024).

Dix évaluateurs ont été recrutés pour annoter manuellement la syllabe qu'ils percevaient comme accentuée dans les mots polysyllabiques lexicaux d'un texte lu par six locuteurs anglophones de langue maternelle japonaise. Chaque mot cible a ensuite été classé comme « correct », « incorrect » ou « sans accent », selon les annotations des participants.



(a) Distribution des annotations de l'accent lexical par les 10 évaluateurs à travers les six enregistrements

(b) Taux d'erreur d'accentuation par locuteur selon PLSPP et les 10 évaluateurs humains. La barre verticale indique l'amplitude des scores humains

FIG. 7.6 : Figures issues de Kimura et al. (2024, p. 675)

La figure 7.6a présente la distribution des annotations par évaluateur. Une forte variabilité apparaît quant à la perception de la présence ou de l'absence d'un accent lexical. Les évaluateurs 8, 9 et 10 rapportent une absence d'accent pour la majorité des mots (respectivement 66, 74 et 82 %), tandis que cinq autres identifient un accent sur 80 à 88 % des mots, et les deux premiers annotateurs sur 98 et 99 %. Il semblerait que le seuil de sensibilité pour percevoir la position de l'accent soit un paramètre fortement dépendant des évaluateurs. Par ailleurs, le taux d'erreur d'accentuation varie entre 3,5 et 17,8 % (moyenne : 7,56 %; écart-type : 4,8), révélant un certain désaccord entre les évaluateurs.

En parallèle, les mêmes enregistrements ont été annotés automatiquement avec PLSPP v2. Les modules de segmentation par locuteur et de reconnaissance de la parole ont été désactivés, et la transcription orthographique du texte a été directement fournie au module d'alignement. La figure 7.6b compare le taux d'erreur obtenu par PLSPP et ceux obtenus par les annotateurs. Seuls les mots pour lesquels la position de l'accent était précisée ont été inclus. Il apparaît que PLSPP rapporte systématiquement davantage d'erreurs que les annotateurs humains, particulièrement pour le locuteur 6. Deux facteurs peuvent expliquer cet écart : d'une part, des erreurs d'alignement, notamment dans l'enregistrement du locuteur 6 qui présente plus de disfluences que les autres ; d'autre part, le fait que PLSPP identifie systématiquement une syllabe proéminente, alors que les annotateurs humains peuvent indiquer l'absence d'accent (dans ce cas, le mot n'est pas comptabilisé comme incorrect). Cela conduit à un jugement plus strict du système automatique.

Pour réduire l'impact des erreurs d'alignement sur les résultats, un filtrage manuel a été effectué pour ne conserver que les mots correctement alignés. Le nombre de mots retenus par locuteur varie entre 33 et 51 (M = 43), pour un total de 258 mots.

	Sp1	Sp2	Sp3	Sp4	Sp5	Sp6
Mots annotés par PLSPP	55	48	49	34	45	46
Mots dont l'alignement est correct	51	45	48	33	44	37

Tab. 7.5 : *Nombre de mots annotés par PLSPP et nombre de mots conservés pour les analyses pour chaque locuteur*

Le tableau 7.5 présente le nombre de mots annotés par PLSPP et ceux retenus après filtrage manuel pour chaque locuteur.

La figure 7.7 est une projection de ces 258 mots en fonction du score de contraste prosodique (C') estimé par PLSPP, et la moyenne des annotations humaines. Le coefficient de corrélation entre les deux mesures est faible ($r = 0{,}29$). Si l'on divise la figure en quatre zones, on obtient la distribution suivante :

- (a) 52 mots (20%) sont évalués corrects par la majorité des évaluateurs mais obtiennent un contraste prosodique négatif d'après PLSPP ($C' \leq 0{,}5$) ;
- (b) 164 mots (64%) sont évalués corrects par la majorité des évaluateurs et obtiennent un contraste prosodique positif ($C' > 0{,}5$) ;
- (c) 25 mots (10%) sont évalués incorrects par la majorité des évaluateurs et obtiennent un contraste prosodique négatif ;
- (d) 17 mots (7%) sont évalués incorrects par la majorité des évaluateurs mais obtiennent un contraste prosodique positif.

Le taux d'accord entre la moyenne des évaluations humaines et les estimations automatiques est de 73,3 % (coefficient κ de Cohen : 0,27). Il apparaît que la majorité des mots (83,7 %) sont jugés corrects par la majorité des annotateurs, bien qu'un consensus total soit rare (seulement 8 mots, soit 3,1 %). De plus, de nombreux mots avec un contraste accentuel légèrement négatif (C' entre 0,4 et 0,5) sont perçus comme correctement accentués par les annotateurs, tandis que les mots avec des contrastes fortement négatifs, bien que rares, sont généralement jugés incorrects.

Ces observations mettent en lumière deux points importants : premièrement, l'accentuation d'un mot ne devrait pas être considérée comme un paramètre binaire (correct/incorrect), mais plutôt comme un continuum reposant sur le contraste accentuel entre syllabes. Un contraste plus élevé est en effet plus facilement perçu par les auditeurs. Deuxièmement, les auditeurs semblent souvent identifier un schéma accentuel correct même lorsque le contraste prosodique est faible ou légèrement négatif (mauvaise syllabe accentuée). Ces résultats corroborent ceux de van Leyden et van Heuven (1996) et Cooper et al. (2002), qui montrent que les locuteurs anglophones

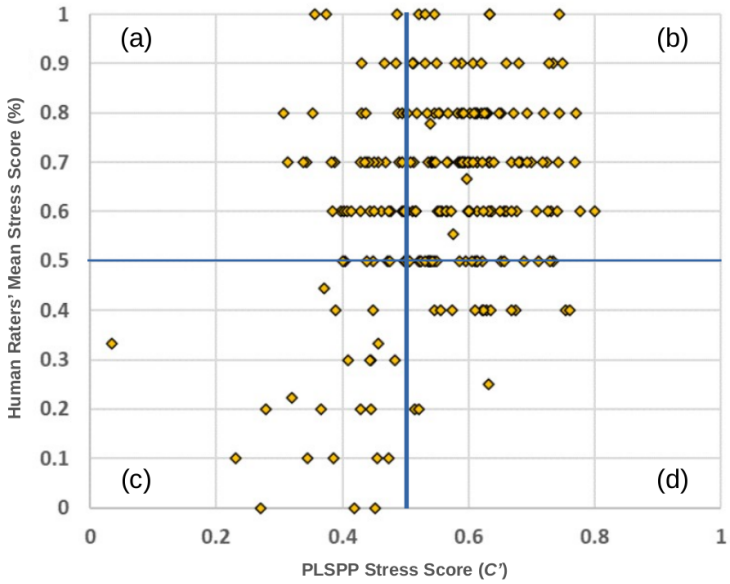


FIG. 7.7 : Score accentuel des 258 mots analysés : score estimé par PLSPP (C') en x et moyenne des évaluateurs humains en y (Kimura et al., 2024, p. 676). Un score de 0,5 en x indique un contraste prosodique nul entre les syllabes ; un score de 0,5 en y indique que 50 % des évaluateurs considèrent que l'accent est correctement positionné.

natifs ont tendance à percevoir un accent sur la syllabe initiale, même en l'absence d'indices lexicaux ou prosodiques. Cette tendance s'explique probablement par le fait que la majorité des mots usuels de l'anglais en parole spontanée sont accentués sur la syllabe initiale, et que ce pattern est donc plus « attendu » que les autres (Cutler & Carter, 1987).

Ainsi, cette étude montre que l'estimation automatique de l'accentuation lexicale par PLSPP est globalement en adéquation avec le jugement des auditeurs natifs, mais également que ce jugement humain varie de manière non négligeable selon les évaluateurs, qui semblent plus ou moins influencés par les tendances d'accentuation de leur langue maternelle. Cette variation inter-évaluateur souligne le fait que la perception de l'accent est un phénomène subjectif et contextuel, et que les paramètres prosodiques syllabiques ne font que participer à cette perception.

3.2 Annotation automatique et conscience phonologique

La deuxième investigation a consisté à confronter les schémas accentuels identifiés par PLSPP avec ceux dont les locuteurs ont conscience. Plus exactement, nous avons cherché à savoir si un locuteur produit effectivement une proéminence acous-

tique sur la syllabe qu'il pense devoir accentuer. Nous avons également investigué l'influence que peuvent avoir les tendances accentuelles de la langue maternelle (L1) des locuteurs, en comparant trois groupes de locuteurs de L1 différentes. Cette section est un résumé d'une présentation donnée à la conférence LabPhon en 2024 (Sugahara et al., 2024).

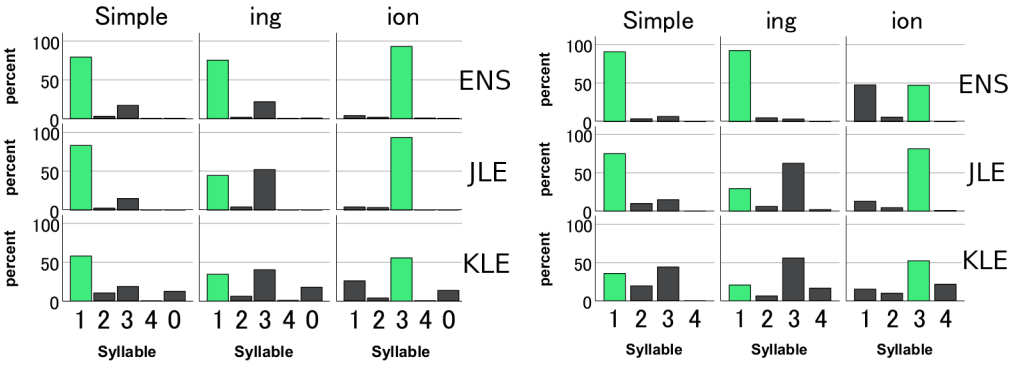
Trois groupes de locuteurs ont donc participé à l'expérience : 12 locuteurs anglophones natifs (ENS), 14 locuteurs japonophones (JLE) et 11 locuteurs coréanophones (KLE) de niveaux CECRL B1 à B2. L'expérience a consisté en deux tâches : une annotation manuelle de la position de l'accent primaire sur une liste de mots cibles, via un questionnaire papier *(stress-assignment task)* – les participants devaient entourer la syllabe qui porte, selon eux, l'accent primaire –, et un enregistrement des mêmes mots cibles dans des phrases porteuses *(production task)*. Ces enregistrements ont ensuite été annotés automatiquement avec PLSPP v2 de la même manière que pour Kimura et al. (2024).

Les mots sélectionnés consistent en 19 triplets composés d'un verbe de 3 syllabes à l'infinitif portant l'accent sur l'initiale (ex. *dominate*), son participe présent (en *-ing*, accent primaire sur l'initiale, ex. *dominating*), et son dérivé substantif en *-ion* (accent primaire sur la 3ème syllabe, ex. *domination*). Pour plus de commodité, nous appellerons par la suite la première syllabe « $\sigma 1$ » et la troisième « $\sigma 3$ ».

À l'issue de la tâche d'annotation manuelle de l'accent, il est apparu que les ENS ont choisi majoritairement la position prescriptive de l'accent pour les trois formes (infinitif, participe présent et substantif). On note toutefois une certaine variabilité inter-annotateur pour l'infinitif et le participe (-ing), pour lesquels $\sigma 3$ a été identifiée comme portant l'accent primaire par un certain nombre de participants (*cf.* figure 7.8a, première ligne). De leur côté, les JLE ont choisi la position prescriptive pour l'infinitif et le substantif, mais se divisent en deux groupes pour le participe, $\sigma 3$ étant sélectionnée dans un peu plus de 50 % des cas (*cf.* figure 7.8a, deuxième ligne). Enfin, les KLE montrent des résultats plus variés, et ne choisissent parfois aucune syllabe (valeur 0, figure 7.8a, troisième ligne).

La figure 7.8b présente les résultats de l'estimation de la position de l'accent par PLSPP. On peut voir que $\sigma 1$ est clairement identifiée comme proéminente chez les ENS pour l'infinitif et le participe. Contrairement à la tâche d'annotation, il n'y a pas de variabilité inter-locuteur (proéminence exclusive sur $\sigma 1$). Du côté des substantifs, la proéminence est détectée tantôt sur $\sigma 1$ et $\sigma 3$, bien que $\sigma 3$ ait été exclusivement sélectionnée dans la tâche d'annotation.

Bien que la proéminence soit détectée sur $\sigma 1$ pour une partie des substantifs chez les locuteurs natifs, le contraste prosodique de $\sigma 3$ permet toutefois de distinguer les participes des substantifs. La figure 7.9a indique la valeur prosodique moyenne



(a) Résultats de l'annotation manuelle de la position de l'accent (stress-assignment task)

(b) Résultats de l'estimation de la position de l'accent par PLSPP (production task)

Fig. 7.8 : *Sur les deux figure, la barre verte représente la position prescriptive de l'accent primaire. Simple, ing et ion font référence à l'infinitif, au participe et au substantif. Le chiffre en x indique la position de la syllabe accentuée. Les figures sont issues de Sugahara et al. (2024).*

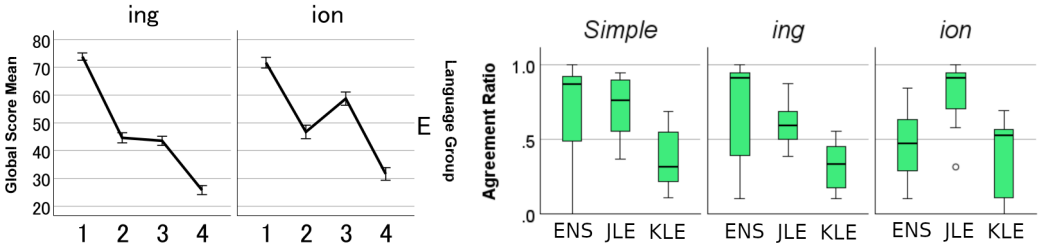
de chaque syllabe pour le participe et le substantif. On peut voir que la première syllabe est plus marquée que les autres dans les deux cas, cependant le contraste avec les autres syllabes est plus fort pour le participe (clairement accentué sur $\sigma 1$), tandis que le substantif présente également une accentuation marquée sur $\sigma 3$, probablement due à la présence de l'accent primaire. Dans le cas du substantif, $\sigma 1$ porte l'accent secondaire, mais elle a peut-être tendance à être accentuée plus fortement sous l'influence d'une tendance à accentuer la première syllabe en anglais. Il peut également s'agir d'un artefact provoqué par la phrase porteuse et le contexte très contraint de la production.

De leur côté, les locuteurs JLE présentent de manière générale un haut degré de corrélation entre l'annotation de la position de l'accent et la détection automatique de la syllabe proéminente. Contrairement aux ENS, un contraste clair est observable sur $\sigma 3$ dans le cas du substantif. Cela pourrait peut-être s'expliquer par le fait que les locuteurs japonais sont habitués à accentuer en médiale dans leur langue maternelle, et à n'accentuer qu'une seule syllabe par mot.

Enfin, dans le cas des KLE, on observe un faible taux d'accord entre les annotations et la détection automatique de proéminence pour les trois formes.

La figure 7.9b indique le taux d'accord par locuteur entre l'annotation et la production. On constate que les ENS et les JNS sont globalement cohérents entre leur annotation et leur production (telle qu'elle est caractérisée par PLSPP). Les ENS sont toutefois moins cohérents dans le cas du substantif, et les JNS dans le cas du participe. Les KLE, quant à eux, apparaissent très peu cohérents sur l'ensemble des mots.

Le haut degré d'accord annotation-production chez les locuteurs ENS et JLE laisse penser que l'estimation de la position de l'accent par PLSPP est généralement



(a) Centile moyen par syllabe pour les participes (-ing) et les substantifs (-ion) dont la proéminence est détectée sur $\sigma 1$ par PLSPP, chez les locuteurs ENS. Les barres d'erreur représentent l'intervalle de confiance à 95 %.

(b) Taux d'accord par locuteur entre l'annotation manuelle de la position de l'accent par le locuteur, et la détection de proéminence par PLSPP. E, J et K correspondent aux groupes de locuteurs anglophones, japonophones et coréanophones.

FIG. 7.9 : Figures issues de Sugahara et al. (2024)

correcte dans ce type de tâche contrôlée. Par ailleurs, l'estimation de la position de l'accent basée sur la moyenne des trois dimensions prosodiques (f_0, intensité, durée) donne de meilleurs résultats que chaque dimension de manière isolée. Il semble donc important de considérer les trois dimensions pour estimer la position de l'accent. Les résultats de l'estimation automatique par dimension et le taux d'accord associé sont présentés en détail par Sugahara et al. (2024).

Nous avons constaté dans cette étude une influence de la L1 des locuteurs, tant au niveau de la tâche d'annotation que de la tâche de production. Si les ENS ont généralement choisi la position prescriptive de l'accent dans la tâche d'annotation, on observe une tendance à accentuer $\sigma 1$ même lorsque l'accent est conscientisé sur $\sigma 3$. Du côté des JLE, on constate une tendance à souvent accentuer $\sigma 3$ plutôt que $\sigma 1$ pour les participes, et à accentuer $\sigma 3$ plus nettement que les ENS pour les substantifs. Cela pourrait être une influence de la plus fréquente accentuation en médiale en japonais. Enfin, les locuteurs KLE présentent effectivement plus de difficultés de manière générale pour placer l'accent, autant pour la tâche d'annotation que de production : le taux d'accord entre annotation et production est généralement en dessous de 50 %. On peut y voir ici l'influence de la présence d'un accent lexical en japonais, contrairement au coréen, qui simplifie la conscientisation et la production de l'accent en anglais.

3.3 Annotation de parole produite par des locuteurs natifs

Dans cette troisième approche pour évaluer les performances de PLSPP en termes d'annotation de l'accentuation lexicale, nous avons analysé l'annotation automatique obtenue à partir de parole produite par des locuteurs natifs. Ici, nous faisons le postulat que les locuteurs natifs ont généralement tendance à accentuer la syllabe prescrite,

	Phrases porteuses		Textes lus	
	Filtré	*Non-filtré*	Filtré	*Non-filtré*
Nombre de mots polysyllabiques lexicaux annotés	541	*954*	4414	*7238*
Score de position de l'accent (S)	79 %	*76 %*	73 %	*71 %*
Contraste prosodique moyen ($\overline{C}$)	28	*27*	17	*16*
Contraste de f_0 ($\overline{C_{f_0}}$)	26	*25*	15	*15*
Contraste d'intensité ($\overline{C_{int}}$)	33	*33*	21	*21*
Contraste de durée ($\overline{C_{dur}}$)	25	*23*	13	*11*
Centile moyen de la syllabe accentuée ($\overline{P_s}$)	70	*70*	61	*60*

Tab. 7.6 : *Résultat des annotations automatiques obtenues sur de la parole native en production de parole contrôlée*

et qu'ils peuvent donc être considérés comme une référence. De cette manière, si la syllabe proéminente identifiée par PLSPP ne correspond pas à la syllabe censée porter l'accent primaire selon le dictionnaire phonologique intégré à l'outil, on peut considérer que l'annotation est incorrecte.

Deux corpus de parole contrôlée ont été analysés avec PLSPP v2 :

- L'enregistrement des 57 mots cibles dans des phrases porteuses lues par 17 locuteurs natifs (dont 12 sont issus de l'étude présentée dans la sous-section précédente) ;
- 92 textes lus en studio par 7 locuteurs natifs, issus de quatre manuels scolaires (Nakanishi et al., 2023a, 2023b, 2024a, 2024b). Les enregistrements vont de 55 s à 3 min 54 s (moyenne : 2 min 6 s), totalisant 3 h 13 min 18 s.

Un total de 954 mots polysyllabiques lexicaux ont été annotés pour le corpus de phrases porteuses, et 7 238 mots pour le corpus de textes lus. Pour réduire l'impact de possibles erreurs d'alignement, nous avons filtré ces mots pour ne conserver que ceux présentant un nombre équivalent de syllabes et de pics d'intensité, ramenant le nombre de mots analysés à respectivement 541 et 4 414. Le tableau 7.6 présente les résultats obtenus avec et sans filtrage.

Dans le cas des phrases porteuses, 79 % des mots analysés sont accentués selon le dictionnaire de référence. Plus précisément, sur les trois types de mots produits, les infinitifs et les participes sont accentués en initiale pour 96 % chacun, tandis que les substantifs sont accentués en $\sigma 3$ seulement pour 46 % (contre 50 % en initiale). On retrouve le constat présenté dans la sous-section précédente, selon lequel les locuteurs natifs produisent une syllabe initiale marquée plus fortement pour la moitié des substantifs analysés.

La figure 7.10 montre le contraste moyen entre les syllabes des trois catégories de mots. Chaque syllabe est représentée par un cercle d'une taille proportionnelle à

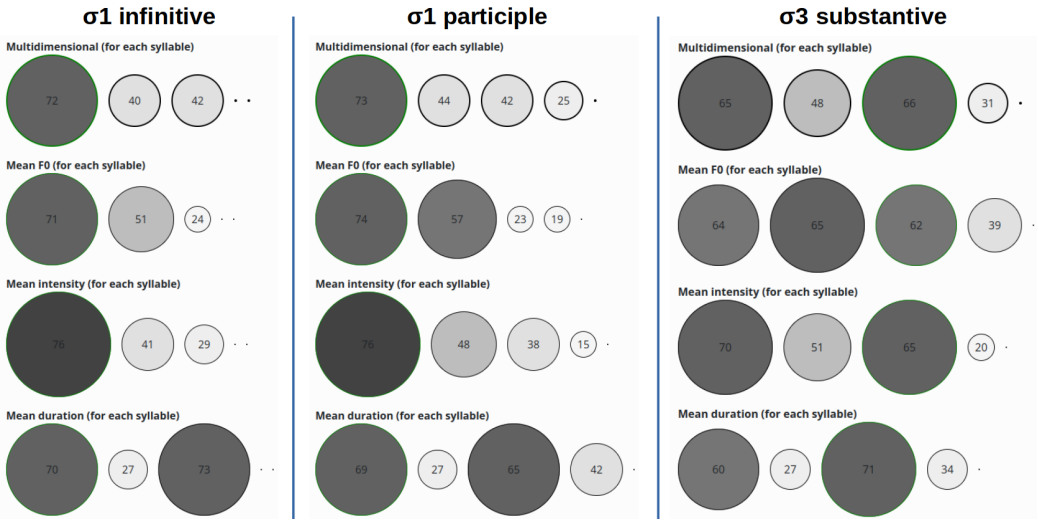


FIG. 7.10 : *Centile moyen par syllabe pour les trois types de mots produits dans les phrases porteuses : infinitifs (n=203), participes (n=155) et substantifs (n=183)*

son poids moyen, c'est-à-dire à la moyenne des centiles de la syllabe en question sur l'ensemble des mots. Aussi, on peut voir que le contraste prosodique entre les syllabes est assez marqué, en particulier au niveau de la f_0 et de l'intensité. Dans le cas de la durée syllabique, $\sigma 3$ est relativement plus longue que les autres syllabes, probablement à cause de la présence d'une diphtongue dans la plupart des mots, dont la durée est intrinsèquement plus longue que les autres voyelles. Le cas du substantif est plus ambigu : $\sigma 1$ et $\sigma 3$ sont fortement marquées sur les trois dimensions prosodiques. Si on peut y voir là une proéminence due à l'accent secondaire, on ne voit pas le cas inverse pour le participe présent, ce qui laisse penser qu'il y a bien une préférence à accentuer la syllabe initiale, au moins dans le cadre de la production de ces phrases porteuses.

Dans le cas de la lecture de textes, le score de position de l'accent est légèrement moins élevé (73 %), de même que le contraste prosodique moyen entre les syllabes ($\overline{C} = 17$, contre 28 dans le cas des phrases porteuses). La figure 7.11 est une visualisation des différents patterns accentuels observés (syllabe proéminente identifiée par PLSPP) pour chaque gabarit accentuel du dictionnaire. Par exemple, les mots dont le gabarit accentuel est Oo (mots de deux syllabes avec accent en initiale), comme "*student*" ou "*wonder*", sont accentués selon PLSPP à 77 % en initiale et 23 % en finale. On peut voir que PLSPP détecte une proéminence sur la syllabe initiale pour une portion non négligeable de mots qui ne sont pas censés être accentués sur cette syllabe : 43 % de mots trisyllabiques censés porter l'accent en finale (ooO), 42 % des mots quadrisyllabiques censés le porter en $\sigma 3$. Dans le premier cas, une analyse approfondie du contraste syllabique par dimension des 87 mots de gabarit théorique ooO

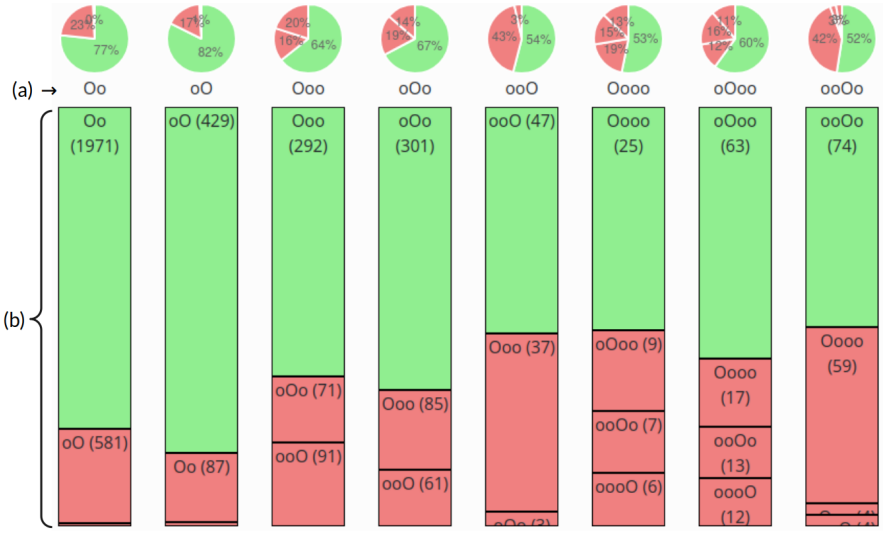


FIG. 7.11 : Position de la proéminence identifiée (b) pour chaque position théorique (a) parmi les 4 414 mots polysyllabiques lexicaux analysés dans les textes lus. "O" représente la syllabe accentuée d'après le dictionnaire (a) ou l'estimation de PLSPP (b). Le nombre entre parenthèses indique le nombre de mots.

révèle que $\sigma 1$ a tendance à être marquée par un allongement de durée et produite avec une intensité plus forte, bien que la proéminence soit détectée sur $\sigma 3$.

Ces résultats montrent que l'annotation de l'accentuation lexicale de PLSPP permet d'obtenir une représentation relativement fiable du contraste prosodique des syllabes, bien qu'elle reste influencée par la durée intrinsèque des voyelles. Il est difficile d'identifier la cause précise des patterns accentuels identifiés comme incorrects par PLSPP : il peut s'agir d'erreurs de mesure, mais aussi d'une réalité acoustique dans la production des locuteurs natifs. Nous retiendrons notamment une tendance de PLSPP à mesurer une proéminence sur la syllabe initiale des mots produits par des locuteurs natifs.

Conclusion

Nous avons proposé une évaluation des différents modules de traitement, afin de vérifier la fiabilité des annotations produites et d'identifier les limites de l'outil. L'évaluation des modules de segmentation en locuteur, de reconnaissance de la parole et d'alignement ont permis de nous assurer que les étapes de prétraitements sont suffisamment performantes pour effectuer une annotation des pauses et de l'accentuation lexicale sur un corpus de conversations spontanées L2. On notera toutefois une précision limitée lorsque les chevauchements entre locuteurs sont nombreux, ainsi

qu'une tendance générale de l'aligneur mot-signal à réduire légèrement la durée des mots. Aucune évaluation de l'analyse syntaxique n'a pu être effectuée faute de corpus de référence, mais elle semble nécessaire au vu des résultats mitigés obtenus lors de l'évaluation de l'annotation des pauses.

Enfin, l'évaluation de l'annotation de l'accentuation lexicale a mis en évidence le fait que l'accent est loin d'être un phénomène absolu et binaire (correct/incorrect), et qu'il est préférable de considérer la mesure des proéminences acoustiques comme un degré de contraste continu entre les syllabes, participant à la perception de l'accent et plus généralement du rythme de la parole. Nous avons constaté que la perception de l'accent varie selon les auditeurs, et qu'elle subit probablement une influence des tendances accentuelles de leur langue maternelle. Cette influence semble également s'observer dans la production des locuteurs natifs, chez qui une importante proportion de mots accentués sur la syllabe initiale a été constatée.

Chapitre 8

Analyses en parole spontanée

Ce chapitre présente les résultats d'analyses des annotations de pauses et d'accentuation produites par PLSPP sur les trois corpus de parole spontanée conversationnelle. Nous présentons d'abord les analyses relatives à la distribution des pauses, puis celles concernant les patterns accentuels. Chaque section débute par les résultats obtenus avec les locuteurs B1 et B2 du corpus CLES-FR, suivis d'une comparaison des groupes de niveau de locuteurs japonophones (CLES-JP) et des locuteurs anglophones natifs (CLES-EN).

Pour le corpus CLES-FR, l'extraction automatique des segments de parole par le premier module de PLSPP a permis d'identifier 1 559 segments, représentant un total de 10 h20 min de parole continue produite par 70 locuteurs B1[1] et 99 locuteurs B2. Les corpus CLES-JP et CLES-EN ont respectivement fourni 275 segments (3 h 37 min, 29 locuteurs) et 113 segments (1 h 55 min, 15 locuteurs[2]). Les analyses présentées dans ce chapitre reposent sur ces ensembles de segments.

8.1 Analyse des patterns de pauses

1.1 Corpus CLES-FR

Sur les 10 h20 min de parole continue extraite du corpus CLES-FR, 72 140 intervalles inter-mots ont été analysés. Parmi ces intervalles, 22 796 (32%) présentent une durée supérieure à 180 ms, 1 085 (1,5%) dépassent 2 s, et 83 (0,1%) excèdent 5 s.

[1] Un des locuteurs n'a pas généré de segments d'au moins 8 s, ce qui explique un effectif final de 70 locuteurs pour le corpus CLES-FR.

[2] Une locutrice bilingue anglais-japonais d'origine américaine a été ajoutée au corpus CLES-EN.

Seuils de durée	p-value	Δ de Cliff	médianes	moyennes	écarts-types
180 ms-2 s	$< 0{,}05$	0,021	$481 - 474$	$600 - 585$	$400 - 390$
180 ms-5 s	$< 0{,}01$	0,024	501	$701 - 675$	$616 - 594$
250 ms-2 s	*ns*	0,009	581	$693 - 683$	$391 - 380$
250 ms-5 s	*ns*	0,014	$602 - 601$	$812 - 791$	$631 - 610$

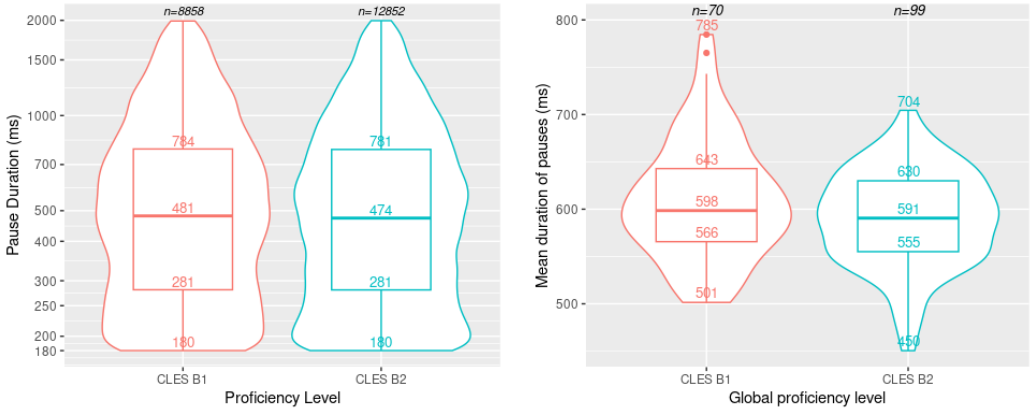
Tab. 8.1 : *Différence de distribution de durée de pauses entre B1 et B2 selon différents seuils de durée. Avec la p-value du test non-paramétrique Wilcoxon-Mann-Whitney, le Δ de Cliff, et la médiane, la moyenne et l'écart type des deux distributions.*

Durées et fréquences des pauses

Pour les analyses, nous avons fixé un seuil minimal de 180 ms. Par ailleurs, un seuil maximal de 2 s a été retenu pour exclure les intervalles potentiellement dus à des erreurs d'alignement, comme détaillé dans le chapitre 5. Ainsi, 21 710 intervalles (30%) ont été identifiés comme des pauses. Ces pauses présentent une durée médiane de 481 ms, avec un premier quartile à 281 ms et un troisième quartile à 782 ms.

La comparaison des durées de pauses entre les locuteurs B1 et B2 montre une différence faible, bien que statistiquement significative ($p < 0{,}05$, médianes respectives de 481 ms et 474 ms, *cf.* figure 8.1a). Cependant, cette différence devient non significative si l'on élève le seuil minimal à 250 ms (médianes à 581 ms). L'inclusion de pauses plus longues (jusqu'à 5 s) n'affecte pas notablement les résultats : la différence reste significative avec un seuil de 180 ms-5 s ($p < 0{,}01$, médianes de 501 ms) mais devient non significative avec un seuil de 250 ms-5 s. Ces résultats suggèrent que les pauses courtes (moins de 250 ms) pourraient contribuer à distinguer les niveaux B1 et B2, tandis que les pauses longues (supérieures à 2 s) ne sont pas discriminantes. Dans l'ensemble, la durée des pauses ne semble pas suffire à différencier les deux groupes, comme en témoigne le Δ de Cliff, toujours proche de 0 (*cf.* tableau 8.1).

En examinant la durée moyenne des pauses par locuteur (*cf.* figure 8.1b), aucune différence significative n'apparaît entre les B1 et B2. Cependant, la forme des distributions révèle des distinctions : la distribution des B1 présente une queue longue, indiquant que certains locuteurs ont des durées moyennes de pauses particulièrement longues (6 locuteurs au-delà de 700 ms), tandis que celle des B2 est plus compacte, et présente au contraire 5 locuteurs dont la durée moyenne des pauses est particulièrement courte.



(a) Distribution des durées de pauses *(b) Durée moyenne des pauses par locuteur*

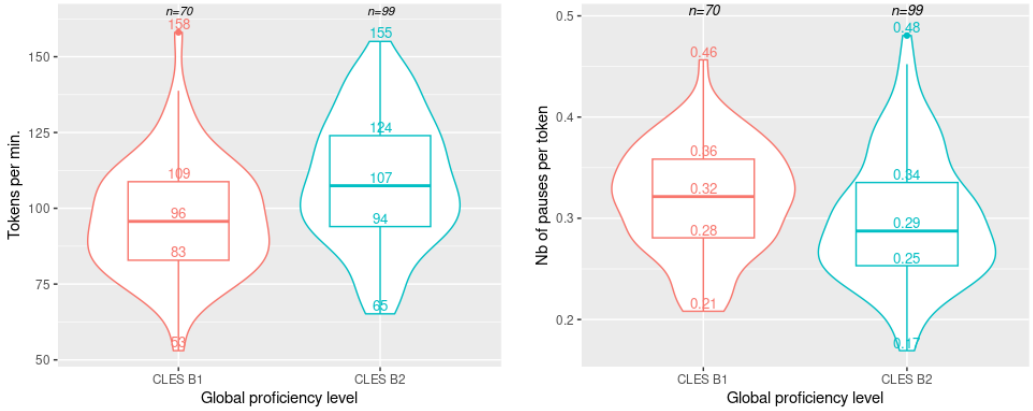
FIG. 8.1 : *Durées des pauses dans le corpus CLES-FR (*180 ms-2 s*)*

Le débit de parole des locuteurs B1 est significativement plus lent que celui des B2 ($p < 0{,}001$, $\Delta = -0{,}35$ (moyen), avec des médianes respectives de 96 et 107 tokens/minute[3], *cf.* figure 8.2a). De plus, nos observations montrent que l'utilisation du nombre de pauses par minute comme indicateur du niveau de compétence en langue n'est pas valide, car nous observons une différence non significative entre les locuteurs B1 et B2, avec des médianes respectives de 32 et 34 pauses par minute. En revanche, le nombre de pauses par token, qui neutralise l'effet de la vitesse d'élocution, révèle quant à lui une différence significative : les B1 font plus de pauses par token que les B2 ($p < 0{,}05$, $\Delta = 0{,}154$ (faible), médianes respectives de 0,32 et 0,29 pauses/token, *cf.* figure 8.2b).

Distribution syntaxique

Bien que les mesures de débit de parole et de fréquence des pauses par token permettent de différencier les locuteurs B1 et B2, ces critères ne suffisent pas à expliquer les variations de compréhensibilité. Le chapitre 3 a souligné l'importance de la distribution syntaxique des pauses et détaillé des études antérieures qui ont montré que la fréquence des pauses situées à l'intérieur des groupes syntaxiques est souvent négativement corrélée avec la perception de fluidité, tandis que les pauses situées entre les groupes semblent avoir un impact moins important (Kahng, 2014 ; Kallio et al., 2022 ; Shea & Leonard, 2019 ; Suzuki & Kormos, 2020).

[3]Les tokens correspondent aux unités de sortie du système de reconnaissance de parole, et équivalent peu ou prou aux mots.



(a) Débit de parole *(b) Nombre de pauses par token*

Fig. 8.2 : *Débit de parole (gauche) et fréquence des pauses (droite) par locuteur dans le corpus CLES-FR (*180 ms-2 s*)*

Pour analyser ces patterns, nous avons distingué deux types de groupes syntaxiques : les propositions et les syntagmes. Une pause est classée comme inter-propositionnelle *(between clauses, BC)* si elle se situe en frontière de proposition, inter-syntagme *(between phrases, BP)* en frontière de syntagme, ou intra-syntagme *(within phrases, WP)* si elle est à l'intérieur d'un syntagme. Par cohérence avec la littérature, nous rapportons aussi les pauses intra-propositionnelles *(within clauses, WC)*, englobant les pauses inter- et intra-syntagmes.

La figure 8.3 montre les proportions de pauses selon le type de frontière syntaxique. À nombre égal de propositions, les B1 effectuent davantage de pauses inter-propositionnelles que les B2 ($p < 0{,}001$, $F_{p,BC}$ médianes : 47 % contre 42 %, $\Delta = 0{,}311$ (faible), $IC = [0{,}132; 0{,}47]$), mais pas significativement plus de pauses intra-propositionnelles (*ns.*, $F_{p,WC}$ médianes : 28 % contre 25 %, $\Delta = 0{,}172$ (faible), $IC = [0; 0{,}334]$). Tous les locuteurs semblent donc privilégier les frontières syntaxiques de haut niveau (entre propositions) pour placer leurs pauses.

Au niveau des syntagmes, les B1 réalisent davantage de pauses inter-syntagmes, mais la différence n'est pas significative (*ns.*, $F_{p,BP}$ médianes : 29 % contre 26 %, $\Delta = 0{,}149$ (faible), $IC = [-0{,}027; 0{,}316]$). En revanche, les pauses intra-syntagmes sont significativement plus fréquentes ($p < 0{,}05$, $F_{p,WP}$ médianes : 21 % contre 18 %, $\Delta = 0{,}187$ (faible), $IC = [0{,}009; 0{,}353]$).

Pour neutraliser l'effet du nombre total de pauses, nous avons calculé la fréquence relative de chaque type ($P_{p,i \in \{BC,WC,BP,WP\}}$, *cf.* figure 8.4). La proportion de pauses inter-propositionnelles est alors légèrement plus faible chez les B1, sans différence significative (*ns.*, $P_{p,BC}$ médianes : 35 % contre 36 %, $\Delta = -0{,}069$). Les proportions

de pauses inter-syntagmes sont également similaires (*ns.*, $P_{p,BP}$ médianes : 52 %, $\Delta = -0{,}069$). La proportion de pauses intra-syntagmes reste significativement plus élevée chez les B1 ($p < 0{,}05$, $P_{p,WP}$ médianes : 12 % contre 11 %, $\Delta = 0{,}216$ (faible), $IC = [0{,}037; 0{,}382]$), mais la différence reste faible.

Ainsi, la principale différence entre les B1 et B2 réside dans la fréquence des pauses en frontières de bas niveau syntaxique. Cependant, cette différence demeure limitée en ampleur.

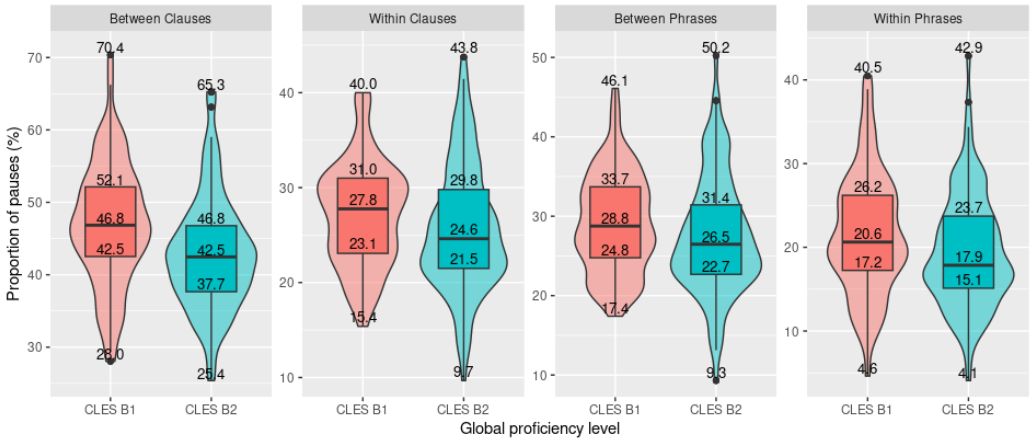


FIG. 8.3 : Proportion de pauses par type de frontière syntaxique ($F_{p,i}$) en fonction du niveau CLES B1 vs. B2

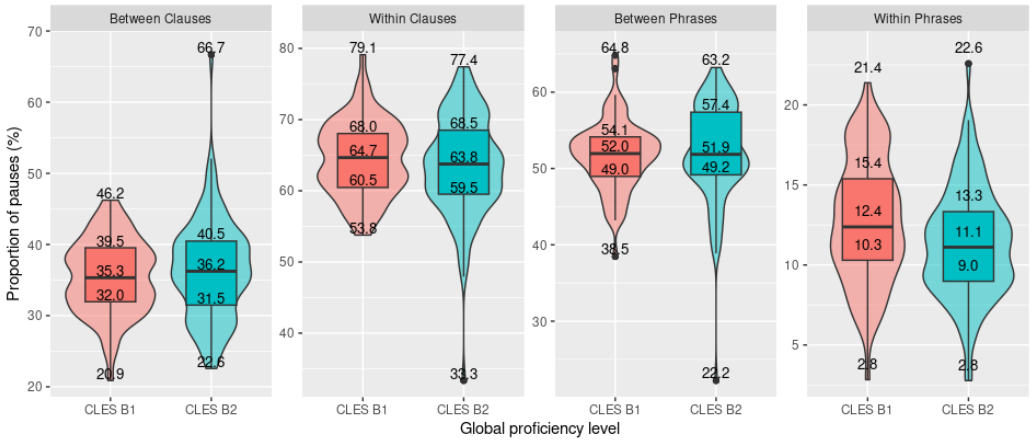
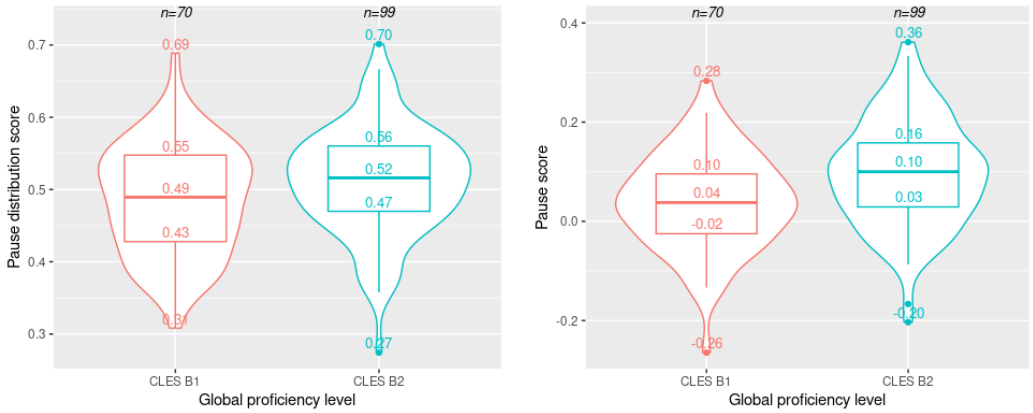


FIG. 8.4 : Proportion de pauses par type de frontière pour 100 pauses ($P_{p,i}$)



(a) Score basé sur le niveau des constituants (proposition ou syntagme, DSP_i)

(b) Score basé sur le nombre de constituants qui s'ouvrent ou se ferment (DSP_n)

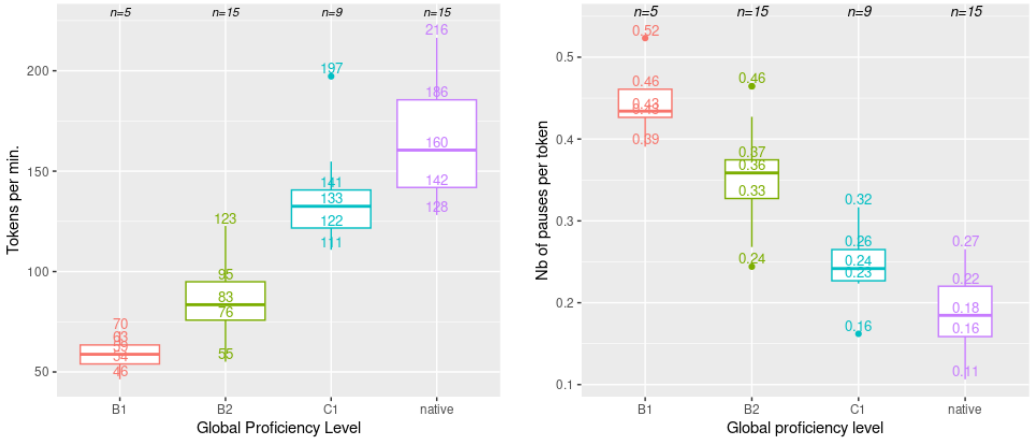
FIG. 8.5 : *Scores de distribution syntaxique des pauses par locuteur (corpus CLES-FR,* 180 ms-2 s*)*

Score de distribution syntaxique des pauses

Le score de distribution syntaxique des pauses (DSP_i) est calculé à partir des pauses inter-propositionnelles, inter-syntagmatiques et intra-syntagmatiques. Ce score est obtenu en normalisant le nombre de pauses de chaque type par le total des pauses par locuteur, puis en effectuant une somme pondérée. Les pauses intra-syntagmatiques sont pénalisées (-1), tandis que celles inter-syntagmatiques et inter-propositionnelles sont respectivement pondérées à +0,5 et +1. Ce calcul reflète la tendance des locuteurs à placer leurs pauses en frontières syntaxiques de haut niveau : plus le score est élevé, plus les pauses respectent ces frontières.

Les résultats montrent que les locuteurs B1 obtiennent en moyenne un score DSP_i plus faible que les locuteurs B2. Cependant, la différence entre les deux groupes reste limitée ($p < 0{,}05$, médianes : 0,49 pour B1 et 0,52 pour B2, $\Delta = -0{,}198$ (faible), $IC = [-0{,}365; -0{,}019]$, *cf.* figure 8.5a).

Une variante de ce score peut également être calculée en fonction du nombre de constituants syntaxiques qui se ferment ou s'ouvrent au moment où survient une pause, plutôt que de s'appuyer uniquement sur le niveau des constituants (propositions ou syntagmes). Ce score DSP_n permet de tenir compte de l'imbrication des groupes syntaxiques et offre davantage de flexibilité dans les paramètres de calcul. Avec cette approche, la différence entre les locuteurs B1 et B2 devient plus marquée ($p < 0{,}001$, médianes : 0,04 pour B1 et 0,10 pour B2, $\Delta = -0{,}301$ (faible), $IC = [-0{,}455; -0{,}13]$, *cf.* figure 8.5b).



(a) Débit de parole *(b) Nombre de pauses par token*

Fig. 8.6 : *Débit de parole (gauche) et fréquence des pauses (droite) par locuteur dans les corpus CLES-JP et CLES-EN (*180 ms-2 s*)*

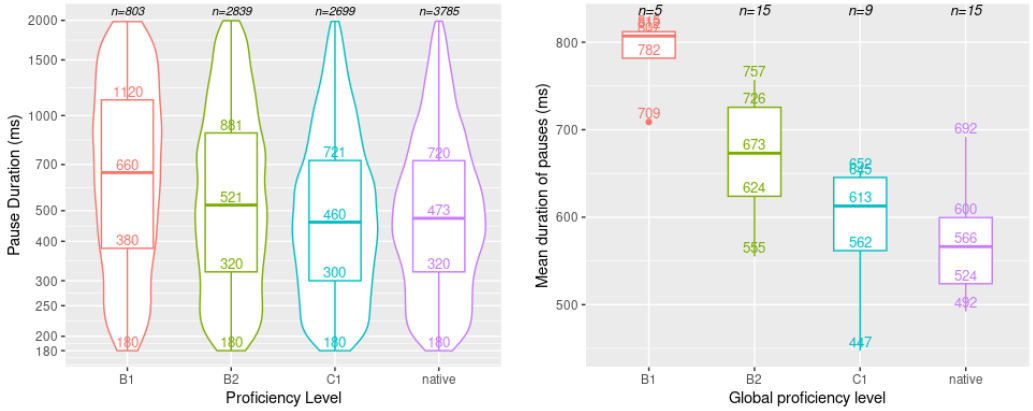
1.2 Corpus CLES-JP et CLES-EN

Pour les corpus CLES-JP et CLES-EN, respectivement 21 631 et 20 486 intervalles ont été analysés. Les différences de durée des pauses entre locuteurs B1 et B2 sont significatives ($p < 0{,}001$) pour tous les seuils étudiés (durée minimale : 180 ou 250 ms, maximale : 2 ou 5 s), bien que les différences soient faibles (le Δ de Cliff varie de 0,145 pour 250 ms-2 s à 0,171 pour 180 ms-5 s). Avec un seuil de 180 ms à 2 s, le corpus contient 6 341 pauses pour les locuteurs japonophones (803 pour les 5 B1, 2 839 pour les 15 B2, 2 699 pour les 9 C1) et 3 785 pauses pour les 15 locuteurs natifs.

Durées et fréquences des pauses

Comme pour le corpus CLES-FR, le débit de parole augmente avec le niveau du locuteur (*cf.* figure 8.6a), entraînant une quantité de parole, et donc de pauses observées, plus importante. En revanche, le nombre de pauses par token suit une tendance inverse, fortement contrastée entre niveaux : les B1 réalisent en moyenne 43 pauses pour 100 mots, les B2 36, les C1 24, et les natifs 18 (*cf.* figure 8.6b).

La figure 8.7 montre la distribution des durées de pauses selon les niveaux. La durée moyenne des pauses par locuteur est fortement contrastée entre les niveaux : **B1 :** médiane à 807 ms, **B2 :** médiane à 673 ms, **C1 :** médiane à 613 ms, et **locuteurs natifs :** médiane à 566 ms.



(a) Distribution des durées de pauses *(b) Durée moyenne des pauses par locuteur*

Fig. 8.7 : *Durées des pauses dans les corpus CLES-JP et CLES-EN (*180 ms-2 s*)*

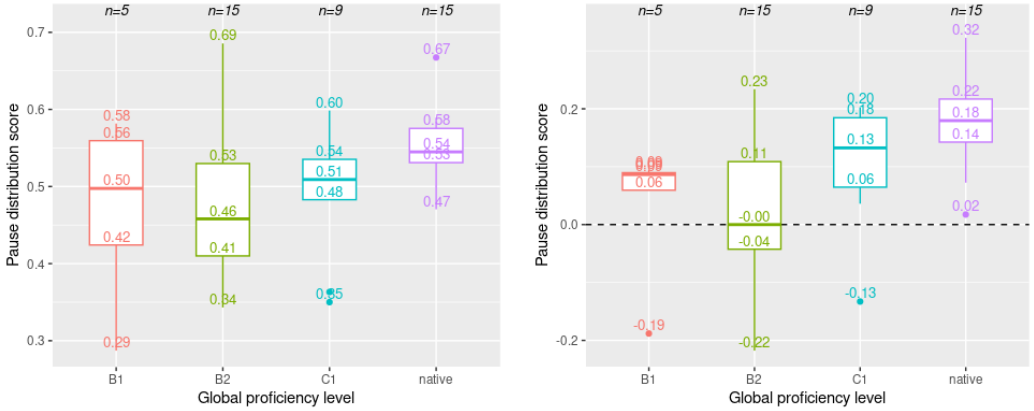
Les différences sont significatives entre **B1 et B2** ($p < 0{,}01$, $\Delta = 0{,}867$ (élevé), $IC = [0{,}338; 0{,}98]$), et **B1/B2 et natifs** ($p < 0{,}001$, $\Delta = 0{,}82$ (élevé), $IC = [0{,}54; 0{,}937]$).

Distribution syntaxique des pauses

La distribution syntaxique des pauses montre des tendances similaires à celles observées pour les locuteurs francophones, mais avec des contrastes plus marqués :

- **Frontières de propositions :** les B1 réalisent significativement plus de pauses que les B2 ($p < 0{,}01$, $F_{p,BC}$ médianes à 58 % pour B1 et 46 % pour B2, $\Delta = 0{,}840$ (élevé), $IC = [0{,}459; 0{,}96]$).
- **Frontières de syntagmes :** différence significative également ($p < 0{,}05$, $F_{p,BP}$ médianes à 43 % pour les locuteurs B1 et 33 % pour B2, $\Delta = 0{,}657$ (élevé), $IC = [0{,}115; 0{,}876]$).
- **Intra-syntagmes :** taille d'effet importante mais pas de différence significative, probablement en raison du faible nombre de données pour les B1 (*ns*, $F_{p,WP}$ médianes à 30 % pour B1 et 23 % pour B2, $\Delta = 0{,}52$ (élevé), $IC = [-0{,}037; 0{,}83]$).

Pour les proportions de pauses par type, aucune différence significative n'est observée entre les groupes de niveau des locuteurs du corpus CLES-JP. Cependant, la comparaison des locuteurs B1+B2 avec les locuteurs natifs du corpus CLES-EN révèle :



(a) Score basé sur le niveau des constituants DSP_i

(b) Score basé sur la profondeur des frontières DSP_n

FIG. 8.8 : *Score de distribution syntaxique des pauses par locuteur, basé sur le niveau des constituants ou le niveau de profondeur des frontières syntaxiques. Corpus CLES-JP et CLES-EN,* 180 ms-2 s

- Davantage de pauses en frontières de propositions chez les natifs ($p < 0{,}05$, $P_{p,BC}$ médianes à 35 % pour L2 et 37 % pour L1, $\Delta = -0{,}4$ (moyen), $IC = [-0{,}687; -0{,}004]$),
- Moins de pauses à l'intérieur des syntagmes ($p < 0{,}05$, $P_{p,WP}$ médianes à 13 % pour L2 et 10 % pour L1, $\Delta = 0{,}483$ (élevé), $IC = [0{,}087; 0{,}747]$).

Score de distribution syntaxique des pauses

Le score de distribution syntaxique des pauses (DSP) révèle une différence non significative entre B1 et B2, quelle que soit la méthode de calcul (niveaux de constituants ou profondeur des frontières). Cependant, la différence entre locuteurs natifs (L1) et non natifs (L2) est significative :

- **Niveaux de constituants (DSP_i) :** $p < 0{,}01$, médianes à 0,48 pour L2 et 0,54 pour L1, $\Delta = -0{,}527$ (élevé), $IC = [-0{,}777; -0{,}131]$.
- **Profondeur des frontières syntaxiques (DSP_n) :** $p < 0{,}001$, médianes à 0,06 pour L2 et 0,18 pour L1, $\Delta = -0{,}707$ (élevé), $IC = [-0{,}891; -0{,}32]$ (*cf.* figure 8.8).

Ces résultats montrent que les pauses sont plus souvent situées en frontières syntaxiques de haut niveau chez les locuteurs natifs du corpus CLES-EN que chez les locuteurs non natifs japonophones du corpus CLES-JP.

8.2 Accentuation lexicale

Dans le chapitre 3, nous avons vu que la précision de la position de l'accent lexical est souvent corrélée avec le jugement de compréhensibilité des locuteurs. Toutefois, la majorité des études mentionnées s'appuient sur des énoncés en parole lue ou sur des annotations manuelles de la position de l'accent. Nous avons cherché ici à déterminer si des mesures automatiques peuvent caractériser les schémas accentuels des locuteurs en parole spontanée, et si une différence significative entre les niveaux B1 et B2 est observable. Par ailleurs, nous avons examiné la qualité de l'accentuation en termes de contraste prosodique entre les syllabes au niveau de la f_0, de l'intensité et de la durée.

2.1 Corpus CLES-FR

Données analysées

Les segments de parole du corpus CLES-FR analysés par PLSPP comprennent un total de 68 515 tokens. Le nombre de tokens par locuteur est légèrement inférieur pour les B1 par rapport aux B2 (médianes à 376 contre 422), mais cette différence n'est pas significative. Les mesures d'accentuation syllabique réalisées ici portent exclusivement sur les mots polysyllabiques lexicaux (noms communs, verbes, adjectifs et adverbes). On compte un total de 14 873 mots polysyllabiques lexicaux, significativement plus nombreux chez les locuteurs B2 ($p < 0{,}05$, médianes à 75 pour B1 et 94 pour B2, $\Delta = -0{,}232$ (faible), $IC = [-0.396; -0.052]$; rapporté au nombre de tokens par locuteur : $p < 0{,}01$, médianes à 21 % et 23 %, $\Delta = -0{,}238$ (faible), $IC = [-0.397; -0.065]$). Cependant, parmi ces mots, seulement 6 468 ont été annotés par PLSPP, soit 43 % des mots ciblés initialement. Cette proportion limitée vient du fait que seuls les mots pour lesquels le nombre de pics d'intensité détectés correspond au nombre de syllabes attendues ont été conservés. Nous reviendrons sur ce constat dans le chapitre 10. Toutes les analyses de cette section portent sur ces 6 468 mots, désignés par la suite comme « mots annotés ».

Nombre de mots annotés par locuteur Le nombre absolu de mots annotés par locuteur est significativement plus élevé pour les B2 ($p < 0{,}01$, médianes à 32 pour B1 contre 41 pour B2). Cependant, cette différence n'est pas significative lorsqu'on rapporte le nombre de mots annotés au nombre total de tokens (médianes à 9 et 10%) ou au nombre de mots polysyllabiques lexicaux (médianes à 42 et 43%). Cela indique que les locuteurs B2 n'utilisent pas proportionnellement plus de mots polysyllabiques que les B1, et que les mots produits ne sont pas mieux reconnus par PLSPP pour les B2 que pour les B1.

	B1		*B2*		*all*	
Position	**Théorique**	**Observée**	**Théorique**	**Observée**	**Théorique**	**Observée**
Initiale	74 % (1636)	23 % (502)	74 % (2796)	28 % (1059)	74 % (4432)	26 % (1561)
Médiale	14 % (301)	6 % (131)	13 % (490)	7 % (264)	13 % (791)	7 % (395)
Finale	12 % (274)	71 % (1578)	13 % (505)	65 % (2468)	13 % (779)	67 % (4046)

Tab. 8.2 : *Position théorique et observée de l'accent lexical dans les mots de 2 à 3 syllabes annotés par PLSPP (corpus CLES-FR, $n = 6\ 002$)*

Caractéristiques des mots annotés Les mots annotés sont principalement des noms communs (57% des occurrences, doublons inclus), suivis des verbes (19%), des adjectifs (12%) et des adverbes (12%). La majorité de ces mots est composée de deux syllabes (73%), mais on trouve également des mots de trois syllabes (21%), quatre syllabes (5%) et cinq ou six syllabes (moins de 1%). Pour limiter l'influence potentielle de l'accent secondaire et étant donné que les mots de plus de trois syllabes représentent moins de 6% des mots annotés, nous concentrons nos analyses sur les mots de deux à trois syllabes ($n = 6\ 002$).

Parmi ces mots, l'accent primaire est attendu en initiale dans 74 % des cas (4 432 mots), en médiale dans 13 % (791 mots) et en finale dans 13 % (779 mots) (*cf.* tableau8.2). Pour les mots à deux syllabes, 84 % d'entre eux sont accentués en initiale contre 16 % en finale. Pour les mots à trois syllabes, la majorité est accentuée sur la syllabe médiale (58%), suivie de l'initiale (38%) et de la finale (4%).

Patterns accentuels observés

Examinons à présent les patterns accentuels produits par les locuteurs. Nous utiliserons désormais le terme « syllabe proéminente » pour désigner la syllabe identifiée par PLSPP comme étant acoustiquement proéminente au sein du mot annoté, et donc perçue, en théorie, comme syllabe accentuée par l'auditeur. Le terme « accent théorique » fera quant à lui référence à la position de l'accent prescrit par le dictionnaire de référence, ici le *CMU Pronouncing Dictionary*[4] (version 0.7b).

Les résultats montrent que 67 % des syllabes proéminentes se situent en finale, contre 26 % en initiale et seulement 7 % en médiale (*cf.* tableau 8.2). De manière générale, la proportion de mots dont la syllabe proéminente correspond à la position de l'accent primaire théorique est relativement faible : 36 % sur l'ensemble du corpus, avec 32 % pour les mots produits par les locuteurs B1 et 38 % pour ceux produits par les B2.

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

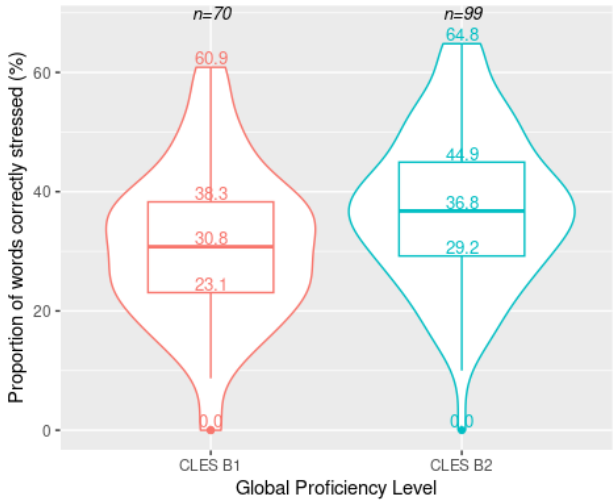


FIG. 8.9 : Score de position de l'accent par locuteur, sur 6 468 mots lexicaux de 2 à 3 syllabes (corpus CLES-FR)

Cependant, étant donné que le nombre de mots annotés varie entre les locuteurs, il est plus pertinent d'examiner la proportion de mots correctement accentués pour chaque locuteur, que nous appellerons « score de position de l'accent » (S). Ces scores individuels présentent une grande variabilité, allant de 0 à 64,8 % (médiane à 34,6 %).

La figure 8.9 montre la distribution des scores par locuteur selon leur niveau. Bien que les scores des locuteurs B1 et B2 se chevauchent largement, une différence significative est observée entre les deux groupes ($p < 0{,}01$, médianes respectives à 30,8 % pour les B1 et 36,8 % pour les B2). Toutefois, cette différence est accompagnée d'une taille d'effet limitée ($\Delta = -0{,}275$ (faible), $CI = [-0{,}432; -0{,}102]$).

La différence entre les deux groupes de locuteurs réside principalement dans le taux d'accentuation correcte des mots à accent en initiale et en médiale, tandis que les deux niveaux ont une proportion de mots correctement accentués de l'ordre de 77 % pour les mots à accent en finale. En effet, lorsqu'on examine le score obtenu en fonction de la position de l'accent théorique, les résultats montrent que les locuteurs B2 ont une proportion de mots correctement accentués plus élevée pour les mots à accent initial ($p < 0{,}01$, médianes à 25 % pour les B1 et 32 % pour les B2, $\Delta = -0{,}269$ (faible), $CI = [-0{,}428; -0{,}094]$) et médial (*ns.*, médianes à 30 % pour les B1 et 36 % pour les B2, $\Delta = -0{,}158$ (faible), $CI = [-0{,}329; -0{,}024]$). En revanche, aucune différence significative n'est observée pour les mots à accent final (*ns.*, médianes à 77 % ; $\Delta = -0{,}032$ (négligeable), $CI = [-0{,}148; 0{,}210]$). De manière générale, l'accentuation en initiale et en médiale est bien moins réalisée, avec moins de 40 % de

Rang	Mot	Gabarit accentuel théorique	Position accent théorique	Catégorie grammaticale	Fréquence	Détection gabarit attendu (%)
1	students	Oo	initial	NOUN	187	24.60
2	maybe	Oo	initial	ADV	186	41.40
3	people	Oo	initial	NOUN	173	20.81
4	computer	oOo	medial	NOUN	155	30.32
5	testing	Oo	initial	VERB	111	26.13
6	also	Oo	initial	ADV	100	28.00
7	really	Oo	initial	ADV	96	33.33
8	computers	oOo	medial	NOUN	95	26.32
9	problem	Oo	initial	NOUN	92	39.13
10	teacher	Oo	initial	NOUN	89	19.10
11	children	Oo	initial	NOUN	87	27.59
12	teachers	Oo	initial	NOUN	75	10.67
13	cameras	Ooo/Oo	initial	NOUN	73	23.29
14	student	Oo	initial	NOUN	59	25.42
15	money	Oo	initial	NOUN	57	21.05
16	very	Oo	initial	ADV	56	44.64
17	paper	Oo	initial	NOUN	53	30.19
18	agree	oO	final	VERB	52	50.00

Tab. 8.3 : *Liste des mots annotés de plus de 50 occurrences dans le corpus CLES-FR ("O" représente la syllabe censée porter l'accent primaire)*

réalisation correcte. Or, l'accent en initial est de loin le plus fréquent, avec 74 % des mots annotés.

Le tableau 8.3 présente les 18 mots les plus fréquents parmi les mots annotés (plus de 50 occurrences). On observe que les mots les plus fréquents ne sont pas nécessairement ceux dont l'accentuation est la mieux maîtrisée. Par exemple, "*students*" n'est accentué sur la syllabe initiale que dans 24,6 % des cas (sur 187 occurrences), "*people*" dans 20,8 % des cas (173 occurrences), "*teacher*" dans 19 % des cas (89 occurrences), et "*teachers*" dans seulement 10,7 % des cas (75 occurrences).

Contraste prosodique

Dans cette section, nous examinons le degré de contraste prosodique entre les syllabes d'un mot. Il ne s'agit pas uniquement d'évaluer si la syllabe proéminente correspond à celle qui porte l'accent lexical d'après le dictionnaire de référence, mais de mesurer à quel point cette syllabe se démarque des autres sur le plan prosodique. Pour cela, nous calculons pour chaque mot un contraste prosodique C, obtenu par la différence entre la valeur acoustique normalisée P_s de la syllabe censée être accentuée et la moyenne $\overline{P_u}$ des autres syllabes du mot. La valeur obtenue est comprise entre -100 et 100, où une valeur positive indique que la syllabe proéminente correspond à la position de l'accent primaire. Plus cette valeur est élevée, plus le contraste prosodique avec les autres syllabes est marqué. Nous nommerons $\overline{C}$ la moyenne des contrastes calculés sur l'ensemble des mots annotés pour un locuteur.

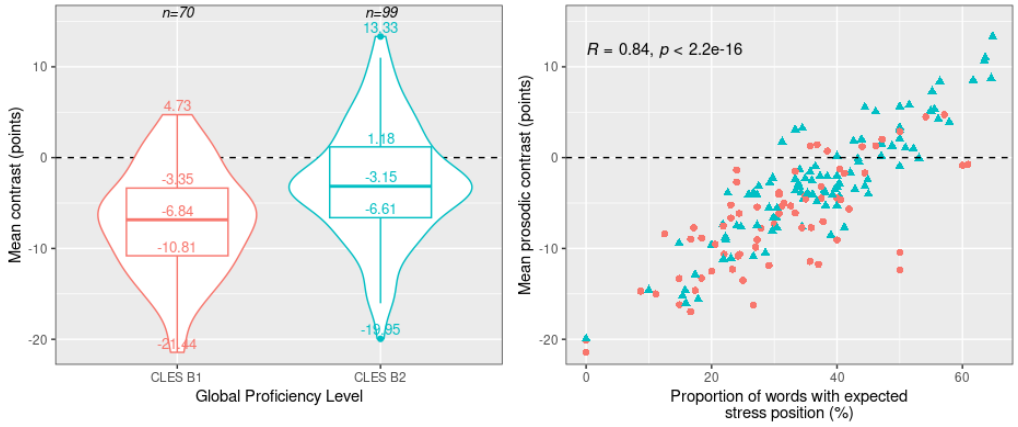


FIG. 8.10 : Contraste prosodique moyen $\overline{C}$ dans le corpus CLES-FR, à gauche en fonction du niveau du locuteur ; à droite en fonction du score de position de l'accent S

Le contraste moyen par locuteur varie entre -21,44 et 13,33 dans le corpus CLES-FR, avec une médiane à -4,11. Cela indique une tendance générale des locuteurs à ne pas accentuer la syllabe censée porter l'accent primaire. Bien que les scores des locuteurs B1 et B2 se chevauchent largement, la différence entre les deux groupes est significative et plus prononcée qu'avec le score de position S décrit dans la section précédente ($p < 0{,}001$, $\overline{C}$ médianes à -6,84 pour les locuteurs B1 et -3,15 pour les B2, $\Delta = -0{,}389$ (moyen) $CI = [-0{,}534; -0{,}222]$, *cf.* figure 8.10 gauche). De plus, il existe une corrélation forte entre le score de position S et le contraste moyen $\overline{C}$ ($R = 0{,}84$, $p < 0{,}001$, *cf.* figure 8.10 droite), montrant que les locuteurs ayant une meilleure maîtrise de la position accentuelle produisent également un contraste prosodique plus marqué[5].

Différences entre les dimensions prosodiques Pour identifier la dimension prosodique différenciant le mieux les locuteurs B1 et B2, nous avons examiné le contraste moyen séparément pour chaque dimension (f_0, intensité, durée). Les résultats, figure 8.11, montrent que la plus grande différence se situe au niveau de l'intensité ($p < 0{,}001$, $\overline{C_{int}}$ médianes à -5,82 pour les B1 et 0,77 pour les B2, $\Delta = -0{,}450$ (moyen), $CI = [-0{,}590; -0{,}283]$). Vient ensuite la fréquence fondamentale, avec une différence moins marquée mais significative ($p < 0{,}001$, $\overline{C_{f_0}}$ médianes à -7,38 et -1,44, $\Delta = -0{,}317$ (faible), $CI = [-0{,}475; -0{,}138]$). En revanche, la durée des syllabes ne présente pas de différence significative entre les deux niveaux (*ns.*, médianes à -9,84 et -9,36, $\Delta = -0{,}018$ (négligeable) $CI = [-0{,}196; 0{,}162]$). Une analyse par position de l'accent (initale, médiale, finale) indique comme attendu un fort contraste de durée en finale pour les deux groupes de locuteurs, et une difficulté générale à placer

[5]Constat toutefois sans surprise, les deux variables étant dépendantes.

l'accent en initiale ou médiale. La différence entre B1 et B2 se joue réellement au niveau de l'utilisation de la f_0 et de l'intensité pour l'accentuation en initiale ($p < 0{,}001$ dans les deux cas, *cf.* figure 8.12).

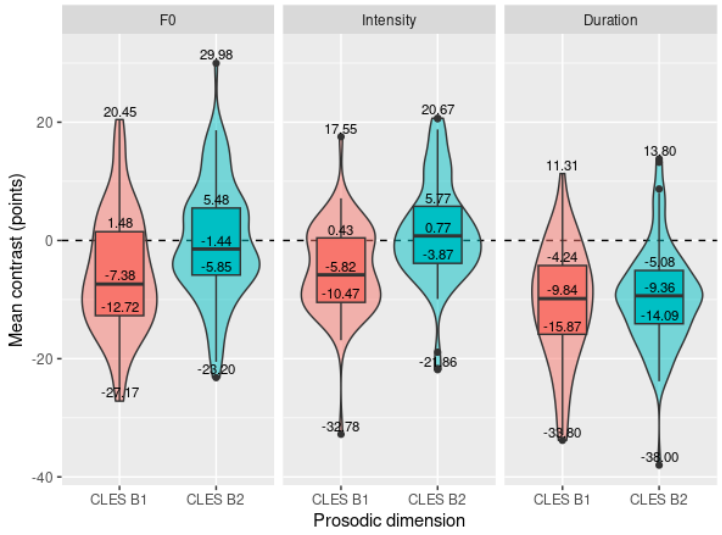


FIG. 8.11 : Contraste moyen par dimension prosodique par locuteur (corpus CLES-FR)

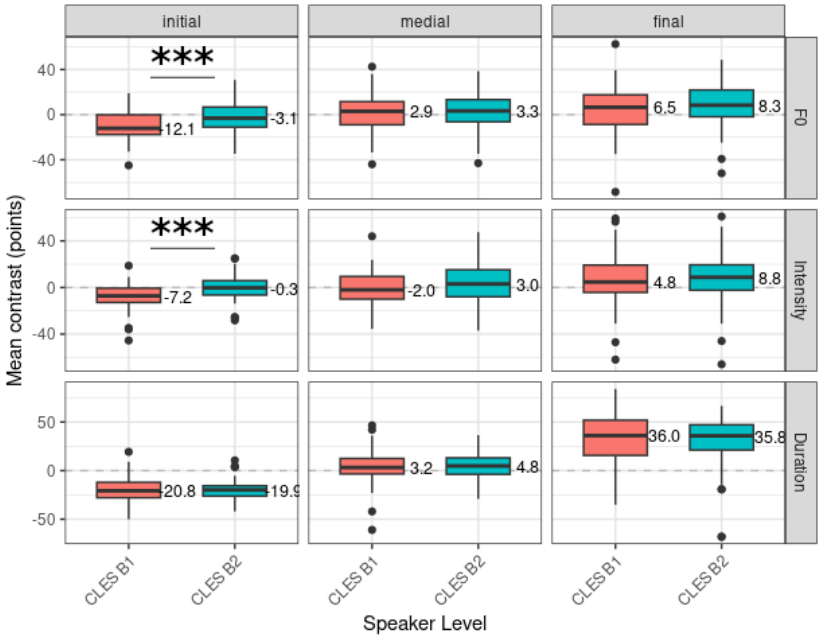


FIG. 8.12 : Contraste moyen par dimension prosodique et par position de l'accent (corpus CLES-FR)

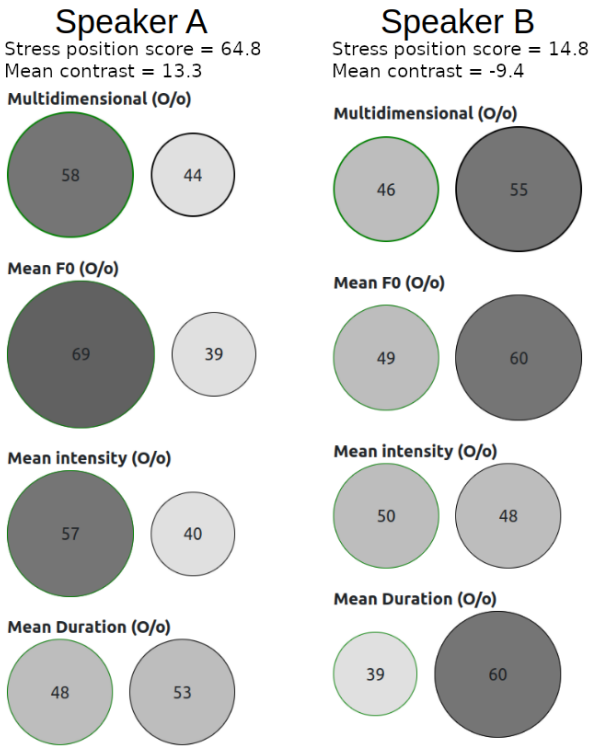


FIG. 8.13 : Contrastes prosodiques de deux locuteurs au profil différent
Écouter un extrait : Speaker A et Speaker B.

Profils d'accentuation des locuteurs La figure 8.13 illustre les profils d'accentuation de deux locuteurs représentatifs, SpeakerA et SpeakerB. La figure indique le contraste prosodique moyen du locuteur pour chaque dimension ($\overline{C}$, $\overline{C_{f_0}}$, $\overline{C_{int}}$ et $\overline{C_{dur}}$). Les cercles représentent les valeurs prosodiques normalisées (P_s pour la syllabe censée être accentuée et $\overline{P_u}$ pour la moyenne des autres syllabes). La taille des cercles et les valeurs inscrites reflètent les centiles des valeurs prosodiques.

- **SpeakerA,** avec un score de position de $S = 64{,}8$ % et un contraste moyen de $\overline{C} = 13{,}3$, est représentatif d'un locuteur de haut niveau en termes d'accentuation. On peut voir que la syllabe accentuée se démarque clairement en termes de f_0 (contraste de 30) et d'intensité (17), mais pas en durée (-5).
- **SpeakerB,** en revanche, présente un score de position de seulement $S = 14{,}8$ % et un contraste moyen de $\overline{C} = -9{,}4$. La syllabe censée être accentuée a tendance à être moins marquée que les autres, particulièrement au niveau de la durée (-21) et de la f_0 (-11), tandis que l'intensité reste peu mobilisée (+2).

Ces profils montrent des tendances générales : les locuteurs avec un score de position et un contraste élevés mobilisent davantage la f_0 et l'intensité, tandis que ceux avec des scores faibles sont principalement influencés par un contraste négatif de durée et de f_0, sans exploitation notable de l'intensité.

2.2 Corpus CLES-JP et CLES-EN

Les corpus CLES-JP et CLES-EN comptent respectivement 275 et 113 segments de parole (21 356 et 20 358 tokens). La proportion de mots polysyllabiques lexicaux par token est similaire dans les deux corpus (médianes : 23 % pour les locuteurs B1/B2 du CLES-JP et 22 % pour les locuteurs natifs). Cependant, la proportion de mots annotés par PLSPP est significativement plus élevée chez les locuteurs japonophones que chez les natifs ($p < 0{,}001$, médianes : 41 % pour les locuteurs B1/B2 et 30 % pour les natifs, $\Delta = 0{,}753$ (élevé) $IC = [0{,}410; 0{,}910]$). Ce résultat inattendu suggère une meilleure reconnaissance des mots pour les locuteurs non natifs. Au total, 1 913 mots ont été annotés dans le CLES-JP et 1 354 dans le CLES-EN. Les distributions des catégories grammaticales et du nombre de syllabes restent similaires entre les deux corpus, comme l'illustre le tableau 8.4. À l'instar du corpus CLES-FR, l'analyse se focalise sur les mots de 2 à 3 syllabes, représentant environ 96 % des mots annotés.

La figure 8.14 montre les scores de position de l'accent pour chaque locuteur. Deux observations majeures émergent. Premièrement, aucune amélioration significative n'est observée entre les niveaux des locuteurs japonophones : les locuteurs B1 obtiennent des scores variant entre 35,3 % et 70,6 % (médiane à 45,8 %), tandis que les locuteurs B2 atteignent des scores compris entre 31,7 % et 72,7 % (médiane à 49,3 %, différence non significative). Les 9 locuteurs C1 n'affichent pas de résultats sensiblement supérieurs. Deuxièmement, les scores des locuteurs natifs présentent une grande variabilité (19,2 % à 69,6 %) et, en moyenne, restent inférieurs à ceux des locuteurs japonophones (médianes à 49,3 % pour le CLES-JP et 41,2 % pour le CLES-EN, différence non significative, $\Delta = 0{,}170$ (faible), $IC = [-0{,}223; 0{,}516]$). Ces résultats suggèrent que la détection de la syllabe proéminente est influencée par un facteur indépendant du niveau de compétence en langue.

	CLES-JP	*CLES-EN*
Noms	53 % (945)	43 % (567)
Verbes	24 % (436)	26 % (342)
Adjectifs	15 % (264)	16 % (211)
Adverbes	9 % (154)	15 % (196)

	CLES-JP	*CLES-EN*
2 syll.	82 % (1483)	83 % (1094)
3 syll.	14 % (247)	14 % (190)
4 syll.	3 % (60)	2 % (30)
5+ syll.	0,5 % (9)	0,2 % (8)

Tab. 8.4 : *Catégories grammaticales et nombres de syllabes des mots annotés par PLSPP dans les corpus CLES-JP (n=1913) et CLES-EN (n=1354)*

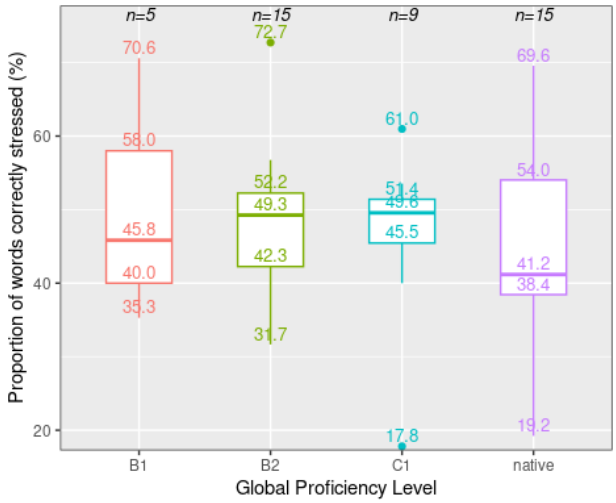


FIG. 8.14 : Scores de position par locuteur, sur 3 014 mots lexicaux de 2 à 3 syllabes (corpus CLES-JP et CLES-EN)

	CLES-JP		*CLES-EN*	
Position	**Théorique**	**Observée**	**Théorique**	**Observée**
Initiale	83 % (1430)	40 % (689)	83 % (1071)	38 % (485)
Médiale	7 % (120)	4 % (76)	6 % (79)	6 % (71)
Finale	10 % (180)	56 % (965)	10 % (134)	57 % (728)

TAB. 8.5 : Position théorique et observée de l'accent lexical dans les mots des corpus CLES-JP (n=1 913) et CLES-EN (n=1 354)

Le tableau 8.5 compare le nombre de mots par position d'accent théorique (attendue) et observée (proéminence détectée par PLSPP). Alors que l'accent est attendu en initiale dans 80 % des cas, en médiale dans 7 %, et en finale dans 10 %, PLSPP détecte une proéminence en initiale dans seulement 40 % des mots, en médiale dans 5 %, et en finale dans près de 60 %. Cette tendance est similaire pour les locuteurs japonophones et natifs, indiquant que la syllabe finale est souvent considérée comme proéminente, même lorsqu'elle ne devrait pas l'être a priori, et ce indépendamment de la langue maternelle du locuteur.

L'évaluation du contraste prosodique moyen $\overline{C}$ indique que la syllabe censée être accentuée est systématiquement plus élevée en intensité, mais plus courte en durée que la moyenne des autres syllabes (*cf.* figure 8.15). Le contraste de f_0 tend à être positif, mais avec une amplitude moindre. Ces résultats reflètent des tendances similaires à celles observées dans le corpus CLES-FR (*cf.* figure 8.10), bien que les valeurs soient ici plus extrêmes : les contrastes d'intensité et de hauteur sont généralement marqués et alignés avec la position d'accent théorique, tandis que le contraste de durée

est fortement négatif. Ce contraste négatif suggère que la dernière syllabe des mots est souvent allongée, quel que soit l'accentuation, et cela influence la détection de la proéminence, au point de rendre le contraste global négatif.

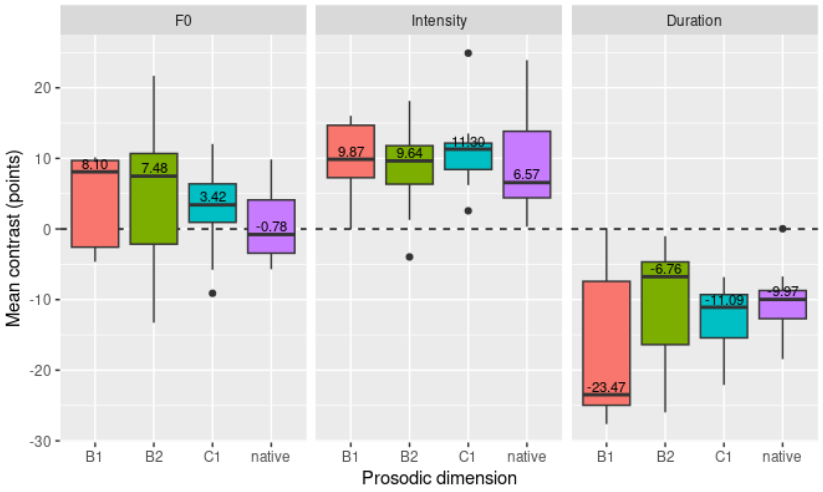


FIG. 8.15 : Contraste moyen par dimension et par locuteur (corpus CLES-JP, CLES-EN)

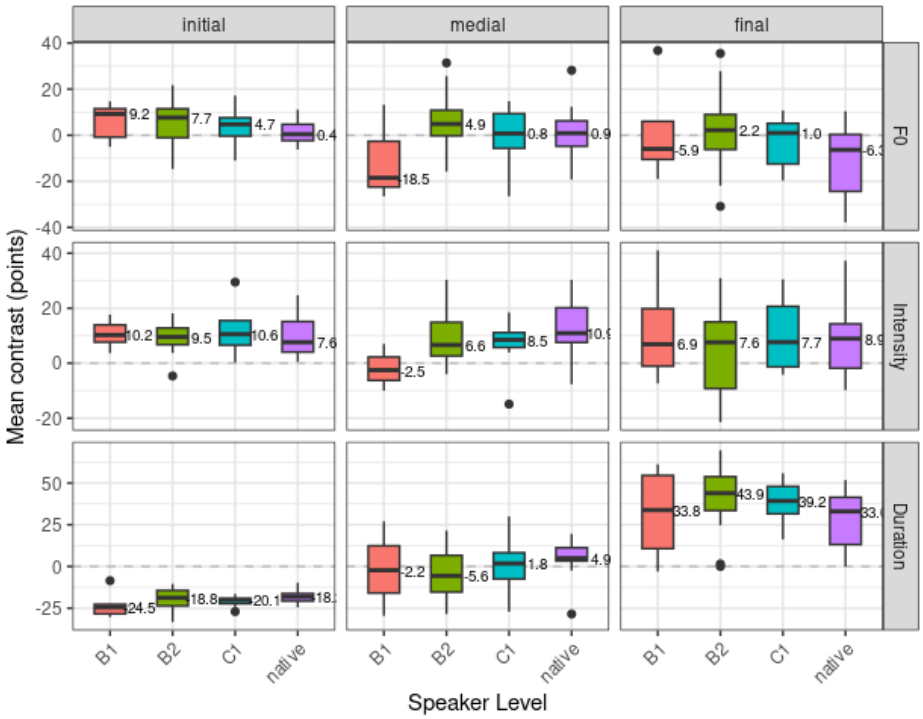


FIG. 8.16 : Contraste moyen par dimension prosodique et par position de l'accent

Enfin, l'analyse du contraste prosodique en fonction de la position de l'accent (*cf.* figure 8.15) montre que, à la différence des locuteurs francophones, l'accentuation en initiale et en médiale ne constitue pas de difficulté particulière ($\overline{C}$ positifs dans l'ensemble, pour la f_0 et l'intensité). Par ailleurs, si le contraste d'intensité moyen pour les locuteurs natifs est généralement élevé, voire en tête pour l'accentuation médiale et finale, le contraste de f_0 est quant à lui proche de 0, voire négatif, quelle que soit la position de l'accent. Ce constat laisse penser qu'un biais externe vient perturber les analyses. Nous en reparlerons en discussion, chapitre 10.

Conclusion

Le tableau 8.6 récapitule l'ensemble des résultats que nous avons mentionnés dans ce chapitre. Pour chaque corpus, il indique la valeur médiane obtenue pour chaque groupe de niveau, ainsi que la différence entre les groupes B1 et B2, et locuteurs japonophones et natifs.

Nous avons montré dans ce chapitre que, si les locuteurs de niveau B1 produisent plus de pauses que ceux de niveau B2 à tous les niveaux syntaxiques (inter-proposition, inter-syntagme et intra-syntagme), ils ont tendance à produire proportionnellement davantage de pauses intra-syntagmes, mais pas significativement plus de pauses inter-propositionnelles. Une tendance similaire est observée chez les locuteurs japonophones par rapport aux locuteurs natifs. Autrement dit, c'est bien au niveau de la fréquence de pauses aux frontières syntaxiques de bas niveau que s'observe la principale différence entre les locuteurs B1 et B2.

Nous avons proposé un score de distribution syntaxique des pauses qui permet de mesurer la tendance de positionnement des pauses dans l'énoncé d'un locuteur. Qu'il soit basé sur les types de constituants ou le niveau de profondeur des frontières syntaxiques, ce score confirme que plus le niveau de compétence du locuteur est élevé, plus les pauses sont positionnées aux frontières syntaxiques de haut niveau, avec les locuteurs natifs atteignant les scores les plus élevés.

L'analyse des annotations automatiques de proéminence syllabique révèle que les locuteurs de niveau B2 francophones positionnent en général mieux l'accent et produisent un contraste acoustique plus marqué entre la syllabe accentuée et les autres syllabes du mot. Toutefois, les mesures individuelles présentent une grande variabilité : le score de position de l'accent s'étend de 0 à 65 % dans le corpus CLES-FR, avec un chevauchement important entre les locuteurs B1 et B2 malgré une différence significative. Les annotations mettent également en évidence l'influence notable des schémas accentuels de la langue maternelle des locuteurs. Les francophones tendent souvent à augmenter la f_0 et à allonger la syllabe finale des mots, tandis que l'intensité

reste stable, quelle que soit la position attendue de l'accent. Cette tendance diminue à mesure que le niveau du locuteur augmente, avec une meilleure maîtrise de la f_0 et de l'intensité chez les locuteurs B2.

Chez les locuteurs japonophones, la tendance à accentuer la syllabe finale est beaucoup moins marquée. Ces locuteurs obtiennent globalement de meilleurs scores de position de l'accent que les francophones (45 % en moyenne, contre 36 % pour les francophones) et mobilisent fortement la f_0 et l'intensité dès le niveau B1. Toutefois, le faible nombre de locuteurs dans ce corpus n'a pas permis de détecter des différences significatives entre les niveaux. Enfin, l'analyse des locuteurs natifs a révélé une forte tendance à l'allongement de la syllabe finale quelle que soit l'accentuation, et un contraste de f_0 limité, entraînant des scores de position globalement faibles (41 % en moyenne). En revanche, le contraste d'intensité est plus marqué chez les locuteurs natifs et très corrélé avec le score de position de l'accent.

		CLES-FR				CLES-JP					CLES-EN		
Metric	**Definition**	**B1**	**B2**	**B1/B2 (p)**	**Cliff Δ**	**B1**	**B2**	**C1**	**B1/B2 (p)**	**Cliff Δ**	**Native**	**JPB1B2/EN (p)**	**Cliff Δ**
N_{tokens}	*Number of tokens*	376	422	B1<B2		357	525	1021	B1<B2		1274	JP<EN ***	-0,92
SR	*Speech rate (token/min)*	96	107	B1<B2 ***	-0,35	59	83	133	B1<B2 **	-0,813	160	JP<EN ***	-1
$\overline{F_p}$	*Number of pauses per token*	0,32	0,29	B1>B2 *	0,154	0,43	0,36	0,24	B1>B2 *	0,733	0,18	JP>EN ***	0,98
$\overline{d_p}$	*Mean duration of pauses*	598	591	B1>B2		807	673	613	B1>B2 **	0,867	566	JP>EN ***	0,82
$F_{p,BC}$	*Number of BC pauses per BC boundary (%)*	47	42	B1>B2 ***	0,311	58	47	36	B1>B2 **	0,84	29	JP>EN ***	0,94
$F_{p,WC}$	*Number of WC pauses per WC boundary (%)*	28	25	B1>B2	0,172	39	31	20	B1>B2 *	0,667	15	JP>EN ***	0,947
$F_{p,BP}$	*Number of BP pauses per BP boundary (%)*	29	26	B1>B2	0,149	43	33	21	B1>B2 *	0,627	17	JP>EN ***	0,93
$F_{p,WP}$	*Number of WP pauses per WC boundary (%)*	21	18	B1>B2 *	0,187	30	23	16	B1>B2	0,52	10	JP>EN ***	0,9
$P_{p,BC}$	*Number of BC pauses per number pauses (%)*	35	36	B1<B2	-0,069	36	35	37	B1>B2		37	JP<EN *	-0,4
$P_{p,WC}$	*Number of WC pauses per number pauses (%)*	65	64	B1>B2		65	65	63	-		63	JP>EN *	-0,4
$P_{p,BP}$	*Number of BP pauses per number pauses (%)*	52	52	-	-0,069	53	53	51	-		52	JP>EN	
$P_{p,WP}$	*Number of WP pauses per number pauses (%)*	12	11	**B1>B2 ***	**0,216**	12	14	12	B1<B2		10	**JP>EN ***	**0,483**
DSP_i	*Pauses syntactic distrib. based on boundary type*	0,49	0,52	B1<B2 *	-0,198	0,5	0,46	0,51	B1>B2		0,54	**JP<EN ****	**-0,527**
DSP_n	*Pauses syntactic distrib. based on boundary depth*	0,04	0,1	**B1<B2 *****	**-0,301**	0,08	0	0,13	B1>B2		0,18	**JP<EN *****	**-0,707**
N_{poly}/N_{tokens}	*Number of lexical polysyllabic words per token*	0,21	0,23	B1<B2 **	-0,238	0,26	0,23	0,24	B1>B2		0,22	JP<EN	
N_{ann}	*Number of words annotated with PLSPP*	32	41	B1<B2 **	-0,245	35	49	103	B1<B2		79	JP<EN ***	-0,707
N_{ann}/N_{tokens}	*Number of N_{ann} per token*	0,09	0,1			0,09	0,1	0,09	B1<B2		0,07	JP>EN **	0,627
N_{ann}/N_{poly}	*Number of N_{ann} per N_{poly}*	0,42	0,43			0,4	0,45	0,39	B1<B2		0,3	JP>EN ***	0,753
S	*Stress position score*	30,8	36,8	**B1<B2 ****	**-0,275**	45,8	49,3	49,6	B1<B2		41,2	JP>EN	
$S_{initial}$	*S for words with initial stress*	25	32	B1<B2 **	-0,269								
S_{medial}	*S for words with medial stress*	30	36	B1<B2	-0,158								
S_{final}	*S for words with final stress*	77	77	-	-0,032								
$\overline{C}$	*Mean prosodic contrast*	-6,84	-3,15	**B1<B2 *****	**-0,389**	-1,29	1,4	-0,12	B1<B2		-2,1	JP>EN	
$\overline{C_{f0}}$	*Mean f_0 constrast*	-7,38	-1,44	B1<B2 ***	-0,317	8,1	7,48	3,42	B1>B2		-0,78	JP>EN	
$\overline{C_{int}}$	*Mean intensity contrast*	-5,82	0,77	B1<B2 ***	-0,45	9,87	9,64	11,3	B1>B2		6,57	JP>EN	
$\overline{C_{dur}}$	*Mean duration contrast*	-9,84	-9,36	B1<B2	-0,018	-24,97	-6,76	-11,09	B1<B2		-9,97	JP<EN	

Tab. 8.6 : *Tableau de résultats d'analyses en parole spontanée.*
Chaque métrique est calculée par locuteur, le tableau indique la valeur médiane par groupe de locuteurs.
*Le niveau de significativité de la différence entre les groupes de locuteurs est indiqué par * pour $p < 0{,}05$, ** pour $0{,}01$ et *** pour $0{,}001$.*

Chapitre 9

Mesure de l'impact du rythme sur la compréhensibilité

Ce chapitre présente les résultats obtenus lors de l'expérience d'évaluation dynamique de la compréhensibilité que nous avons organisée en février 2024. Ces résultats ont d'abord été présentés lors du colloque English Pronunciation : Issues and Practice (EPIP8) en mai 2024 (Coulange et al., 2024b), puis ont fait l'objet d'une publication présentée à InterSpeech 2024 (Coulange et al., 2024c). Une analyse approfondie des 800 commentaires libres recueillis durant l'expérience est en cours et sera présentée prochainement, à la conférence New Sounds 2025 (Nakanishi & Coulange, 2025). Dans ce chapitre, nous présentons d'abord le comportement des évaluateurs qui ont participé à l'expérience, avant d'exposer les résultats des évaluations globales des enregistrements, puis l'analyse de l'évaluation dynamique de la compréhensibilité.

9.1 Comportements des évaluateurs

Nous avons recueilli l'évaluation de 16 stimuli audio par 60 participants. Les données récoltées consistent en une suite de *timestamps* correspondant aux temps où les évaluateurs ont cliqué pendant la lecture de chaque stimulus pour signaler qu'ils éprouvent une difficulté de compréhension, ainsi qu'une évaluation globale sur trois dimensions pour chaque stimulus (qualité de la prononciation, fluidité et facilité de compréhension). L'évaluation des 16 stimuli a duré en moyenne 26 min 59 s pour les 60 participants, allant de 12 min 42 s à 1 h 3 min 2 s, dont 4 évaluateurs excédant 45 min.

Le coefficient de corrélation intra-classe (ICC) pour l'accord absolu entre les évaluateurs pour l'évaluation globale est de 0,97, avec un intervalle de confiance à 95

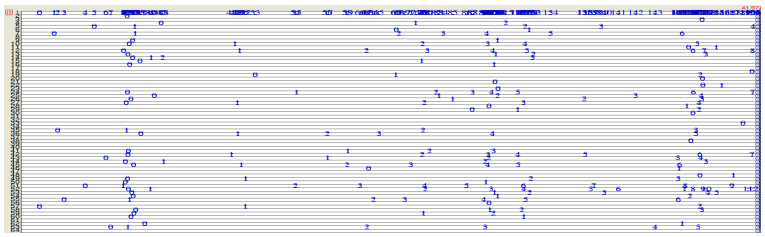


FIG. 9.1 : Aperçu des clics enregistrés par les 60 participants sur l'un des stimuli (format TextGrid, un point représente un clic, un évaluateur par tier, la première tier est la somme de l'ensemble des clics)

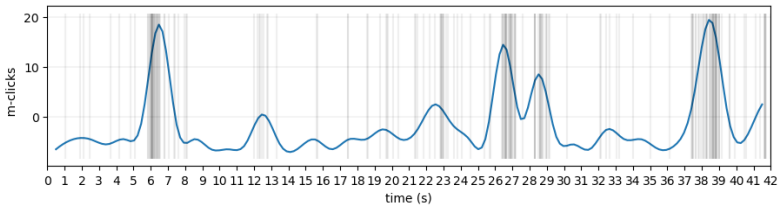


FIG. 9.2 : Somme des m-clics sur une fenêtre glissante d'une seconde, pour le même enregistrement (les clics bruts sont représentés par une barre verticale grise)

% de $[0{,}95; 0{,}99]$. Cet ICC élevé indique une forte cohérence entre les évaluateurs ($F(55, 15) = 55$, $p < 0{,}001$), confirmant la fiabilité des évaluations sur les dimensions évaluées. L'alpha de Cronbach calculé pour évaluer la cohérence interne des 60 évaluations sur les trois dimensions est de 0,93 ($IC = [0{,}93; 0{,}94]$) et indique une forte fiabilité des évaluations. Si l'une des dimensions était retirée, l'alpha de Cronbach resterait au-dessus de 0,89 pour chacune, ce qui montre que chaque dimension contribue de manière significative à la cohérence interne globale des évaluations. Les coefficients de corrélation item-total pour chaque dimension sont également élevés (prononciation : 0,86 ; fluidité : 0,85 ; compréhensibilité : 0,88), ce qui montre que chaque dimension est bien corrélée avec l'ensemble des évaluations. Ces valeurs sont proches de celles obtenues dans des études antérieures, comme celle de Kahng (2018) qui avait un accord absolu de 0,93 et un alpha de Cronbach de 0,98 dans une expérience mobilisant 46 évaluateurs et 80 extraits de parole. Ces mesures renforcent la validité des scores obtenus, indiquant que les différences observées entre les enregistrements reflètent bien des différences perçues dues aux signaux de parole, et non à des variations inter- ou intra-individuelles dans le jugement des évaluateurs.

Comme prévu, une grande variabilité de comportement a été observée parmi les évaluateurs, avec un nombre total de clics par évaluateur allant de 12 à 272 sur les 16 enregistrements (moyenne de 76,7, écart type de 48,65). Cinq évaluateurs ont présenté une fréquence de clics particulièrement élevée, totalisant plus de 120 clics chacun.

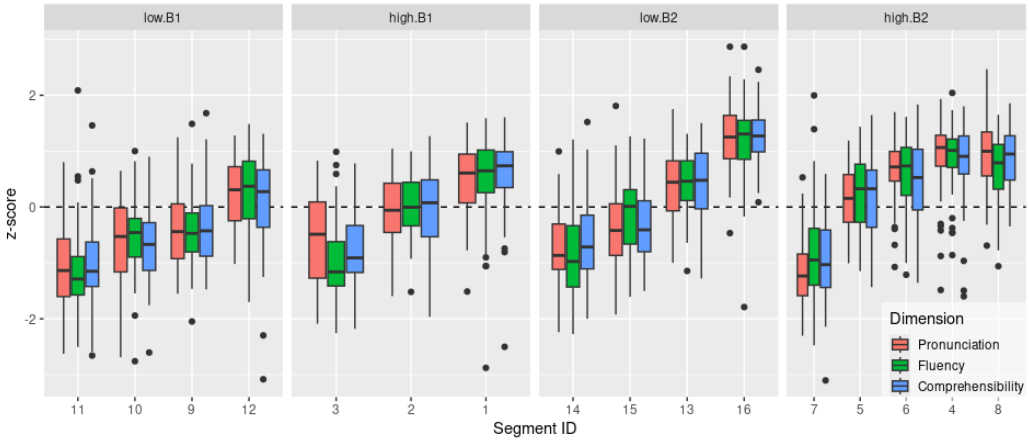


FIG. 9.3 : Évaluation globale normalisée des 16 segments en termes de qualité de prononciation (rouge) de fluidité (vert) et de compréhensibilité (bleu), en fonction du niveau du locuteur et de la catégorie du segment (low : haute proportion de pauses WP et bas score accentuel moyen ; high : basse proportion de pauses WP et haut score accentuel moyen ; un point de donnée correspond à l'évaluation d'un segment par un évaluateur sur une dimension)

Si la fréquence de clic varie d'un participant à l'autre, une tendance claire à cliquer dans les mêmes zones est toutefois observable, se traduisant par des pics de clics relativement bien contrastés comme l'illustrent les figures 9.1 et 9.2.

9.2 Évaluations globales

Commençons par analyser les évaluations globales des enregistrements. Une fois normalisés, les scores de qualité de prononciation, de fluidité et de compréhensibilité apparaissent de manière générale très corrélés entre eux, comme le montre la figure 9.3. On peut voir que les locuteurs B2 ont tendance à obtenir un score global plus élevé que les B1 ($p < 0{,}001$, médianes à -0,31 pour B1 et +0,38 pour B2, $\Delta = -0{,}328$ (faible), $IC = [-0{,}367; -0{,}288]$, figure 9.4 gauche), bien que tous ne reçoivent pas un score positif. De la même façon, les segments catégorisés "high" par PLSPP (c'est à dire avec proportionnellement peu de pauses de type intra-syntagme et un score accentuel élevé) reçoivent un score généralement plus élevé que les segments "low" , bien que le contraste soit moins important que pour le niveau du locuteur ($p < 0{,}001$, médianes à -0,22 pour low et +0,35 pour high, $\Delta = -0{,}202$ (faible), $IC = [-0.243; -0.16]$, figure 9.4 milieu). Notons par ailleurs que les locutrices ont tendance à obtenir un meilleur score ($p < 0{,}001$, médianes à -0,28 pour les hommes et +0,28 pour les femmes, $\Delta = -0{,}245$ (faible), $IC = [-0{,}286; -0{,}204]$, figure 9.4 droite).

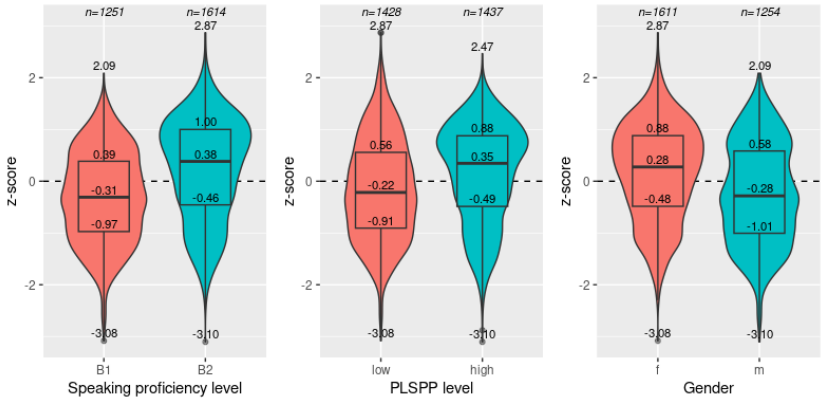


FIG. 9.4 : Évaluation globale normalisée en fonction du niveau, de la catégorie et du genre du locuteur (3 dimensions confondues, un point de donnée correspond à l'évaluation d'un segment par un évaluateur sur une dimension)

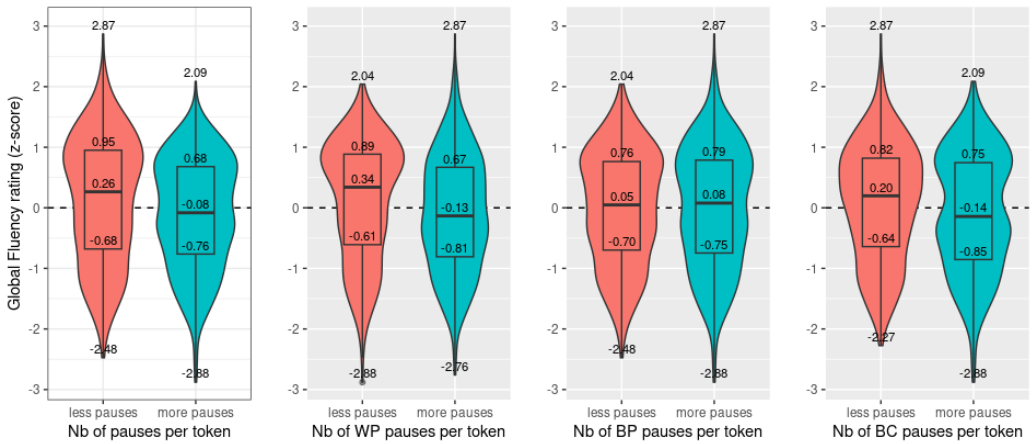


FIG. 9.5 : Évaluation globale de la fluidité selon la fréquence des différents types de pauses

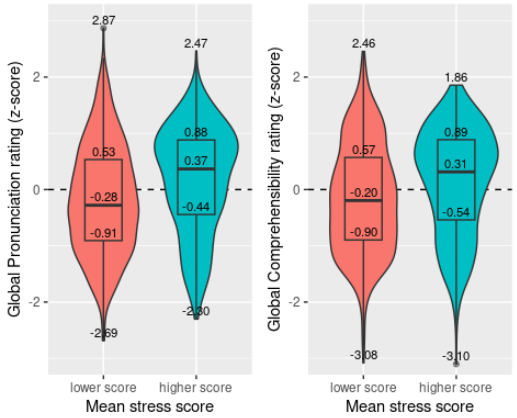


FIG. 9.6 : Évaluation globale de la prononciation et de la compréhensibilité en fonction du score accentuel moyen

Nous avons ensuite regroupé les segments en fonction de leurs tendances pour chaque dimension : ceux qui ont plus de pauses intra-syntagme, ceux qui ont les meilleurs scores accentuels etc.

Commençons par les pauses. On constate tout d'abord que les segments qui contiennent globalement moins de pauses de manière générale (les 8 segments dont le nombre de pauses par token est inférieur à la fréquence médiane) obtiennent un meilleur score de fluidité que les segments qui contiennent plus de pauses ($p < 0{,}001$, médianes à +0,26 contre -0,08, $\Delta = -0{,}138$ (faible), $IC = [-0{,}063; -0.212]$, *cf.* figure 9.5). Ce n'est pas une surprise : les pauses sont souvent perçues comme une disfluence de la parole, il n'est donc pas étonnant que les enregistrements qui en présentent plus soient jugés moins fluides. Mais regardons ce qu'il en est si l'on considère les pauses en fonction de leur position syntaxique : seules les pauses intra-syntagme (WP) présentent une différence significative sur la perception de fluidité des segments ($p < 0{,}001$, médianes à +0,34 contre -0,13, $\Delta = 0{,}156$ (faible), $IC = [0{,}082; 0{,}227]$) ; tandis que la fréquence des pauses inter-syntagme (BP) ne distinguent pas les segments (*ns.*, médianes à 0,05 et 0,08, $\Delta = -0{,}003$ (négligeable), $IC = [-0{,}076; 0.07]$), et que celle des pauses inter-proposition (BC) reste assez floue : les segments qui en ont moins ont tendance à être légèrement mieux notés, mais la taille de l'effet reste négligeable ($p < 0{,}05$, médianes à +0,2 contre -0,14, $\Delta = 0{,}088$ (négligeable), $IC = [0{,}014; 0{,}16]$). On observe les mêmes tendances avec le jugement de compréhensibilité.

Du côté de l'accentuation lexicale, on constate que les segments dont le score accentuel est bas ont tendance à être moins bien notés en termes de prononciation ($p < 0{,}001$, médianes à -0,28 contre +0,37, $\Delta = -0{,}225$ (faible), $IC = [-0{,}296; -0{,}153]$)

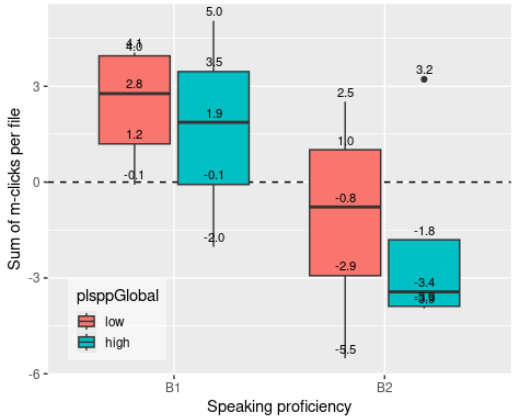


FIG. 9.7 : Somme des clics normalisés par enregistrement, groupés par niveau du locuteur et catégorie de segment

comme de compréhensibilité ($p < 0{,}001$, médianes à -0,20 contre +0,31, $\Delta = -0{,}189$ (faible), $IC = [-0{,}26; -0{,}115]$, *cf.* figures 9.6).

Enfin, nous nous sommes intéressés au nombre total de clics par enregistrement. La figure 9.7 présente le nombre de clics normalisés reçus pour chaque segment audio en fonction du niveau du locuteur et de la catégorie du segment. La différence entre les locuteurs B1 et B2 apparaît très clairement ($p < 0{,}05$, médianes à 1,9 pour B1 et -2,1 pour B2, $\Delta = 0{,}651$ (large) $IC = [0{,}088; 0{,}899]$). Les enregistrements catégorisés *low* obtiennent également plus de clics en moyenne que ceux catégorisés *high*, mais la différence n'est pas significative (médianes à 1,1 pour low et -1,9 pour high, $\Delta = -0{,}188$ (faible), $IC = [-0{,}679; 0{,}42]$).

À ce stade, nous avons vu que les enregistrements qui contiennent plus de pauses en général obtiennent de moins bons scores de fluidité et de compréhensibilité, mais que cette différence n'est significative que lorsqu'on regroupe les segments par fréquence de pauses intra-syntagme ; un plus grand nombre de pauses inter-syntagme ou inter-proposition n'affecte pas autant le jugement des évaluateurs. Nous avons vu également que les segments au score accentuel élevé obtiennent de meilleurs scores de prononciation et de compréhensibilité. Toutefois, s'il y a corrélation, il n'y a pas nécessairement causalité : on ne peut pas affirmer que les pauses ou les mots mal accentués ont un impact direct sur la compréhensibilité. Une évaluation dynamique de la compréhensibilité nous permet d'observer les tendances de fluctuation du jugement à la suite de ces phénomènes, et ainsi mieux comprendre leur impact sur la perception de la parole.

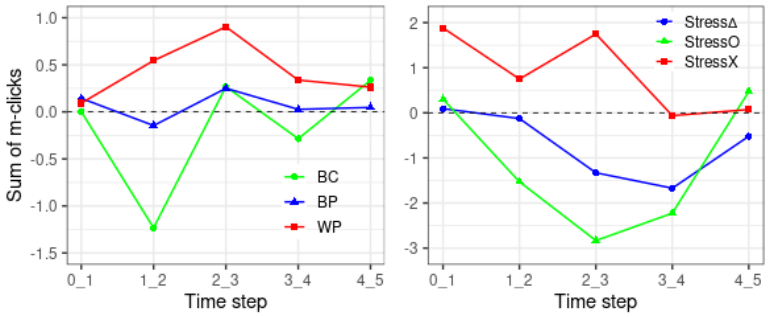


FIG. 9.8 : Nombre de m-clics moyen sur les 5 secondes suivant l'onset d'une pause (gauche) ou d'un mot pattern accentuel (droite) ; les valeurs positives indiquent une activité de clics supérieure à la moyenne

9.3 Analyse des patterns de clics

Nous présentons dans cette section les résultats de l'analyse des variations du nombre de clics normalisé (m-clics) à la suite des pauses et des mots polysyllabiques. Nous utilisons une fenêtre glissante d'une seconde sur les 5 secondes suivant l'onset de l'événement qui nous intéresse. La moyenne des m-clics est calculée dans chaque fenêtre, et comparée selon le type de pause ou de pattern accentuel.

La figure 9.8 (gauche) montre le nombre moyen de m-clics sur les 5 secondes suivant l'onset d'une pause. Les valeurs positives indiquent une activité de clic supérieure à la moyenne. Entre 0 et 1 seconde après l'onset de la pause, le nombre de m-clics est proche de 0 pour tous les types de pauses, indiquant que l'activité de clics est normale. À partir d'une seconde après l'onset, on constate que le nombre de clics tend à augmenter lorsqu'il s'agit d'une pause intra-syntagme (WP), atteignant son maximum (+0,8) entre 2 et 3 secondes après l'onset de la pause. Parallèlement, le nombre de m-clics à la suite d'une pause inter-proposition (BC) décroît nettement (-1,23) entre 1 et 2 secondes, puis revient rapidement vers 0 dès la troisième fenêtre. Dans le cas des pauses inter-syntagme (BP), enfin, le nombre de m-clics semble stagner autour de 0, n'indiquant aucune variation observable de l'activité. Le test de rangs montre une différence significative entre le nombre moyen de m-clics après les pauses WP et BC seulement sur la deuxième fenêtre, entre 1 et 2 secondes, *cf.* tableau 9.1.

En ce qui concerne l'évolution du nombre de clics à la suite des 139 mots polysyllabiques cibles, on constate une tendance assez similaire. Bien que la moyenne de m-clics après les mots dont le pattern accentuel est jugé incorrect par PLSPP (StressX, $C'' < -0{,}2$) reste globalement supérieure à celle des mots jugés corrects (StressO, $C'' > 0{,}2$) ou ambigus (StressΔ), on observe une augmentation locale entre 2 et 3 secondes après l'onset du mot, mais une diminution évidente des clics après les mots

	Rank tests		*Pearson correlations*	
	BC vs. WP	**StressO vs. StressX**	**Stress score**	
window	p-value	p-value	R	p-value
0-1s	–	–	-0.13	–
1-2s	*	*	-0.1	–
2-3s	–	**	-0.25	**
3-4s	–	*	-0.062	–
4-5s	–	–	-0.027	–

Tab. 9.1 : *Tests de rangs comparant le nombre moyen de m-clics après les pauses inter-proposition (BC) et intra-syntagme (WP), et après les patterns accentuels corrects (StressO, $C'' > 0{,}2$) et incorrects (StressX, $C'' < -0{,}2$), et coefficient de corrélation entre le nombre de m-clics et la valeur de C'' (– : non significatif, * : $p < .05$, ** : $p < .01$)*

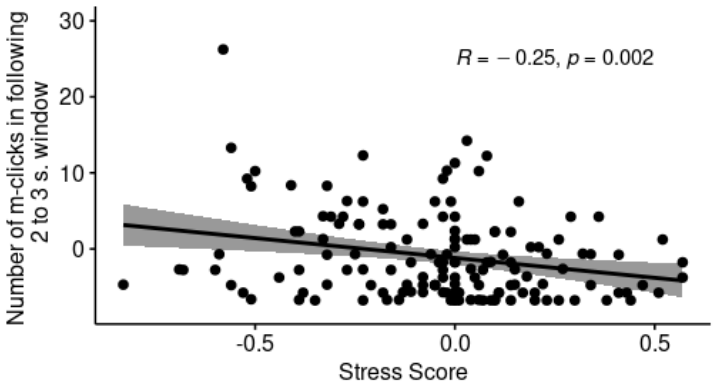


Fig. 9.9 : *Projection des 139 mots cibles en fonction de leur score accentuel C'' et du nombre de m-clics enregistrés dans la fenêtre de 2 à 3 secondes après l'onset du mot*

StressO jusqu'à la troisième seconde (atteignant -2,83). La différence de nombre de m-clics après StressX et StressO est significative entre 1 et 4 secondes après l'onset du mot, *cf.* tableau 9.1.

Comme le score accentuel C'' est une valeur continue, nous avons également mesuré la corrélation entre celui-ci et le nombre de m-clics observés dans chaque fenêtre. La corrélation est négative de 0 à 5 secondes après l'onset, indiquant que plus le score est élevé, moins on observe de clics. La corrélation la plus forte, et la seule qui est significative, est observée entre 2 et 3 secondes : elle reste toutefois relativement faible (-0,25, $p < 0{,}01$, *cf.* tableau 9.1 et figure 9.9).

Conclusion

Le protocole expérimental mis en place dans cette étude a montré que les auditeurs natifs ont tendance à signaler des difficultés de compréhension dans les mêmes zones et que l'évaluation dynamique permet d'identifier ces zones et ainsi quantifier l'impact potentiel de certains phénomènes sur la compréhensibilité du locuteur.

Nous nous sommes concentrés sur l'impact des pauses et des patterns accentuels. Les analyses montrent une augmentation moyenne du nombre de clics sur les trois secondes suivant les pauses intra-syntagme, mais une diminution à la suite des pauses inter-proposition. Les enregistrements présentant plus de pauses en général ont été moins bien notés en termes de fluidité et de compréhensibilité globale ; mais seule la proportion de pauses intra-syntagme semble faire la différence. Nous avons également mesuré une augmentation significative du nombre de clics entre deux et trois secondes après un mot au pattern accentuel incorrect, et à l'inverse une diminution sur trois secondes après un mot accentué sur la bonne syllabe. Ces résultats étaient attendus, puisque plusieurs études ont déjà démontré la corrélation entre la distribution des pauses et des accents lexicaux, mais notre approche propose un moyen de mesurer leur impact en temps réel sur la perception des auditeurs.

Chapitre 10

Discussion

Dans cette thèse, nous avons cherché à concevoir un outil permettant d'évaluer certains aspects de la production orale d'apprenants en langue seconde (L2), en nous appuyant sur des phénomènes linguistiques susceptibles d'entraver la compréhension des auditeurs. L'état de l'art présenté dans la première partie de la thèse a mis en évidence la fluence et le rythme de la parole comme deux paramètres clés influençant la compréhensibilité des locuteurs. Parmi les facteurs sous-jacents, les pauses et l'accent lexical sont apparus comme des composantes essentielles de ces deux paramètres.

Les outils existants pour l'évaluation automatique des pauses et de l'accentuation lexicale présentent toutefois des limites importantes : ils se concentrent souvent sur le calcul d'une fréquence globale des pauses et ne permettent d'évaluer l'accent que sur des mots ou des phrases isolés, généralement produits dans des contextes contrôlés. En tirant parti d'outils open-source de dernière génération, nous avons développé une chaîne de traitements automatiques capable d'annoter la distribution des pauses et de l'accentuation lexicale dans des conversations spontanées en anglais L2.

Si la fréquence des pauses est souvent utilisée comme un indicateur de fluidité, toutes les pauses ne perturbent pas nécessairement la compréhension des auditeurs. Au contraire, certaines pauses jouent un rôle structurant dans l'organisation de l'énoncé et facilitent la segmentation du flux de parole. Par exemple, les pauses en frontière de haut niveau syntaxique (e.g., entre des propositions) contribuent à cette structuration, tandis que celles situées en frontière de bas niveau (e.g., à l'intérieur des syntagmes) sont plus souvent perçues comme des disfluences perturbant la compréhension du message. De manière similaire, l'accent lexical en anglais revêt une importance particulière pour la segmentation du flux de parole : il est généralement porté par la syllabe initiale des mots lexicaux, tandis que les autres syllabes et les mots grammaticaux sont réduits. Accentuer une syllabe qui ne porte pas l'accent lexical,

ou produire un contraste prosodique insuffisant entre syllabes accentuées et non accentuées, peut donc créer un rythme de parole déstabilisant, demandant un effort de compréhension supplémentaire de la part de l'auditeur.

Nous avons testé notre chaîne de traitements sur des corpus de différents types de parole, de langues maternelles et de niveaux d'anglais. Notre analyse a porté en particulier sur trois corpus de conversations spontanées enregistrés dans le cadre de cette thèse, impliquant des locuteurs de niveaux CECRL B1 et B2. Les principales grilles d'évaluation de la production orale en anglais mettent en évidence un seuil de compétence notable en termes de compréhensibilité au niveau B2 ou équivalent, et il nous a ainsi paru pertinent d'analyser les patterns de pauses et d'accentuation entre locuteurs B1 et B2. Par ailleurs, étant donné l'influence notable de la langue maternelle (L1) sur les tendances accentuelles, nous avons comparé deux L1 aux caractéristiques accentuelles distinctes : le français et le japonais.

Le premier corpus, appelé CLES-FR, se compose de jeux de rôles argumentatifs d'une dizaine de minutes, impliquant deux ou trois candidats francophones enregistrés lors d'épreuves d'interaction orale pour la certification CLES. Ce corpus compte 170 locuteurs et totalise environ 16 heures de parole. Un corpus similaire a été enregistré avec 29 locuteurs japonais (CLES-JP, 4 h de parole) et un autre avec 14 locuteurs anglophones natifs (CLES-EN, 2 h de parole).

Enfin, nous avons exploré l'impact des pauses et des patterns accentuels sur la perception de la difficulté de compréhension. Pour cela, nous avons adapté un protocole d'évaluation dynamique de la compréhensibilité, que nous avons soumis à 60 auditeurs anglophones natifs.

Ce chapitre final récapitule les principaux résultats obtenus dans cette thèse, met en lumière les apports de ce travail dans le domaine de l'évaluation en L2 et propose une discussion critique visant à identifier ses limites et à en dégager des perspectives d'amélioration.

10.1 Principaux résultats obtenus

1.1 Annotation des pauses

Nous avons développé un système permettant de calculer la fréquence des pauses en fonction de leur position syntaxique. Ce système repose sur un alignement temporel des mots avec le signal de parole et une analyse grammaticale par constituants des transcriptions. Chaque intervalle identifié comme une pause dans l'alignement est caractérisé selon le type de frontière syntaxique, la taille des constituants adjacents, et la

profondeur syntaxique depuis la racine de l'énoncé. Dans cette thèse, nous avons utilisé Wav2Vec2.0 pour effectuer l'alignement mot-signal. Ce système a la particularité de produire un intervalle vide entre chaque mot de l'énoncé. Cette caractéristique nous a permis de régler précisément les seuils de durée souhaités pour distinguer pauses et simples frontières de mots. Les analyses des trois corpus de parole spontanée ont été conduites avec un seuil de durée minimum de 180 ms et un maximum de 2 s, car les pauses inférieures à 250 ms, seuil commun dans la littérature, se sont révélées efficaces pour discriminer les locuteurs B1 et B2.

Nous avons également proposé une métrique, le score de distribution syntaxique des pauses (DSP), qui reflète la tendance de distribution syntaxique des pauses dans un énoncé. Ce score, variant entre -1 et 1, est calculé comme une somme pondérée des pauses à différents niveaux syntaxiques, normalisée par le nombre total de pauses. Une valeur élevée indique une concentration des pauses aux frontières syntaxiques de haut niveau.

L'analyse des trois corpus de parole conversationnelle a montré une tendance chez les locuteurs de niveau B1 à produire proportionnellement plus de pauses en frontières syntaxiques de bas niveau que les locuteurs B2.

Plus spécifiquement, dans le corpus CLES-FR, les locuteurs B1 produisent davantage de pauses en moyenne, indépendamment du type de frontière syntaxique (inter-propositions, inter-syntagmes ou intra-syntagmes). Cependant, seule la proportion de pauses intra-syntagmes diffère significativement entre B1 et B2 (B1 > B2, $p < 0{,}05$). La proportion de pauses inter-propositions est légèrement inférieure pour les B1, mais cette différence n'est pas significative. Nous avons observé la même tendance entre les locuteurs japonophones B1 et B2, sans toutefois obtenir de différence significative entre les groupes de locuteurs, en raison du faible nombre de locuteurs B1. La différence entre les locuteurs japonophones (B1 et B2 rassemblés) et les locuteurs natifs du corpus CLES-EN a montré quant à elle que la proportion de pauses inter-propositions est significativement plus élevée pour les locuteurs natifs ($p < 0{,}05$), tandis que les pauses intra-syntagmes sont significativement moins fréquentes par rapport aux locuteurs japonais ($p < 0{,}05$).

Nous avons calculé deux scores de distribution syntaxique des pauses, l'un basé sur le type de frontière syntaxique (DSP_i), l'autre sur le niveau de profondeur des frontières estimé à partir du nombre de constituants se fermant ou s'ouvrant au niveau de la pause (DSP_n). Les deux scores révèlent des tendances similaires : les pauses sont plus souvent placées en frontières de bas niveau syntaxique chez les B1 par rapport aux B2, et chez les locuteurs japonais par rapport aux locuteurs natifs. En outre, le DSP_n s'est montré plus discriminant que le DSP_i (entre B1 et B2 francophones : $p < 0{,}001$, $\Delta = -0{,}301$, contre $p < 0{,}05$, $\Delta = -0{,}198$ pour DSP_i ; entre japonophones et

anglophones natifs : $p < 0{,}001$, $\Delta = -0{,}707$, contre $p < 0{,}01$, $\Delta = -0{,}527$ pour DSP_i).

Ces résultats corroborent les conclusions de l'état de l'art :

- La position des pauses est fortement contrainte par la syntaxe, avec une préférence marquée pour les frontières de haut niveau.
- Les locuteurs non-natifs ont tendance à produire plus de pauses en frontières de bas niveau (Fauth & Trouvain, 2018), et plus le niveau de compétence augmente, plus les pauses ont tendance à se concentrer aux frontières de haut niveau (de Jong, 2016).
- Nos observations concordent avec celles de Kahng (2018) et Suzuki et Kormos (2020), selon lesquelles un nombre élevé de pauses intra-propositions est associé à une perception de fluence réduite.
- Enfin, conformément à Kallio et al. (2022), nous avons constaté que la fréquence des pauses intra-syntagmes est un indicateur plus discriminant que celle des pauses inter- et intra-propositions.

1.2 Annotation de l'accent lexical

Nous avons proposé une méthode permettant d'annoter automatiquement les syllabes en fonction de leur degré de proéminence acoustique. Cette méthode repose sur l'alignement mot-signal et sur la détection des noyaux syllabiques à partir des pics d'intensité. Pour chaque noyau identifié, une extraction de la fréquence fondamentale (f_0), de l'intensité et de la durée est réalisée. Ces valeurs sont ensuite normalisées pour estimer la syllabe la plus proéminente, dont la position est comparée à celle de la syllabe portant l'accent primaire selon un dictionnaire phonologique de référence.

Cette méthode permet de calculer deux scores principaux : le score de position de l'accent, qui mesure la proportion de mots dont la syllabe proéminente correspond à celle attendue, et le score de contraste prosodique, qui représente le degré de contraste entre la syllabe accentuée et les autres syllabes sur chaque dimension prosodique. Ces scores permettent ainsi d'établir un profil accentuel pour chaque locuteur, caractérisant sa manière d'accentuer les mots.

Une seconde version de l'outil, intégrant un alignement phonologique, a permis d'affiner les mesures acoustiques en ciblant les voyelles des syllabes. Cependant, cette version s'est avérée moins performante pour la parole spontanée, où les hésitations nuisent à la précision de l'alignement. Nous avons donc conservé la première version pour analyser les corpus CLES.

Les annotations du corpus CLES-FR montrent des résultats variés entre locuteurs, mais des scores globalement plus élevés pour les B2 par rapport aux B1. Toutefois, les scores de position restent faibles, avec une médiane à 30,8 % pour les B1 et 36,8 % pour les B2, et les contrastes prosodiques (f_0, intensité, durée) demeurent limités. Les locuteurs ayant un score de position élevé marquent principalement la syllabe accentuée en f_0 et en intensité, tandis que ceux ayant un score faible tendent à ne pas produire de contraste d'intensité, et produire un allongement marqué de la syllabe finale accompagné d'une montée de f_0.

Dans le corpus CLES-JP, la taille limitée de l'échantillon n'a pas permis de distinguer significativement les niveaux B1 et B2. Néanmoins, les locuteurs japonais obtiennent en général de meilleurs scores que les francophones (45,8 % pour les B1, 49,3 % pour les B2, et 49,6 % pour les C1). Le contraste de f_0 est celui qui est le plus fortement corrélé au score de position de l'accent ($R = 0{,}65$, $p < 0{,}001$). Les locuteurs japonophones montrent par ailleurs une tendance moins marquée à accentuer la syllabe finale (56 % des mots, contre 71 % et 65 % chez les B1 et B2 francophones). De plus, la f_0 et l'intensité sont mobilisées dès le niveau B1.

La comparaison avec les anglophones natifs du corpus CLES-EN révèle les limites de l'outil pour évaluer la parole native en contexte spontané. Le contraste d'intensité est le seul paramètre systématiquement positif, tandis que ceux de f_0 et de durée apparaissent influencés par des facteurs externes, limitant ainsi l'interprétation des résultats. Nous revenons sur ces limitations en section 3.2.

Les résultats confirment l'influence des patterns accentuels de la L1 sur la production en L2. Les locuteurs francophones accentuent davantage cette syllabe, principalement par allongement de la durée, une tendance bien documentée (Astesano, 2001; Tortel & Hirst, 2010). Bien que cette tendance diminue avec l'augmentation du niveau de compétence, elle reste notable comparée aux locuteurs japonais ou anglophones. Les locuteurs japonais, habitués à un accent lexical en L1, positionnent quant à eux l'accent avec plus de précision et produisent des contrastes prosodiques plus marqués que les francophones, comme le rapportent également Dupoux et al. (1997), Sugahara (2016) ou Cutler (2015).

1.3 Impact des pauses et de l'accent sur l'auditeur

Nous avons également exploré l'impact des pauses et de l'accentuation sur la perception des auditeurs. Pour cela, nous avons extrait 16 segments de parole issus du corpus CLES-FR et les avons soumis à l'évaluation de 60 auditeurs anglophones natifs. Ces segments se répartissaient en deux groupes : l'un présentant un taux élevé de pauses intra-syntagmes et un contraste prosodique négatif, et l'autre un faible taux

de pauses intra-syntagmes et un contraste prosodique positif. Lors de l'évaluation, les auditeurs devaient cliquer chaque fois qu'ils ressentaient un effort pour comprendre le locuteur. Ces clics ont permis de mesurer l'évolution du degré de difficulté perçue au fil de chaque enregistrement.

L'analyse des résultats a mis en évidence un lien direct entre les pauses, l'accentuation et la perception de l'effort de compréhension. Plus précisément, la fréquence de clics normalisée augmente dans les trois secondes suivant le début des pauses intra-syntagmes, tandis qu'elle diminue dans les deux secondes suivant les pauses inter-propositions. Par ailleurs, les mots présentant un contraste prosodique négatif entraînent une augmentation de la fréquence de clics entre deux et trois secondes après leur occurrence, tandis qu'un contraste prosodique positif est associé à une diminution notable de la fréquence de clics dans les trois secondes qui suivent.

Ces résultats confirment que les pauses de bas niveau syntaxique et les patterns accentuels incorrects ont un effet négatif sur la compréhensibilité du locuteur, en accord avec les travaux de Isaacs et Trofimovich (2012), Saito et al. (2015) et Suzuki et Kormos (2020). Nos observations montrent qu'une évaluation dynamique de la compréhensibilité est possible et pertinente, à condition de simplifier le protocole d'évaluation et de normaliser les patterns de clics pour tenir compte des variations individuelles des auditeurs.

10.2 Apports de notre travail

Ce travail de recherche apporte plusieurs contributions significatives à l'évaluation en L2.

Premièrement, nous avons démontré qu'il est possible d'évaluer automatiquement certains aspects de la production orale spontanée influençant directement la compréhensibilité des locuteurs. La chaîne de traitement que nous avons développée constitue un premier prototype combinant des outils d'identification des locuteurs et de reconnaissance de la parole pour évaluer la production orale dans des contextes de communication réalistes.

Ensuite, notre méthodologie ne repose pas sur la comparaison avec un modèle ou avec des tendances observées chez les locuteurs natifs, mais se base sur l'identification de phénomènes susceptibles d'entraver la compréhension. L'outil PLSPP développé dans le cadre de cette thèse est principalement destiné à annoter automatiquement la parole spontanée. À partir de ces annotations, nous avons proposé plusieurs métriques permettant de comparer les productions de différents groupes de locuteurs.

Notre chaîne de traitement, bien qu'adaptable, n'a pas vocation à être utilisée en l'état dans des applications finales. Son aspect modulaire lui permet de s'adapter à des données et des types de parole variés. Ainsi, PLSPP a déjà évolué en différentes versions pour s'adapter à des corpus incluant des données de parole lue, récitée, ou spontanée, produites par des locuteurs enfants et adultes de diverses langues maternelles (japonais, coréen, slovaque, portugais ou anglais) (Erickson et al., 2025 ; Kimura et al., 2024 ; Nakanishi & Coulange, 2024 ; Nishioka et al., 2025 ; Raso et al., 2024 ; Sugahara et al., 2023, 2024). Le code source de PLSPP est accessible ici[1].

En termes d'évaluation de la fluence, nous avons proposé une métrique de distribution syntaxique des pauses qui, à notre connaissance, est sans équivalent dans la littérature. Cette métrique synthétise en une valeur la capacité d'un locuteur à concentrer ses pauses aux frontières syntaxiques de haut niveau, reflétant ainsi sa compétence à structurer son discours de manière fluide et compréhensible.

Concernant l'évaluation du rythme, notre apport a consisté à proposer un outil qui 1) est utilisable en parole spontanée, 2) mesure le contraste prosodique entre les syllabes, de manière à pouvoir évaluer le degré de marquage de la syllabe accentuée, ou au contraire le degré de réduction des syllabes non accentuées, et 3) caractérise l'accentuation des syllabes en termes de f_0, d'intensité et de durée. Nous n'avons trouvé aucun autre système automatique qui permette cette évaluation : ceux-ci se limitent généralement à identifier la position des syllabes accentuées, parfois en proposant trois classes d'accentuation comme Shahin et al. (2016) ou Ferrer et al. (2015), mais sans caractériser la façon dont l'accentuation est réalisée ni le degré avec lequel les syllabes sont accentuées.

Nous avons également développé une application de visualisation des patterns de pauses et d'accentuation. Bien qu'encore perfectible, cet outil facilite l'exploration et la visualisation des annotations générées par PLSPP. Son code source est également mis à disposition[2].

Les trois corpus de conversations spontanées élaborés dans cette thèse représentent une contribution précieuse, dans la mesure où ils mettent à disposition de la communauté une grande quantité d'enregistrements d'apprenants, accompagnés d'une évaluation précise du niveau de production orale par des évaluateurs experts effectuée sur la base de ces productions. Les corpus sont accessibles aux adresses suivantes : CLES-FR[3], CLES-JP[4] et CLES-EN[5].

[1] https://gricad-gitlab.univ-grenoble-alpes.fr/lidilem/plspp
[2] https://gricad-gitlab.univ-grenoble-alpes.fr/lidilem/plsppviz
[3] https://hdl.handle.net/11403/cles-spontaneous-english
[4] https://hdl.handle.net/11403/cles-jp_corpus
[5] https://hdl.handle.net/11403/cles-en_corpus

Enfin, nos résultats sur l'évaluation dynamique de la compréhensibilité confirment que la position syntaxique des pauses et la réalisation de l'accent lexical ont un impact direct sur la perception de l'effort de compréhension chez l'auditeur, mais aussi que cet impact est mesurable. Le protocole d'évaluation que nous avons adapté de Nagle et al. (2019) devrait permettre de mesurer l'impact d'autres phénomènes linguistiques sur l'auditeur, et faire avancer notre compréhension des liens entre production orale et compréhension. Une étude est en cours (Nakanishi & Coulange, 2025) pour identifier les causes des pics d'effort perçus par les 60 auditeurs de notre expérimentation, d'une part à partir d'analyses syntaxiques, lexicales et acoustiques des enregistrements, mais aussi à partir des quelques 800 commentaires recueillis durant l'expérience. Par ailleurs, notre méthodologie de normalisation des patterns de clics a inspiré une autre étude similaire (Frost et al., 2024), visant à identifier les facteurs de difficulté de compréhension sur un autre corpus. Enfin, nous avons mis à disposition le code source de l'application web[6] utilisée pour l'expérience, avec l'espoir qu'elle puisse être adaptée pour de futures recherches.

10.3 Limites & perspectives

Ce travail de recherche présente plusieurs limites que nous avons identifiées au fil des analyses. Dans cette section, nous proposons de revenir sur les limites des corpus de conversations que nous avons analysés, les limites méthodologiques et techniques de l'outil de mesure et des métriques d'évaluation utilisées, ainsi que sur les limitations inhérentes au protocole utilisé pour l'évaluation dynamique de la compréhensibilité.

3.1 Limitations de corpus

Les sessions d'interaction orale du CLES B2 offrent l'avantage d'évaluer la production orale en contexte conversationnel. Cependant, cette conversation n'est pas écologique, dans la mesure où elle est simulée dans le cadre d'un examen. Le stress associé à cette situation, et le fait qu'il s'agisse d'un jeu de rôle où les candidats doivent parfois défendre des points de vue différents des leurs, peut influencer la production des participants. Toutefois, cette difficulté étant partagée par tous les participants des trois corpus, elle ne devrait pas avoir un impact significatif sur les comparaisons que nous avons faites entre les groupes de locuteurs.

Par ailleurs, il est difficile d'estimer dans quelle mesure les candidats adaptent leur discours en fonction du niveau de leur binôme. Si ce dernier éprouve des difficultés à comprendre ou à interagir, il est naturel que le candidat simplifie ses énoncés,

[6] https://gricad-gitlab.univ-grenoble-alpes.fr/lidilem/dynamic-rater

voire adopte une prononciation moins authentique afin d'être mieux compris. Comme l'ont souligné Nagle et al. (2022), une conversation implique nécessairement un alignement entre les interlocuteurs. D'après eux, ce phénomène de convergence linguistique se traduit par l'appropriation mutuelle d'expressions, de structures syntaxiques, mais aussi, dans une certaine mesure, du rythme de parole du partenaire. Une conversation entre le candidat et un examinateur aurait permis d'atténuer dans une certaine mesure la variable du niveau de l'interlocuteur. Cela dit, ce biais d'adaptation à l'interlocuteur est probablement limité par le contexte d'évaluation. Bien que la conversation ait lieu entre les candidats, le véritable destinataire des locuteurs reste l'évaluateur, présent dans la salle.

Enfin, le choix d'un contexte conversationnel pour analyser les patterns de pauses s'est révélé être une autre limite. Il est en effet difficile de distinguer les pauses destinées à gérer les tours de parole de celles ayant pour fonction de structurer l'énoncé ou résultant d'une hésitation. L'analyse du corpus de Mareková et Beňuš (2024) a d'ailleurs montré l'importante proportion des pauses inter-tours dans leur contexte conversationnel. Dans un contexte monologal, des patterns de pauses différents auraient probablement été observés.

3.2 Limitations des analyses de parole

Les différentes études réalisées à l'aide de l'outil PLSPP ont permis de mettre en évidence plusieurs limitations techniques et méthodologiques, qui doivent être prises en compte lors de l'interprétation des résultats.

Pré-traitements

Le premier module de PLSPP a pour objectif de découper la conversation en segments de parole, correspondant approximativement aux tours de parole ou à leurs sous-unités, pour les analyser indépendamment dans les étapes suivantes. Bien que cette approche simplifie les traitements en aval, une recontextualisation des segments s'avère nécessaire pour interpréter les mesures au regard du contexte conversationnel.

De plus, la compilation des segments pourrait être optimisée pour inclure systématiquement tous les segments formant un tour de parole complet. La position d'un segment dans le tour de parole influence probablement les patterns de pauses observés. Nous avons également introduit un seuil paramétrable pour tolérer, dans une certaine mesure, les réactions de l'interlocuteur dans un segment de parole. Une adaptation dynamique de ce seuil, en fonction des caractéristiques spécifiques de la conversation (par exemple, une faible fréquence de chevauchements de parole), serait bénéfique.

Par ailleurs, il serait pertinent d'identifier les zones de chevauchement de parole, de manière à pouvoir signaler les annotations effectuées en contexte de chevauchement, et ainsi permettre de les isoler si besoin.

Au niveau de la reconnaissance automatique de la parole, la question qui se pose est de savoir s'il vaut mieux transcrire l'ensemble de la conversation puis la découper en segments de parole, au lieu de segmenter d'abord, puis transcrire segment par segment comme nous l'avons fait. La première option permettrait au système de reconnaissance de bénéficier du contexte global de la conversation, ce qui pourrait améliorer la précision des transcriptions. Cependant, cette approche implique la transcription de longs enregistrements contenant les voix de plusieurs locuteurs. Les systèmes récents, comme les dernières versions de WhisperX (Bain et al., 2023), semblent capables de gérer efficacement ce type de transcription de dialogues, ce qui en fait une piste intéressante à explorer.

L'analyse syntaxique, enfin, repose sur la transcription orthographique des segments de parole, sans intégrer le contexte global de la conversation ni les informations prosodiques. Pourtant, ces dernières peuvent fournir des indices syntaxiques précieux. À l'avenir, il serait utile d'employer un modèle d'analyse syntaxique capable de travailler directement sur le signal de parole, sans passer par la transcription (Pupier et al., 2024). De plus, les modèles utilisés dans cette étude, comme `en_core_web_md v3.6`, sont principalement entraînés sur des corpus écrits ou, dans une moindre mesure, sur des corpus oraux de parole contrôlée (par exemple, OntoNotes5, Weischedel et al., 2013). Ces modèles ne sont pas bien adaptés à l'analyse de la parole spontanée et conversationnelle, ce qui constitue une limite pour l'interprétation des résultats obtenus.

Annotations des pauses

Le module d'annotation des pauses a été conçu de manière à ne pas imposer de seuils prédéfinis pour la durée minimale et maximale des pauses. Cependant, pour les analyses, un seuil commun a dû être adopté pour tous les segments étudiés. Bien que cette approche ait permis une certaine uniformité, il apparaît, comme souligné dans la première partie de cette thèse, qu'un ajustement des seuils en fonction du débit de parole des locuteurs serait plus pertinent. À l'instar de la normalisation des mesures acoustiques des syllabes, une démarche consistant à normaliser la durée des pauses pourrait permettre de définir des seuils minimaux adaptés à chaque locuteur, voire à chaque segment.

Lors de l'analyse des distributions des pauses, l'attention a d'abord été portée sur les types de constituants (propositions et syntagmes). Toutefois, cette approche s'est révélée limitée en raison de la rareté des frontières intra-syntagmes en anglais et

du grand nombre de frontières inter-syntagmes laissées de côté. Le calcul du score de distribution syntaxique des pauses (DSP_n), reposant sur l'importance relative des frontières syntaxiques (estimée à partir du nombre de constituants se fermant ou s'ouvrant), a permis de pallier ces limites en intégrant l'imbrication des propositions et des syntagmes les uns dans les autres. Cependant, cette méthode nécessite de définir des seuils d'importance pour le calcul du score. Il serait intéressant de réfléchir à une approche plus continue et moins catégorielle.

Par ailleurs, l'analyse actuelle se concentre exclusivement sur la position des pauses par rapport aux constituants syntaxiques et ne tient pas compte des pauses d'emphase ou stylistiques. Ces dernières, bien que parfois situées en frontières de bas niveau, ne perturbent pas nécessairement la compréhension de l'auditeur (Cao & Chen, 2019). La prise en compte de ces pauses constitue un défi méthodologique pour lequel nous n'avons pas encore identifié de piste d'exploration.

Annotations de l'accent lexical

L'annotation de l'accent lexical représente actuellement la partie de PLSPP qui a le plus de pistes de développement. Les limitations identifiées concernent en particulier les mesures de durée, de f_0, et la méthode de normalisation.

Mesures de durée L'un des principaux problèmes observés jusque-là est une tendance de l'outil à détecter la proéminence sur la syllabe finale en parole spontanée. Cette tendance n'a pourtant pas été constatée en parole lue chez les locuteurs natifs (*cf.* chapitre 7). La cause identifiée est l'allongement fréquent des syllabes finales, parfois accompagné d'une montée intonative, qui semble particulièrement présent dans ce type de parole conversationnelle argumentative. La première version de PLSPP, utilisée pour annoter les corpus CLES, est particulièrement sensible aux allongements de durée, car celle-ci est mesurée sur l'ensemble de la syllabe, et non spécifiquement sur l'intervalle vocalique comme c'est le cas dans les versions suivantes de PLSPP. Toutefois, même en ne mesurant que l'intervalle vocalique, une tendance à l'allongement en finale reste observable, et impacte la précision de détection de la proéminence. Une solution à ce problème pourrait consister à pondérer la durée de la syllabe finale par une constante calculée à partir de l'allongement moyen observé sur l'ensemble des mots d'un locuteur.

Cette limitation concernant la mesure de durée est également due à la précision de l'alignement mot-signal effectué par Wav2Vec2.0 (Baevski et al., 2020), dans la première version de PLSPP. En effet, si cet aligneur s'est révélé plus robuste à la parole spontanée que les autres aligneurs testés, il a toutefois tendance à raccourcir légèrement la durée des mots, en tronquant une partie du début ou de la fin. Ce constat

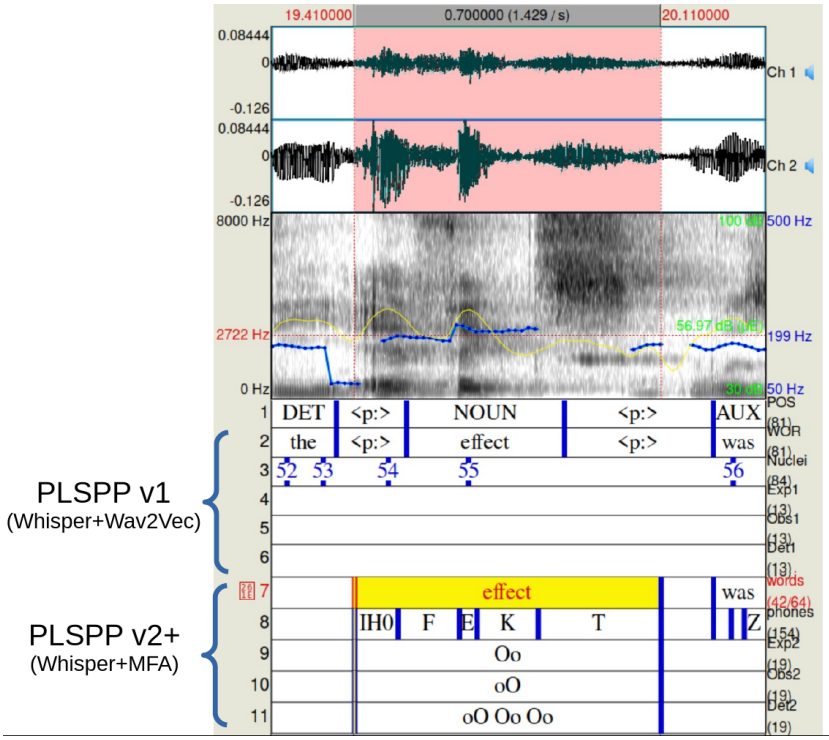


FIG. 10.1 : Illustration d'un cas d'alignement de mot trop court avec Wav2Vec2.0, bloquant l'analyse de proéminence. En comparaison, l'alignement avec MFA est plus adéquat ici. La tier 9 correspond au gabarit accentuel attendu, la 10 à celui qui est observé (accent en finale), la 11 détaille la position de proéminence détectée selon la f_0, l'intensité et la durée.

était déjà fait lors de l'évaluation du module d'alignement, lors duquel Wav2Vec2.0 a obtenu une meilleure précision mais un moins bon rappel que les autres systèmes (*cf.* chapitre 1.3). Cela engendre deux effets majeurs : une sous-estimation de la durée des syllabes initiales et finales, qui impacte alors la détection de la proéminence ; et une diminution importante des mots annotés, puisque les mots pour lesquels le nombre de pics d'intensité détectés ne correspond pas au nombre de syllabes attendues sont éliminés. Ce problème est particulièrement prononcé chez les locuteurs au débit rapide, dont la proportion de mots annotés est considérablement affectée (annotation de seulement 30 % des mots polysyllabiques lexicaux en moyenne chez les locuteurs natifs, contre 43 % chez les francophones). La figure 10.1 illustre ce phénomène : on peut constater que le pic d'intensité de la première syllabe du mot "*effect*" (n°54, tier 3) est positionné en dehors des frontières du mot aligné par Wav2Vec2.0 (tier 2), l'annotation n'est donc pas effectuée. Trois solutions sont envisagées : (1) remplacer Wav2Vec2.0 par un autre aligneur, ce que nous avons fait à partir de la deuxième version de PLSPP avec Montreal Forced Aligner (MFA, McAuliffe et al., 2017), et le résultat est illustré figure 10.1 ; (2) introduire une marge de tolérance avant et après le mot pour inclure des pics d'intensités qui se situeraient en dehors des frontières du mot, bien que cela risque de provoquer le problème inverse, à savoir trop de pics d'intensité affectés au mot ; (3) combiner l'alignement mot-signal de Wav2Vec2.0 et l'alignement phonème-signal de MFA, en contraignant ce dernier à s'ajuster à celui du premier avec une marge de tolérance si nécessaire. Cette troisième solution nécessiterait cependant d'exécuter MFA sur chaque mot individuellement, ce qui pourrait significativement allonger les temps de traitement.

Nous avons tout de même choisi d'utiliser la première version de PLSPP pour l'analyse de la parole spontanée, car la précision obtenue grâce à Wav2Vec2.0 reste meilleure que celle de MFA en présence de disfluences. Cette limitation de MFA pourrait s'expliquer par le fait que les hésitations ne sont pas transcrites par le système de reconnaissance de parole (Whisper, Radford et al., 2022), et provoquent un décalage global dans l'alignement de MFA. Une combinaison des alignements de Wav2Vec2.0 et de MFA pourrait probablement résoudre ce problème.

Une autre limitation a été identifiée à propos de la mesure de durée des syllabes, concernant cette fois toutes les versions de PLSPP. Le système actuel ne prend pas en compte le type de voyelle dans la comparaison des durées syllabiques. Or, en anglais, certaines voyelles comme /iː/ ou /uː/ ont une durée intrinsèquement plus longue que d'autres, comme /ɪ/ ou /ʊ/. Une pondération des durées en fonction des types de voyelles permettrait de neutraliser ces différences.

Mesures de f_0 Les résultats mitigés obtenus sur la parole spontanée des locuteurs anglophones natifs ont révélé que la mesure de f_0 est influencée par la tendance au dévoisement de certaines voyelles, particulièrement fréquente chez les jeunes locu-

teurs de l'anglais. Ce dévoisement résulte en une absence de détection de fréquence fondamentale au niveau du noyau syllabique. Cette absence est alors comblée par une interpolation linéaire de la f_0 à partir des mesures les plus proches, mais peut parfois aboutir à des résultats incohérents.

Ce problème peut être également constaté dans PLSPP v2, mais dans une moindre mesure. Dans cette version, la moyenne des mesures de f_0 est calculée sur l'intervalle vocalique, toujours avec interpolation lorsqu'il y a dévoisement. Sur la figure 10.1, on peut voir que la proéminence est détectée sur la syllabe finale par PLSPP v2 (tier 10). Pourtant, la proéminence est détectée en initiale en termes d'intensité et de durée (tier 11). On peut constater que la f_0 est effectivement légèrement plus haute sur la deuxième syllabe (courbe bleue sur le spectrogramme), mais cette différence semble faible pour pouvoir contrebalancer la proéminence détectée en initiale sur les autres dimensions. La cause du problème semble être le fait que la f_0 moyenne est sous-estimée à cause du premier point de mesure de f_0, probablement issu d'une erreur de détection. Une solution possible ici serait de ne pas prendre en compte la mesure de f_0 à partir d'un certain taux de dévoisement de la voyelle.

Normalisation par locuteur La méthode actuelle de normalisation des mesures acoustiques consiste à transformer les valeurs absolues de f_0, d'intensité et de durée en centiles spécifiques à chaque locuteur et à chaque dimension. Cette méthode s'est montrée adaptée à nos analyses de corpus, mais elle nécessite une certaine quantité de parole pour chaque locuteur pour pouvoir faire un calcul précis des centiles. Une alternative pourrait consister à normaliser les mesures acoustiques à l'aide de z-scores (soustraction de chaque mesure par la moyenne des mesures de la dimension et du locuteur en question, divisée par son écart-type).

Nombre de syllabes Les versions actuelles de PLSPP mesurent la proéminence des syllabes sur la base d'un nombre prédéfini de syllabes par mot. Autrement dit, si le dictionnaire de référence attribue trois syllabes à un mot, l'annotation est réalisée sur trois noyaux syllabiques. La première version de PLSPP repose sur une détection acoustique des noyaux syllabiques, et l'annotation du mot est bloquée lorsque le nombre de noyaux détectés ne correspond pas au nombre de syllabes attendu. Dans les versions ultérieures, le nombre de noyaux syllabiques est déterminé par le dictionnaire phonologique utilisé par MFA, et l'annotation est systématiquement réalisée sur le nombre de syllabes attendu, indépendamment des noyaux détectés. Il serait toutefois pertinent d'envisager une annotation même lorsque le nombre de noyaux détectés diverge du nombre attendu. En effet, l'ajout ou la suppression d'une syllabe peut modifier significativement le rythme de la parole. Ce phénomène est particulièrement observable chez les locuteurs japonophones, qui ont tendance à insérer des voyelles dans certains groupes consonantiques (Kenworthy, 1987 ; Labrune, 2006).

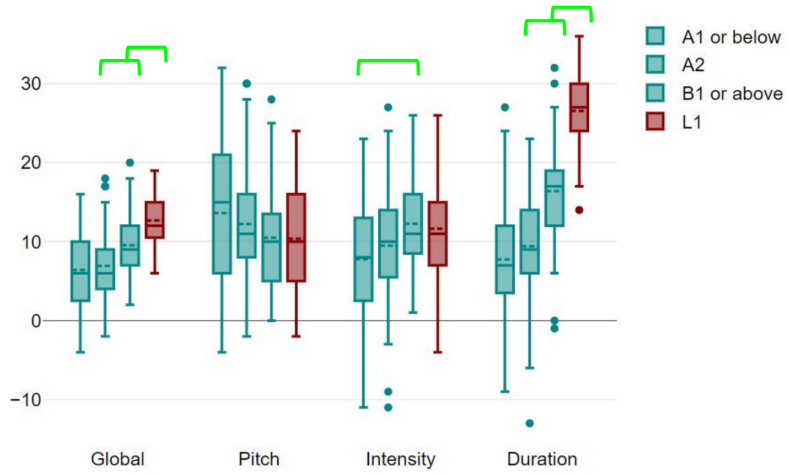


FIG. 10.2 : Degré de contraste prosodique entre les mots grammaticaux et lexicaux à travers un échantillon aléatoire de 55 textes par groupe de niveau (A1, A2, B1 et plus, natifs, dans l'ordre d'apparition). Les crochets verts indiquent les différences significatives (ANOVA, $p < 0{,}001$)

Phénomène de réduction vocalique Notre travail s'est principalement concentré sur la détection et la caractérisation de la syllabe proéminente par rapport aux autres syllabes du mot. Toutefois, il semble important d'accorder une attention particulière au phénomène de réduction vocalique, et de mesurer un degré de réduction vis-à-vis des syllabes non réduites, indépendamment du type d'accentuation. La non-réduction constitue un problème distinct de celui de l'accentuation et mérite une analyse spécifique. Par exemple, un locuteur peut accentuer la syllabe attendue, produire un contraste prosodique marqué, mais toutefois ne pas réduire les voyelles qui sont censées l'être. Cette tendance a un impact significatif sur le rythme de la parole, en particulier lorsque les mots grammaticaux ne sont pas réduits par rapport aux mots lexicaux (Tortel, 2021).

Afin de caractériser le contraste prosodique entre mots grammaticaux et lexicaux, nous avons développé une version adaptée de PLSPP (v3) permettant d'annoter tous les mots du corpus, quel que soit leur nombre de syllabes. Nous avons ensuite comparé la valeur prosodique moyenne des mots grammaticaux à celle de la syllabe accentuée des mots lexicaux dans un corpus de 34 h de parole lue par 42 locuteurs japonais (niveaux A1 à B2) et 7 locuteurs anglophones natifs (Nakanishi & Coulange, 2024). Les résultats montrent une forte influence de la durée syllabique sur la réalisation de ce contraste entre les groupes de locuteurs (*cf.* figure 10.2). Ces analyses mettent en évidence une tendance chez les locuteurs natifs à produire un contraste de durée beaucoup plus marqué entre mots grammaticaux et lexicaux que chez les locuteurs japonais. Ce contraste augmente par ailleurs avec le niveau de compétence en langue.

3.3 Évaluation dynamique de la compréhensibilité

Le protocole d'évaluation dynamique de la compréhensibilité a permis de confirmer nos hypothèses de recherche, notamment l'augmentation de la perception de l'effort de compréhension à la suite de pauses de bas niveau syntaxique et de patterns accentuels inappropriés. Cependant, la significativité des résultats obtenus reste limitée. Cela s'explique en partie par la multiplicité des facteurs susceptibles d'altérer la compréhensibilité : les pauses et les patterns accentuels ne représentent que deux paramètres parmi d'autres, et ces différents paramètres influencent simultanément la perception de l'auditeur. Nous pensons toutefois que des résultats plus marqués auraient pu être obtenus si l'échantillon de segments de parole analysés avait inclus une proportion plus importante de pauses intra-syntagmes et de mots présentant un contraste prosodique positif et élevé. En effet, les pauses intra-syntagmes ne représentaient que 14 % des pauses analysées (53 sur 382), tandis que les patterns accentuels avec un contraste élevé ($C'' \geq 0{,}2$) concernaient seulement 17 % des mots annotés (23 sur 139).

Pour pallier cette limitation, il aurait été possible d'adapter la définition des catégories de pauses et d'accentuation. Par exemple, à l'instar du calcul du score de distribution syntaxique des pauses DSP_n, les catégories de pauses pourraient être définies en fonction de l'importance des frontières syntaxiques plutôt que de leur type, avec des seuils ajustés afin d'obtenir un nombre équivalent de pauses dans chaque catégorie. De même, pour l'accentuation, il serait envisageable de moduler les seuils de contraste C'' afin d'obtenir une répartition plus équilibrée entre les catégories. Cette approche comporte toutefois le risque de diminuer le contraste entre les catégories, ce qui pourrait atténuer les différences observées dans l'effort moyen pour chacune d'elles.

Une autre limitation du protocole réside dans la durée des segments de parole évalués, qui étaient relativement courts (entre 26 et 66 secondes). Selon les retours recueillis, ce manque de contexte a rendu la tâche d'évaluation plus complexe. Dans l'étude de Nagle et al. (2019), les segments de parole duraient environ trois minutes chacun. Néanmoins, une telle durée peut soulever des questions quant à la capacité des évaluateurs à maintenir leur attention sur une période aussi longue. Une durée intermédiaire, d'environ une minute, semble constituer un compromis idéal : suffisamment longue pour permettre à l'auditeur de saisir le sujet de la conversation, tout en étant suffisamment courte pour éviter un déclin de l'attention ou une familiarisation excessive avec la prononciation du locuteur.

Conclusion générale

La production et la compréhension de l'oral sont deux compétences étroitement liées. Dans une situation de communication, produire de la parole vise nécessairement à être compris. Évaluer la compétence de production orale à travers la capacité du locuteur à se faire comprendre prend alors tout son sens. Dès lors, il semble plus pertinent de considérer cette compétence comme un tout, sans isoler l'aspect spécifique de la prononciation, qui n'est finalement qu'une forme donnée à la parole et ne peut être évaluée de manière intrinsèque.

Si la compréhensibilité du locuteur dépend pour une grande part de la capacité de compréhension de l'auditeur, elle peut être facilitée par certains aspects spécifiques de la production orale. Par exemple, une segmentation stratégique du flux de parole peut aider l'auditeur à structurer et à traiter l'information reçue, mais également lui permettre de réagir, de signaler au locuteur qu'il écoute ou qu'il comprend. La communication est ainsi une co-construction permanente entre le locuteur et l'auditeur, et évaluer les compétences de l'apprenant devrait porter sur sa capacité à co-construire cette communication, tant sur les aspects de compréhension (comprendre) que de compréhensibilité (se faire comprendre).

Mais alors, est-ce qu'évaluer la production orale de manière automatique est une entreprise viable ? Comment évaluer la production d'un apprenant autrement que par le biais de la compréhension de son interlocuteur ?

Il y a trois ans, au début de cette thèse, l'idée de proposer un environnement de conversation où l'apprenant pourrait dialoguer avec un avatar numérique capable d'interagir avec lui de manière similaire à un humain semblait encore appartenir à un futur lointain. Si ce n'est pas encore tout à fait une réalité, l'émergence récente des grands modèles de langage laisse entrevoir des perspectives réalistes d'environnements numériques pour la pratique et l'évaluation de la production orale spontanée en L2. Des entreprises comme LangX[7] proposent déjà des systèmes conversationnels spécialisés dans l'évaluation des compétences en langues. Toutefois, ces technologies ne

[7] https://speaking.langx.ai/

permettent que de proposer un environnement de production de parole : les questions de quoi et comment évaluer restent entières.

L'évaluation peut porter sur des aspects propres à la production du locuteur, mais également sur la dynamique des échanges, sur la capacité du locuteur à se faire comprendre et à s'adapter en fonction des réactions de son interlocuteur. Elle peut se faire dans différentes situations de communication, plus ou moins formelles, avec des interlocuteurs plus ou moins familiers avec la parole L2, ou encore avec le sujet abordé.

Concernant plus spécifiquement les aspects de production, il apparaît que la segmentation du flux de parole (à travers la distribution des pauses) et le rythme général (via les phénomènes d'accentuation) ont un impact direct sur la compréhension. Toutefois, ces deux paramètres doivent être appréhendés comme des variables continues et relatives. Il s'est avéré en effet plus pertinent de considérer la position des pauses en fonction du degré d'importance de la frontière syntaxique où elles se trouvent, plutôt que vis-à-vis du type de constituants grammaticaux qui s'y terminent ou commencent. De même, la mesure du contraste accentuel entre les syllabes s'est révélée plus informative que la simple identification des syllabes proéminentes. En outre, une pause de bas niveau syntaxique ou une accentuation non prescriptive ne constituent pas en elles-mêmes des obstacles à la compréhension ; c'est leur récurrence ou leur accumulation, combinée à d'autres facteurs, qui peut accroître l'effort de compréhension requis par l'auditeur. Ce phénomène illustre bien le « cocktail de l'intelligibilité » décrit par Zielinski (2006). La distribution des pauses et l'accentuation font partie d'un ensemble de paramètres qu'il convient de prendre en compte dans l'évaluation, mais celle-ci doit se faire à un niveau global, en combinant les aspects lexicaux, grammaticaux ou encore discursifs, mais également en tenant compte du contexte de la situation de communication et des réactions de l'auditeur.

Perspectives pour SELF

Quelles sont les perspectives pour le dispositif d'évaluation SELF ? Deux objectifs principaux se dessinent. Le premier est d'intégrer la production orale dans le test de positionnement actuellement déployé. Cela passera par l'ajout de courtes tâches d'élicitation de parole, à partir desquelles plusieurs mesures seront effectuées. Ces mesures combineront des aspects de fluence (par exemple, le score de distribution syntaxique des pauses DSP_n proposé dans cette thèse), de rythme (notamment via la mesure du contraste prosodique moyen $\overline{C}$), ainsi que des aspects lexicaux et syntaxiques (précision, diversité) ou pragmatiques (adéquation au contexte). Ces indicateurs contribueront à estimer un niveau de compétence qui viendra enrichir les scores actuellement calculés.

Cependant, si l'objectif immédiat est d'estimer le niveau des apprenants afin de les orienter vers des groupes adaptés, l'ambition à plus long terme est de développer un module d'évaluation formative de la production orale. Ce module devra permettre d'établir un diagnostic sur la base de ces différentes mesures, dans une situation de communication plus réaliste et sur une durée d'évaluation plus longue. L'objectif principal sera d'offrir à l'apprenant une estimation de son niveau de compréhensibilité dans un contexte donné et d'identifier les paramètres prioritaires à travailler pour améliorer ses performances en communication.

Si ces avancées technologiques permettent d'imaginer une évaluation entièrement automatisée, il convient de souligner que celle-ci ne pourrait porter que sur des aspects superficiels de la communication. Les mesures évoquées ici constituent un ensemble d'indices susceptibles de favoriser la réussite de la communication ou de refléter une communication réussie, permettant ainsi d'estimer la compétence de production orale du locuteur. En revanche, une évaluation approfondie ne peut être envisagée que grâce à l'expertise d'un évaluateur humain, capable, contrairement à la machine, d'interpréter la production en profondeur – tant sur le plan du sens que par une combinaison plus nuancée des différents paramètres.

Une évaluation automatisée, limitée à des paramètres de surface, peut néanmoins représenter un complément précieux, en allégeant la charge de l'évaluateur humain et en lui permettant de se concentrer sur l'interprétation globale et l'établissement du jugement final. Cela explique pourquoi une automatisation totale demeure inadéquate dans le cadre d'une évaluation certificative, où les enjeux sont souvent importants. En revanche, dans un contexte de formation, le potentiel de l'évaluation automatique est considérable. Il nous revient alors de nous approprier ces technologies et de les adapter à nos besoins afin d'en tirer pleinement parti.

Nous conclurons par les mots d'un chercheur qui a profondément influencé notre réflexion tout au long de ce parcours doctoral : "*Rather than a false fight over an already too small piece of the language teaching pie,* [*technology*] *is a way to expand the pie so that more teachers and learners can enjoy their own use of spoken language*" (Levis, 2007, p. 197).

Références

Abercrombie, D. (1967). *Elements of General Phonetics*. Edinburgh Univ. Press.

Alazard, C. (2013). *Rôle de la prosodie dans la fluence en lecture oralisée chez des apprenants de Français Langue Étrangère* [thèse de doct., Univ. Toulouse 2].

Amengual-Pizarro, M., & García-Laborda, J. (2017). Analysing test-takers' views on a computer-based speaking test. *Profile : Issues in Teachers' Professional Development*, *19*, 23-38. https://doi.org/10.15446/profile.v19n_sup1.68447

Arias, J. P., Yoma, N. B., & Vivanco, H. (2010). Automatic intonation assessment for computer aided language learning. *Speech Commun.*, *52*(3), 254-267.

Astesano, C. (2001). *Rythme et accentuation en français : invariance et variabilité stylistique*. L'Harmattan.

Baayen, H., Piepenbrock, R., & Gulikers, L. (1995). The CELEX Lexical Database (CD-ROM).

Bada, I., Fohr, D., & Illina, I. (2020). Reconnaissance automatique de la parole : génération des prononciations non natives pour l'enrichissement du lexique. In C. Benzitoun, C. Braud, L. Huber, D. Langlois, S. Ouni, S. Pogodalla & S. Schneider (Éd.), *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 1 : Journées d'Études sur la Parole* (p. 27-35). ATALA. https://hal.archives-ouvertes.fr/hal-02798511

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0 : A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, *33*, 12449-12460.

Bain, M., Huh, J., Han, T., & Zisserman, A. (2023). WhisperX : Time-Accurate Speech Transcription of Long-Form Audio. *Interspeech 2023*.

Baker, A. A. (2011). ESL teachers and pronunciation pedagogy : Exploring the development of teachers' cognitions and classroom practices. *Research Online*, 82. https://ro.uow.edu.au/edupapers/368

Bard, E., & Lickley, R. (1997). On not remembering disfluencies. *Proc. of Eurospeech 97*, 2855-2858.

Bauer, D. F. (1972). Constructing Confidence Sets Using Rank Statistics. *Journal of the American Statistical Association*, *67*(339), 687-690. https://doi.org/10.1080/01621459.1972.10481279

Bernstein, J. (2003). Objective measurement of intelligibility. *ICPhS*, 1581-1584.

Bernstein, J., Cohen, M., Murveit, H., Rtischev, D., & Weintraub, M. (1990). Automatic evaluation and training in English pronunciation. *The First International Conference on Spoken Language Processing, ICSLP 1990*.

Bertinetto, P. (1989). Reflections on the dichotomy "stress" vs. "syllable-timing". *Revue de Phonétique Appliquée*, *91*, 99-130.

Bhat, S., & Yoon, S.-Y. (2015). Automatic assessment of syntactic complexity for spontaneous speech scoring. *Speech Communication*, *67*, 42-57. https://doi.org/10.1016/j.specom.2014.09.005

Bredin, H. (2023). pyannote.audio 2.1 speaker diarization pipeline : principle, benchmark, and recipe. *Interspeech 2023*, 1983-1987. https://doi.org/10.21437/Interspeech.2023-105

Bredin, H., Yin, R., Coria, J. M., Gelly, G., Korshunov, P., Lavechin, M., Fustes, D., Titeux, H., Bouaziz, W., & Gill, M.-P. (2020). pyannote.audio : neural building blocks for speaker diarization. *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*.

Burgess, J., & Spencer, S. (2000). Phonology and Pronunciation in Integrated Language Teaching and Teacher Education. *System*, *28*, 191-215.

Butcher, A. (1981). *Aspects of the Speech Pause : Phonetic Correlates and Communicative Functions*. Inst. f. Phonetik.

Calbris, G., & Montredon, J. (1975). *Approche rythmique, intonative et expressive du Français langue étrangère : sketches-exercices-illustrations-photos-cartes d'expression : les exercices ont éné expérimentés au Centre de linguistique appliquée de Besançon*. CLES International.

Campione, E., & Véronis, J. (2002). A large-scale multilingual study of silent pause duration. *Speech Prosody 2002*, 199-202.

Candea, M. (2000). *Contribution à l'étude des pauses silencieuses et des phénomènes dits « d'hésitation » en français oral spontané. Etude sur un corpus de récits en classe de français.* [thèse de doct., Univ. Sorbonne Nouvelle - Paris III]. https://theses.hal.science/tel-00290143

Cao, Y., & Chen, H. (2019). World Englishes and Prosody : Evidence from the Successful Public Speakers. *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2048-2052.

Cardwell, R., Naismith, B., LaFlair, G. T., & Nydick, S. (2024). *Duolingo English Test : Technical Manual* (Duolingo Research Report). Duolingo, Inc. 5900 Penn Ave, Pittsburgh, PA 15206, USA. https://go.duolingo.com/dettechnicalmanual

Catford, J. C. (1987). Phonetics and the teaching of pronunciation : A systemic description of English phonology. In J. Morley (Éd.), *Current Perspectives on pronunciation : Practices anchored in theory* (p. 87-100). Tesol Press.

Chen, J.-Y., & Wang, L. (2010). Automatic lexical stress detection for Chinese learners' of English. *2010 7th International Symposium on Chinese Spoken Language Processing*, 407-411.

Chen, L., & Zechner, K. (2011). Applying rhythm features to automatically assess non-native speech. *Interspeech 2011*.

Chen, L.-Y., & Jang, J.-S. (2012). Stress Detection of English Words for a CAPT System Using Word-Length Dependent GMM-Based Bayesian Classifiers. *Interdisciplinary Information Sciences*, *18*, 65-70.

Cliff, N. (1993). Dominance statistics : Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, *114*(3), 494-509.

Collard, P. (2009). *Disfluency and listeners' attention : An investigation of the immediate and lasting e ects of hesitations in speech* [thèse de doct., Univ. of Edinburgh]. http://hdl.handle.net/1842/3234

Conseil de l'Europe. (2001). *Un cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer : apprentissage des langues et citoyenneté européenne / Conseil de l'Europe, Division des politiques linguistiques*.

Conseil de l'Europe. (2018). *Cadre Européen Commun de Référence pour les Langues – Volume complémentaire*.

Cooper, N., Cutler, A., & Wales, R. (2002). Constraints of lexical stress on lexical access in English : evidence from native and non-native listeners. *Language and Speech*, *45*(Pt 3), 207-228. https://doi.org/10.1177/00238309020450030101

Corley, M., MacGregor, L. J., & Donaldson, D. I. (2007). It's the way that you, er, say it : Hesitations in speech affect language comprehension. *Cognition*, *105*(3), 658-668. https://doi.org/10.1016/j.cognition.2006.10.010

Coulange, S. (2023). Computer-aided pronunciation training in 2022 : When pedagogy struggles to catch up. In A. Henderson & A. Kirkova-Naskova (Éd.), *Proc. of the 7th International Conference on English Pronunciation : Issues and Practices* (p. 11-22). Univ. Grenoble-Alpes. https://doi.org/10.5281/zenodo.8137754

Coulange, S., & Kato, T. (2023). Pause position analysis in spontaneous speech for L2 English fluency assessment. *2023 Autumn Meeting of the Acoustic Society of Japan*. https://hal.science/hal-04253964

Coulange, S., Kato, T., Rossato, S., & Masperi, M. (2023). フランス人学習者による自発英語発話における語彙アクセント自動測定 [Automatic Measurement of Lexical Stress in Spontaneous L2 English Speech of French Learners]. *Proc. of the 37th General Meeting of the Phonetic Society of Japan*. https://hal.science/hal-04253927

Coulange, S., Kato, T., Rossato, S., & Masperi, M. (2024a). Enhancing Language Learners' Comprehensibility through Automated Analysis of Pause Positions and Syllable Prominence. *Languages*, *9*(3). https://doi.org/10.3390/languages9030078

Coulange, S., Kato, T., Rossato, S., & Masperi, M. (2024b). Dynamic Approach to Comprehensibility Assessment in Foreign Language Pronunciation Training. *8th International Conference on English Pronunciation : Issues and Practices*. https://hal.science/hal-04666118

Coulange, S., Kato, T., Rossato, S., & Masperi, M. (2024c). Exploring Impact of Pausing and Lexical Stress Patterns on L2 English Comprehensibility in Real Time. *Interspeech 2024*, 1030-1034. https://doi.org/10.21437/Interspeech.2024-1627

Coutinho, E., Hönig, F., Zhang, Y., Hantke, S., Batliner, A., Nöth, E., & Schuller, B. (2016). Assessing the Prosody of Non-Native Speakers of English : Measures and Feature Sets. *Proc. of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 1328-1332. https://aclanthology.org/L16-1211

Crowther, D., Trofimovich, P., Saito, K., & Isaacs, T. (2017). Linguistic dimensions of L2 accentedness and comprehensibility vary across speaking tasks. *Studies in Second Language Acquisition*, *40*(2), 443-457. https://doi.org/10.1017/S027226311700016X

Cruttenden, A. (1997). *Intonation* (2e éd.). Cambridge Univ. Press. https://doi.org/10.1017/CBO9781139166973

Crystal, D. (2008). *A dictionary of linguistics and phonetics* (6e éd.). Wiley-Blackwell.

Cucchiarini, C., Strik, H., & Boves, L. (2000). Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. *The Journal of the Acoustical Society of America*, *2*(107), 989-99.

Cucchiarini, C., Strik, H., & Boves, L. (2002). Quantitative assessment of second language learners' fluency : Comparisons between read and spontaneous speech. *The Journal of the Acoustical Society of America*, *111*, 2862-2873. https://doi.org/10.1121/1.1471894

Cutler, A. (2015). Lexical Stress in English Pronunciation. In *The Handbook of English Pronunciation* (p. 106-124). John Wiley & Sons, Inc.

Cutler, A., & Carter, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language*, *2*(3), 133-142. https://doi.org/10.1016/0885-2308(87)90004-0

Cutler, A., & Jesse, A. (2021). Word Stress in Speech Perception. In *The Handbook of Speech Perception* (p. 239-265). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119184096.ch9

Dauer, R. (1987). Phonetic and Phonological Components of Language Rhythm. *Proc. of the 11th International Congress of Phonetic Sciences*, *5*, 445-450.

Davis, L., & Papageorgiou, S. (2021). Complementary strengths ? Evaluation of a hybrid human-machine scoring approach for a test of oral academic English. *Assessment in Education : Principles, Policy & Practice*, *28*(4), 437-455. https://doi.org/10.1080/0969594X.2021.1979466

de Jong, N. (2016). Predicting pauses in L1 and L2 speech : the effects of utterance boundaries and word frequency. *International Review of Applied Linguistics in Language Teaching*, *54*(2), 113-132. https://doi.org/10.1515/iral-2016-9993

de Jong, N., & Bosker, H. R. (2013). Choosing a threshold for silent pauses to measure second language fluency. In R. Eklund (Éd.), *Proc. of the 6th Workshop on Disfluency in Spontaneous Speech, DiSS* (p. 17-20).

Delattre, P. (1963). Comparing the prosodic features of English, German, Spanish and french. *International Review of Applied Linguistics in Language Teaching*, *1*(1).

Delattre, P. (1966). *Studies in French and Comparative Phonetics*. De Gruyter Mouton. https://doi.org/10.1515/9783112416105

Demol, M., Verhelst, W., & Verhoeve, P. (2007). The duration of speech pauses in a multilingual environment. *Interspeech 2007*, 990-993. https://doi.org/10.21437/Interspeech.2007-350

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility : Evidence from Four L1s. *Studies in Second Language Acquisition*, *19*(1), 1-16. https://doi.org/10.1017/S0272263197001010

Derwing, T. M., & Munro, M. J. (2009). Putting accent in its place : Rethinking obstacles to communication. *Language Teaching*, *42*(4), 476-490.

Derwing, T. M., & Munro, M. J. (2015). *Pronunciation Fundamentals : Evidence-based perspectives for L2 teaching and research*. John Benjamins. https://www.jbe-platform.com/content/books/9789027268594

Deshmukh, O., & Verma, A. (2009). Nucleus-level clustering for word-independent syllable stress classification. *Speech Communication*, *51*, 1224-1233.

Di Cristo, A., & Hirst, D. (1997). L'accentuation non emphatique en français : stratégies et paramètres. In *Polyphonie pour Ivan Fónagy* (p. 71-101). L'Harmattan.

Di Cristo, A. (1998). Intonation in French. In D. Hirst & A. Di Cristo (Éd.), *Intonation Systems : A Survey of Twenty Languages*. Cambridge Univ. Press.

Di Cristo, A. (2013). *La prosodie de la parole / Albert Di Cristo*. De Boeck-Solal.

Di Cristo, A., & Hirst, D. (1993). Rythme syllabique, rythme mélodique et représentatin hiérarchique de la prosodie du français. *Travaux de l'Institut de phonétique d'Aix*.

Didelot, M., Racine, I., Zay, F., & Prikhodkine, A. (2019). Enseignement et évaluation de la prononciation aujourd'hui : l'intelligibilité comme enjeu. *Recherches en didactique des langues et des cultures*, *16*(1). https://doi.org/10.4000/rdlc.4333

Ding, H., Lin, B., Wang, L., Wang, H., & Fang, R. (2020). A Comparison of English Rhythm Produced by Native American Speakers and Mandarin ESL Primary School Learners. *Interspeech 2020*, 4481-4485. https://doi.org/10.21437/Interspeech.2020-2207

Dodane, C., & Hirsch, F. (2018). L'organisation spatiale et temporelle de la pause en parole et en discours. *Langages*, *N°211*(3), 5-12.

Duan, R., Kawahara, T., Dantsuji, M., & Zhang, J. (2017). Articulatory modeling for pronunciation error detection without non-native training data based on DNN transfer learning. *IEICE Transactions on Information and Systems*, *E100.D*(9), 2174-2182.

Duez, D. (1982). Silent and Non-Silent Pauses in Three Speech Styles. *Language and Speech*, *25*(1), 11-28. https://doi.org/10.1177/002383098202500102

Duez, D. (1985). Perception of Silent Pauses in Continuous Speech. *Language and Speech*, *28*(4), 377-389. https://doi.org/10.1177/002383098502800403

Duez, D. (1991). *La pause dans la parole de l'homme politique*. Éd. du Centre national de la recherche scientifique.

Duez, D. (1993). Acoustic correlates of subjective pauses. *Journal of psycholinguistic research*, *22*(1), 21-40.

Duez, D. (1995). Perception of hesitations in spontaneous french speech. *ICPhS*, 498-501.

Dupoux, E., Pallier, C., Sebastian, N., & Mehler, J. (1997). A Destressing "Deafness" in French ? *Journal of Memory and Language*, *36*(3), 406-421. https://doi.org/10.1006/jmla.1996.2500

Ellis, N. C., & Bogart, P. S. H. (2007). Speech and language technology in education : the perspective from SLA research and practice. *SLaTE*.

Erickson, D., Raso, T., Lundmark, M. S., Frid, J., & Coulange, S. (2025). The Many Colors of Prominence : A Pilot Study of Topic Prosodic Forms. *Journal of Speech Sciences*.

Evain, S. (2024). *Dimensions de variation de la parole spontanée pour l'étude inter-corpus des performances de systèmes de reconnaissance automatique de la parole* [thèse de doct., Univ. Grenoble Alpes].

Evanini, K., & Wang, X. (2013). Automated speech scoring for non-native middle school students with multiple task types. *Interspeech 2013*.

Evanini, K., & Zechner, K. (2019). Overview of automated speech scoring. In K. Zechner & K. Evanini (Éd.), *Automated Speaking Assessment* (p. 3-20). Routledge. https://doi.org/10.4324/9781315165103-1

Fauth, C., & Trouvain, J. (2018). Détails phonétiques dans la réalisation des pauses en français : étude de parole lue en langue maternelle vs en langue étrangère. *Langages*, *N° 211*(3), 81-95.

Ferrer, L., Bratt, H., Richey, C., Franco, H., Abrash, V., & Precoda, K. (2015). Classification of lexical stress using spectral and prosodic features for computer-assisted language learning systems. *Speech Communication*, *69*, 31-45. https://doi.org/10.1016/j.specom.2015.02.002

Field, J. (2005). Intelligibility and the Listener : The Role of Lexical Stress. *TESOL Quarterly*, *39*(3), 399-423. https://doi.org/10.2307/3588487

Fletcher, J. (1987). Some micro and macro effects of tempo change on timing in French. *Linguistics*, *25*(5), 951-968. https://doi.org/10.1515/ling.1987.25.5.951

Fónagy, I. (1980). L'accent français : accent probabilitaire (dynamique d'un changement prosodique). *Studia Phonetica Montréal*, *15*, 123-233.

Fontan, L., Le Coz, M., & Detey, S. (2018). Automatically Measuring L2 Speech Fluency without the Need of ASR : A Proof-of-concept Study with Japanese Learners of French. *Interspeech 2018*, 2544-2548.

Fox, B. A., Hayashi, M., & Jasperson, R. (1996). Resources and repair : a cross-linguistic study of syntax and repair. In E. Ochs, E. A. Schegloff & S. A. Thompson (Éd.), *Interaction and Grammar* (p. 185-237). Cambridge Univ. Press.

Fox Tree, J. (2001). Listeners' uses ofum anduh in speech comprehension. *Memory & Cognition*, *29*(2), 320-326.

Franco, H., Neumeyer, L., Kim, Y., & Ronen, O. (1997). Automatic pronunciation scoring for language instruction. *Acoustics, Speech, and Signal Processing*, *2*, 1471-1474.

Frost, D. (2023). Prosody in English pronunciation : embodiment and metacognition [Rapport de synthèse en vue d'obtenir l'habilitation à diriger des recherches]. https://doi.org/10.13140/RG.2.2.20000.15369

Frost, D., & O'Donnell, J. (2018). Evaluating the essentials : the place of prosody in oral production. In J. Volín & R. Skarnitzl (Éd.), *The Pronunciation of English by Speakers of Other Languages* (p. 228-259). Cambridge Scholars Publishing. https://hal.science/hal-02085252

Frost, D., Skarnitzl, R., Coulange, S., & Hosseini, H. (2024). Perceived ease of understanding in French-accented academic discourse : and the chief culprits

are...? *The 17th International Conference on Native and Non-native Accents of English*.

Fu, J., Chiba, Y., Nose, T., & Ito, A. (2020). Automatic assessment of English proficiency for Japanese learners without reference sentences based on deep neural network acoustic models. *Speech Communication*, *116*, 86-97. https://doi.org/10.1016/j.specom.2019.12.002

Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., Dahlgren, N., & Zue, V. (1993). *TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1*. Philadelphia : Linguistic Data Consortium.

Gass, S., & Varonis, E. M. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Language Learning*, *34*(1), 65-87. https://doi.org/10.1111/j.1467-1770.1984.tb00996.x

Gibbon, D., & Gut, U. (2001). Measuring speech rhythm. *Proc. 7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, 95-98. https://doi.org/10.21437/Eurospeech.2001-36

Gilquin, G., Bestgen, Y., & Granger, S. (2022). Assessing EFL speech : A teacher-focused perspective. *Journal of Second Language Teaching & Research*, *9*(1), 33-57.

Goldman, J.-P., François, T., Roekhaut, S., & Simon, A.-C. (2010). Étude statistique de la durée pausale dans différents styles de parole. *Actes des 28èmes journées d'étude sur la parole (JEP)*, 161-164. http://hdl.handle.net/2078.1/81909

Goldman-Eisler, F. (1968). *Psycholinguistics : Experiments in Spontaneous Speech*. Academic Press Inc.

Goronzy, S., Rapp, S., & Kompe, R. (2004). Generating non-native pronunciation variants for lexicon adaptation [Adaptation Methods for Speech Recognition]. *Speech Communication*, *42*(1), 109-123. https://doi.org/10.1016/j.specom.2003.09.003

Grosjean, F., & Deschamps, A. (1972). *Phonetica*, *26*(3), 129-156. https://doi.org/10.1159/000259407

Grosjean, F. (1980). Comparative studies of temporal variables in spoken and sign languages : A short review. In H. W. Dechert & M. Raupach (Éd.), *Studies in Honour of Frieda Goldman-Eisler* (p. 307-312). De Gruyter Mouton. https://doi.org/10.1515/9783110816570.307

Grosjean, F., & Deschamps, A. (1975). Analyse contrastive des variables temporelles de l'anglais et du français : vitesse de parole et variables composantes, phénomènes d'hésitation. *Phonetica*, *31*(3-4), 144-184. https://doi.org/10.1159/000259667

Grosman, I., Simon, A. C., & Degand, L. (2018). Variation de la durée des pauses silencieuses : impact de la syntaxe, du style de parole et des disfluences. *Langages*, *N° 211*(3), 13-40. https://doi.org/10.3917/lang.211.0013

Hahn, L. D. (2004). Primary Stress and Intelligibility : Research to Motivate the Teaching of Suprasegmentals. *TESOL Quarterly*, *38*(2), 201-223. https://doi.org/10.2307/3588378

Hayes-Harb, R., Smith, B. L., Bent, T., & Bradlow, A. R. (2008). The interlanguage speech intelligibility benefit for native speakers of Mandarin : Production and perception of English word-final voicing contrasts. *Journal of Phonetics*, *36*(4), 664-679. https://doi.org/10.1016/j.wocn.2008.04.002

Heldner, M., & Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, *38*(4), 555-568. https://doi.org/10.1016/j.wocn.2010.08.002

Henderson, A., Frost, D., Tergujeff, E., Kautzsch, A., Murphy, D., Kirkova-Naskova, A., Waniek-Klimczak, E., Levey, D., Cunnigham, U., & Curnick, L. (2012). The English pronunciation teaching in Europe survey : Selected results. *Research in Language*, *10*(1), 5-27. https://doi.org/10.2478/v10015-011-0047-4

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy : Industrial-strength Natural Language Processing in Python. https://doi.org/10.5281/zenodo.1212303

Huensch, A., & Nagle, C. (2021). The Effect of Speaker Proficiency on Intelligibility, Comprehensibility, and Accentedness in L2 Spanish : A Conceptual Replication and Extension of Munro and Derwing (1995a). *Language Learning*, *71*(3), 626-668. https://doi.org/10.1111/lang.12451

Inoue, Y., Kabashima, S., Saito, D., Minematsu, N., Kanamura, K., & Yamauchi, Y. (2018). A Study of Objective Measurement of Comprehensibility through Native Speakers' Shadowing of Learners' Utterances. *Interspeech 2018*, 1651-1655. https://doi.org/10.21437/Interspeech.2018-1860

Isaacs, T. (2018). Fully automated speaking assessment : Changes to proficiency testing and the role of pronunciation. In O. Kang, R. Thomson & J. Murphy (Éd.), *The Routledge handbook of contemporary English pronunciation* (p. 570-584). Routledge.

Isaacs, T., & Thomson, R. (2013). Rater Experience, Rating Scale Length, and Judgments of L2 Pronunciation : Revisiting Research Conventions. *Language Assessment Quarterly*, *10*(2), 135-159. https://doi.org/10.1080/15434303.2013.769545

Isaacs, T., & Thomson, R. (2020). Reactions to second language speech. *Journal of Second Language Pronunciation*, *6*(3), 402-429. https://doi.org/10.1075/jslp.20018.isa

Isaacs, T., & Trofimovich, P. (2011). Phonological memory, attention control, and musical ability : Effects of individual differences on rater judgments of second language speech. *Applied Psycholinguistics*, *32*(1), 113-140. https://doi.org/10.1017/S0142716410000317

Isaacs, T., & Trofimovich, P. (2012). Deconstructing Comprehensibility : Identifying the Linguistic Influences on Listeners' L2 Comprehensibility Ratings. *Studies in Second Language Acquisition*, *34*(3), 475-505. https://doi.org/10.2307/26328952

Isaacs, T., Trofimovich, P., & Foote, J. A. (2018). Developing a user-oriented second language comprehensibility scale for English-medium universities. *Language Testing*, *35*(2), 193-216. https://doi.org/10.1177/0265532217703433

Isbell, D. R., & Lee, J. (2022). Self-Assessment of Comprehensibility and Accentedness in Second Language Korean. *Language Learning*, *72*(3), 806-852. https://doi.org/10.1111/lang.12497

James, A. L. (1929). *Historical Introduction to French Phonetics*. Univ. of London Press, Ltd.

Johnson, D. O., & Kang, O. (2015). Automatic prominent syllable detection with machine learning classifiers. *Int. J. Speech Technol.*, *18*(4), 583-592. https://doi.org/10.1007/s10772-015-9299-z

Kahng, J. (2014). Exploring Utterance and Cognitive Fluency of L1 and L2 English Speakers : Temporal Measures and Stimulated Recall. *Language Learning*, *64*(4), 809-854. https://doi.org/10.1111/lang.12084

Kahng, J. (2018). The effect of pause location on perceived fluency. *Applied Psycholinguistics*, *39*(3), 569-591. https://doi.org/10.1017/S0142716417000534

Kallio, H., Kuronen, M., & Koivusalo, L. (2022). The role of pause location in perceived fluency and proficiency in L2 Finnish. *ISAPh 2022, 4th International Symposium on Applied Phonetics*, 22-27. https://doi.org/10.21437/ISAPh.2022-5

Kang, O., & Johnson, D. (2018). The roles of suprasegmental features in predicting English oral proficiency with an automated system. *Language Assessment Quaterly*, *15*(2), 150-168.

Kennedy, S., & Trofimovich, P. (2008). Intelligibility, Comprehensibility, and Accentedness of L2 Speech : The Role of Listener Experience and Semantic Context. *The Canadian Modern Language Review*, *64*(3), 459-489. https://doi.org/10.3138/cmlr.64.3.459

Kenworthy, J. (1987). *Teaching English Pronunciation*. Longman.

Kernou, H. (2022). Les disfluences dans le discours radiophonique : signification(s) et fonction (s) communicative (s). *Multilinguales*, (17).

Kim, A.-Y., & di Gennaro, K. (2012). Scoring Behavior of Native vs. Non-native Speaker Raters of Writing Exams. *Language Research*, *48*(2), 319-342.

Kimura, T., Coulange, S., & Kato, T. (2024). 日本人小学生による英語暗唱音声における語彙強勢位置の自動推定と母語話者評価 [Automatic estimation and native speakers' evaluation of lexical stress positions in English recitation speech produced by Japanese elementary school children]. 日本音響学会第 *151* 回研究発表会 [*2024 Spring Meeting of the Acoustical Society of Japan*], *2024 Spring Meeting of the Acoustical Society of Japan*, 673-676. https://hal.science/hal-04510493

Kirsner, K., Dunn, J., & Hird, K. (2005). Language Production : A complex dynamic system with a chronometric footprint. *7th International Conference on Cognitive Systems*.

Kirsner, K., Dunn, J., & Hird, K. (2003). Fluency : Time for a Paradigm Shift. *Disfluency in Spontaneous Speech (DiSS 2003)*, 13-16.

Kisler, T., Reichel, U., & Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language*, *45*, 326-347.

Kitaev, N., Cao, S., & Klein, D. (2019). Multilingual Constituency Parsing with Self-Attention and Pre-Training. *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics*, 3499-3505. https://doi.org/10.18653/v1/P19-1340

Korzekwa, D., Barra-Chicote, R., Zaporowski, S., Beringer, G., Lorenzo-Trueba, J., Serafinowicz, A., Droppo, J., Drugman, T., & Kostek, B. (2021). Detection of Lexical Stress Errors in Non-Native (L2) English with Data Augmentation and Attention. *Interspeech 2021*. https://arxiv.org/abs/2012.14788

Krivokapić, J. (2007). Prosodic planning : Effects of phrasal length and complexity on pause duration. *Journal of Phonetics*, *35*(2), 162-179.

Kubozono, H. (2006). Where does loanword prosody come from ? : A case study of Japanese loanword accent [Loanword Phonology : Current Issues]. *Lingua*, *116*(7), 1140-1170. https://doi.org/10.1016/j.lingua.2005.06.010

Labrune, L. (2006). *La phonologie du japonais* (C. Linguistique, Éd. ; Société de Linguistique de Paris). Peeters.

Lacheret-Dujour, A., & Victorri, B. (2002). La période intonative comme unité d'analyse pour l'étude du français parlé : modélisation prosodique et enjeux linguistiques. *Verbum : Analecta Neolatina*, *1-2*(24), 55-72. https://shs.hal.science/halshs-00009487

Lay, C. H., & Paivio, A. (1969). The effects of task difficulty and anxiety on hesitations in speech. *Canadian Journal of Behavioural Science*, *1*(1), 25-37.

Lee, A., & Glass, J. (2015). Mispronunciation Detection Without Non-Native Training Data. *Interspeech 2015*, 643-647.

Lennon, P. (1990). Investigating Fluency in EFL : A Quantitative Approach. *Language Learning*, *40*(3), 387-417. https://doi.org/10.1111/j.1467-1770.1990.tb00669.x

Levin, H., & Silverman, I. (1965). Hesitation Phenomena in Children's Speech. *Language and Speech*, *8*(2), 67-85. https://doi.org/10.1177/002383096500800201

Levin, H., Silverman, I., & Ford, B. L. (1967). Hesitations in children's speech during explanation and description. *Journal of Verbal Learning and Verbal Behavior*, *6*(4), 560-564.

Levis, J. (2007). Computer technology in teaching and researching pronunciation. *Annual Review of Applied Linguistics*, *27*, 184-202. https://doi.org/10.1017/S0267190508070098

Levis, J. (2018). *Intelligibility, Oral Communication, and the Teaching of Pronunciation*. Cambridge Univ. Press. https://doi.org/10.1017/9781108241564

Li, C., Liu, J., & Xia, S. (2007). English sentence stress detection system based on HMM framework. *Applied Mathematics and Computation*, *185*(2), 759-768.

Li, K., Mao, S., Li, X., Wu, Z., & Meng, H. (2018). Automatic lexical stress and pitch accent detection for L2 English speech using multi-distribution deep neural networks. *Speech Communication*, *96*, 28-36. https://doi.org/10.1016/j.specom.2017.11.003

Li, W., Siniscalchi, S. M., Chen, N. F., & Lee, C.-H. (2016). Improving non-native mispronunciation detection and enriching diagnostic feedback with DNN-based speech attribute modeling. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Lickley, R. (1995). Missing disfluencies. *ICPhS*, *4*, 192-195.

Lickley, R. (2015). Fluency and Disfluency. In M. A. Redford (Éd.), *The Handbook of Speech Production* (p. 445-469). Chichester : Wiley Online Library. https://doi.org/10.1002/9781118584156.ch20

Lin, Z., Inoue, Y., Trisitichoke, T., Ando, S., Saito, D., & Minematsu, N. (2019). Native Listeners' Shadowing of Non-native Utterances as Spoken Annotation Representing Comprehensibility of the Utterances. *8th ISCA Workshop on Speech and Language Technology in Education (SLaTE 2019)*, 43-47. https://doi.org/10.21437/SLaTE.2019-8

Lin, Z., Takashima, R., Saito, D., Minematsu, N., & Nakanishi, N. (2020). Shadowability Annotation with Fine Granularity on L2 Utterances and its Improvement with Native Listeners' Script-Shadowing. *Interspeech 2020*, 3865-3869. https://doi.org/10.21437/Interspeech.2020-2550

Loukina, A., & Yoon, S.-Y. (2019, novembre). Scoring and filtering models for automated speech scoring. In K. Zechner & K. Evanini (Éd.), *Automated speaking assessment : Using language technologies to score spontaneous speech* (p. 75-98). Routledge.

Loukina, A., Zechner, K., Chen, L., & Heilman, M. (2015). Feature selection for automated speech scoring. *Proc. of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*.

Lounsbury, F. (1954). Transitional probability, linguistic structure, and systems of habit-family hierarchies. In C. Osgood & T. Sebeok (Éd.), *Psycholinguistics : A survey of theory and research problems* (p. 93-101). Waverley Press.

Lundholm Fors, K. (2015). *Production and Perception of Pauses in Speech* [thèse de doct., Univ. of Gothenburg]. http://hdl.handle.net/2077/39346

MacGregor, L. (2008). *Disfluencies affect language comprehension : evidence from event-related potentials and recognition memory* [thèse de doct., Univ. of Edinburgh]. http://hdl.handle.net/1842/3311

MacIntyre, P. D. (2012). The Idiodynamic Method : A Closer Look at the Dynamics of Communication Traits. *Communication Research Reports*, *29*(4), 361-367. https://doi.org/10.1080/08824096.2012.723274

Maclay, H., & Osgood, C. E. (1959). Hesitation Phenomena in Spontaneous English Speech. *WORD*, *15*(1), 19-44. https://doi.org/10.1080/00437956.1959.11659682

Mareková, L., & Beňuš, Š. (2024). Task complexity and pausing behavior in L1 and L2 task-oriented dialogues. *Speech Prosody 2024*, 517-521. https://doi.org/10.21437/SpeechProsody.2024-105

Martin, J., & Strange, W. (1968). The perception of hesitation in spontaneous speech. *Perception & Psychophysics*, *3*(6), 427-438.

Martin, L., Degand, L., & Simon, A.-C. (2014). Forme et fonction de la périphérie gauche dans un corpus oral multigenres annoté. *Corpus*, (13), 243-265.

Martin, V., Beaumard, C., Rouas, J.-L., & Wu, Y. (2024). Is automatic phoneme recognition suitable for speech analysis ? Temporal and performance evaluation of an Automatic Speech Recognition model in spontaneous French. *Speech Prosody 2024*, 1120-1124. https://doi.org/10.21437/SpeechProsody.2024-226

Matzinger, T., Ritt, N., & Fitch, W. T. (2020). Non-native speaker pause patterns closely correspond to those of native speakers at different speech rates. *PLoS One*, *15*(4), e0230710.

Maynard, S. K. (1989). *Japanese conversation–self-contextualization through structure and interactional management*. Praeger.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner : Trainable Text-Speech Alignment Using Kaldi. *Interspeech 2017*, 498-502. https://doi.org/10.21437/Interspeech.2017-1386

Minematsu, N., Guo, C., & Hirose, K. (2003). CART-based factor analysis of intelligibility reduction in Japanese English. *8th European Conference on Speech Communication and Technology (Eurospeech 2003)*, 2069-2072. https://doi.org/10.21437/Eurospeech.2003-599

Minematsu, N., Okabe, K., Ogaki, K., & Hirose, K. (2011). Measurement of objective intelligibility of Japanese accented English using ERJ (English read by Japanese) database. *Interspeech 2011*, 1481-1484. https://doi.org/10.21437/Interspeech.2011-310

Minematsu, N., Tomiyama, Y., Yoshimoto, K., Shimizu, K., Nakagawa, S., Dantsuji, M., & Makino, S. (2004). Development of English speech database read by Japanese to support CALL research. *Proc. ICA*, *1*(2004), 557-560.

Moustroufas, N., & Digalakis, V. (2007). Automatic pronunciation evaluation of foreign speakers using unknown text. *Computer Speech & Language*, *21*, 219-230. https://doi.org/10.1016/j.csl.2006.04.001

Muñoz, C. (2014). Exploring young learners' foreign language learning awareness. *Language Awareness*, *23*(1-2), 24-40. https://doi.org/10.1080/09658416.2013.863900

Munro, M. J., & Derwing, T. M. (1995). Foreign Accent, Comprehensibility, and Intelligibility in the Speech of Second Language Learners. *Language Learning*, *45*(1), 73-97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x

Munro, M. J., & Derwing, T. M. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech The Role of Speaking Rate. *Studies in Second Language Acquisition*, *23*(4), 451-468. https://doi.org/10.1017/S0272263101004016

Munro, M. J., & Derwing, T. M. (2006). The functional load principle in ESL pronunciation instruction : An exploratory study. *System*, *34*(4), 520-531. https://doi.org/10.1016/j.system.2006.09.004

Munro, M. J., Derwing, T. M., & Holtby, A. K. (2012). Evaluating Individual Variability in Foreign Accent Comprehension. *Pronunciation in Second Language Learning and Teaching Proceedings*, *3*(1).

Munro, M. J., Derwing, T. M., & Morton, S. L. (2006). The Mutual Intelligibility of L2 Speech. *Studies in Second Language Acquisition*, *28*(1), 111-131. Récupérée décembre 9, 2024, à partir de http://www.jstor.org/stable/44487040

Nagle, C., Trofimovich, P., & Bergeron, A. (2019). Toward a dynamic view of second language comprehensibility. *Studies in Second Language Acquisition*, *41*(4), 647-672. https://doi.org/10.1017/S0272263119000044

Nagle, C., Trofimovich, P., O'Brien, M. G., & Kennedy, S. (2022). Comprehensible to Whom ? Examining Rater, Speaker, and Interlocutor Perspectives on Comprehensibility in an Interactive Context. *The Modern Language Journal*. https://doi.org/10.1111/modl.12809

Naijo, S., Ito, A., & Nose, T. (2021). Improvement of Automatic English Pronunciation Assessment with Small Number of Utterances Using Sentence Speakability. *Interspeech 2021*, 4473-4477. https://doi.org/10.21437/Interspeech.2021-1132

Nakanishi, N., & Coulange, S. (2024). Measuring speech rhythm through automated analysis of syllabic prominences. "*Prosodic features of language learners' fluency*" *Satellite Workshop of Speech Prosody*. https://hal.science/hal-04666098

Nakanishi, N., & Coulange, S. (2025). Beyond Intuition : Identifying Key Factors Affecting L2 Speech Comprehensibility. *The 11th International Symposium on the Acquisition of Second Language Speech*.

Nakanishi, N., Musty, N., Ōtake, S., Ying, T. S., Ebihara, Y., & Fujimura, K. (2023a, janvier). *Global perspectives listening & speaking Book 1*. Seibidō.

Nakanishi, N., Musty, N., Ōtake, S., Ying, T. S., Ebihara, Y., & Fujimura, K. (2023b, février). *Global perspectives listening & speaking Book 2*. Seibidō.

Nakanishi, N., Musty, N., Ōtake, S., Ying, T. S., & Ellis, M. (2024a, janvier). *Global Perspectives Reading & Writing Book 1*. Seibidō.

Nakanishi, N., Musty, N., Ōtake, S., Ying, T. S., & Ellis, M. (2024b, février). *Global Perspectives Reading & Writing Book 2*. Seibidō.

Neri, A., Cucchiarini, C., & Strik, H. (2002). Feedback in computer assisted pronunciation training : technology push or demand pull ? *7th International Conference on Spoken Language Processing*, 1209-1212. https://doi.org/10.21437/ICSLP.2002-246

Neumeyer, L., Franco, H., Digalakis, V., & Weintraub, M. (2000). Automatic scoring of pronunciation quality. *Speech Communication*, *30*(2), 83-93. https://doi.org/10.1016/S0167-6393(99)00046-1

Neumeyer, L., Franco, H., Weintraub, M., & Price, P. (1996). Automatic text-independent pronunciation scoring of foreign language student speech. *Proc. 4th International Conference on Spoken Language Processing (ICSLP 1996)*, 1457-1460.

Nishioka, M., Kato, T., Sugahara, M., & Coulange, S. (2025). 日本人小学生による英語復唱音声における語彙強勢の分析 [Analysis of Lexical Stress in English Repetition Speech Produced by Japanese Primary School Children]. 日本音響学会第 *153* 回研究発表会 [*Spring Meeting of the Acoustic Society of Japan*], *2025 Spring Meeting of the Acoustical Society of Japan*.

Nord, L., Kruckenberg, A., & Fant, G. (1990). Some timing studies of prose, poetry and music [Neuropeech '89]. *Speech Communication*, *9*(5), 477-483. https://doi.org/10.1016/0167-6393(90)90023-3

Ou, S., Yeh, R., & Chuang, Z. (2012). Units of analysis, intelligibility evaluation and phonological cores of EIL. *Poster presented at the 4th annual conference on Pronunciation in Second Language Learning and Teaching, Vancouver*.

Owoicho, P., Camp, J., & Kenter, T. (2024). A Study of the Sensitivity of Subjective Listening Tests to Inter-sentence Pause Durations in English Speech. *Speech Prosody 2024*, 462-466. https://doi.org/10.21437/SpeechProsody.2024-94

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech : An ASR corpus based on public domain audio books. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5206-5210. https://doi.org/10.1109/ICASSP.2015.7178964

Pearson Education, I. (2022). *Versant english test*. https://www.pearson.com/english/versant/our-tests/english-speaking-test.html

Pearson PTE. (2024). Scoring [Consulté le 27 novembre 2024]. https://www.pearsonpte.com/scoring

Piccardo, E. (2016). Common European Framework of Reference for Languages : Learning, Teaching, Assessment. Phonological Scale Revision Process Report. https://rm.coe.int/phonological-scale-revision-process-report-cefr/168073fff9

Pike, K. (1945). *The Intonation of American English*. Univ. of Michigan Press.

Pommée, T., Balaguer, M., Mauclair, J., Pinquier, J., & Woisard, V. (2022). Intelligibility and comprehensibility : A Delphi consensus study. *International Journal of Language & Communication Disorders*, *57*(1), 21-41. https://doi.org/10.1111/1460-6984.12672

Pupier, A., Coavoux, M., Lecouteux, B., & Goulian, J. (2024). Une approche par graphe pour l'analyse syntaxique en dépendances de bout en bout de la parole. In M. Balaguer, N. Bendahman, L.-M. Ho-dac, J. Mauclair, J. G Moreno & J. Pinquier (Éd.), *Actes de la 31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1 : articles longs et prises de position* (p. 234-244). ATALA ; AFPC. https://aclanthology.org/2024.jeptalnrecital-taln.16/

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision. https://doi.org/10.48550/arXiv.2212.04356

Raso, T., Erickson, D., Coulange, S., Lundmark, M. S., & Frid, J. (2024). Acoustic, articulatory and perceptual characteristics of Topic Prosodic Forms in English utterances. *Seminário Internacional de Fonologia*.

Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J., Chou, J.-C., Yeh, S.-L., Fu, S.-W., Liao, C.-F., Rastorgueva, E., Grondin, F., Aris, W., Na, H., Gao, Y., ... Bengio, Y. (2021). SpeechBrain : A General-Purpose Speech Toolkit. https://doi.org/10.48550/arXiv.2106.04624

Rogerson-Revell, P. M. (2021). Computer-Assisted Pronunciation Training (CAPT) : Current Issues and Future Directions. *RELC Journal*, *52*(1), 189-205. https://doi.org/10.1177/0033688220977406

Romano, J., Kromrey, J., Coraggio, J., & Skowronek, J. Appropriate statistics for ordinal level data : Should we really be using t-test and Cohen'sd for evaluating group differences on the NSSE and other surveys ? In : In *annual meeting of the Florida Association of Institutional Research*. 2006.

Sacks, H. (1992). *Lectures on Conversation* (G. Jefferson, Éd.). Blackwell.

Saito, K. (2021). What Characterizes Comprehensible and Native-like Pronunciation Among English-as-a-Second-Language Speakers ? Meta-Analyses of Phonological, Rater, and Instructional Factors. *TESOL Quarterly*, *55*(3), 866-900. https://doi.org/10.1002/tesq.3027

Saito, K., Macmillan, K., Kachlicka, M., Kunihara, T., & Minematsu, N. (2022). Automated assessment of second language comprehensibility : Review, training, validation, and generalization studies. *Studies in Second Language Acquisition*, *45*(1), 234-263. https://doi.org/10.1017/S0272263122000080

Saito, K., Trofimovich, P., Abe, M., & In'nami, Y. (2020). Dunning-Kruger effect in second language speech learning : How does self perception align with other perception over time ? *Learning and Individual Differences*, *79*, 101849. https://doi.org/10.1016/j.lindif.2020.101849

Saito, K., Trofimovich, P., & Isaacs, T. (2015). Using Listener Judgments to Investigate Linguistic Influences on L2 Comprehensibility and Accentedness : A Validation and Generalization Study. *Applied Linguistics*, *38*(4), 439-462. https://doi.org/10.1093/applin/amv047

Segalowitz, N. (2010). *Cognitive bases of second language fluency*. Routledge.

Shahin, M. A., Epps, J., & Ahmed, B. (2016). Automatic Classification of Lexical Stress in English and Arabic Languages Using Deep Learning. *Interspeech*, 175-179.

Shea, C., & Leonard, K. (2019). Evaluating measures of pausing for second language fluency research. *Canadian Modern Language Review*, *75*(3), 216-235. https://doi.org/10.3138/cmlr.2018-025

Shen, Y., Yasukagawa, A., Saito, D., Minematsu, N., & Saito, K. (2021). Optimized prediction of fluency of L2 English based on interpretable network using quantity of phonation and quality of pronunciation. *2021 IEEE Spoken Language Technology Workshop (SLT)*.

Shibata, T., & Shibata, R. (1990). To what extent can accents distinguish homophones ? *Keiryoo Kokugogaku*, *17*(7), 311-323.

Shigemitsu, Y. (2007). A pause in conversation for Japanese native speakers : a case study of successful and unsuccessful conversation in terms of pause though intercultural communication. *Academic Report, Tokyo Polytechnic Univ.*, *30*(2), 11-18. https://cir.nii.ac.jp/crid/1520290882531293440

Shobaki, K., Hosom, J.-P., & Cole, R. (2000). The OGI kids' speech corpus and recognizers. *Sixth International Conference on Spoken Language Processing, ICSLP 2000 / Interspeech 2000*, 258-261.

Siegman, A., & Feldstein, S. (1979). *Of speech and time*. John Wiley & Sons.

Siegman, A., & Pope, B. (1966). Ambiguity and verbal fluency in the TAT. *Journal of Consulting Psychology*, *30*(3), 239-245. https://doi.org/10.1037/h0023374

Simon, A.-C., & Christodoulides, G. (2016). Frontières prosodiques perçues : corrélats acoustiques et indices syntaxiques. *Langue française*, *N°191*(3), 83-106. https://doi.org/10.3917/lf.191.0083

Smiljanić, R., & Bradlow, A. R. (2005). Production and perception of clear speech in Croatian and Englisha). *The Journal of the Acoustical Society of America*, *118*(3), 1677-1688. https://doi.org/10.1121/1.2000788

Sugafuji, M. (1996). 大阪・東京アクセント音声辞典 [*Ōsaka-Tōkyō Accent Dictionary*. Maruzen. https://ci.nii.ac.jp/ncid/BB07124137

Sugahara, M. (2011). Identification of English primary stress and bias toward strong word-initial syllables : native vs. Japanese listeners. *ICPhS*, 1918-1921.

Sugahara, M. (2016). Is Japanese listeners' perception of English stress influenced by the antepenultimate accent in Japanese ? Comparison with English and Korean listeners. *Doshisha Studies in English*, *96*, 61-111.

Sugahara, M. (2020). Assignment of English Lexical Stress by Japanese and Seoul Korean Learners of English [Presented at the 28th Japanese/Korean Linguistics Conference Satellite Workshop : Experimental Phonetics and Phonology]. https://researchmap.jp/read0122533/presentations/32070371

Sugahara, M., Coulange, S., & Kato, T. (2023). 意識されている強勢 *vs.* 発話における強勢 — 日本人と韓国人の大学生による英単語への主強勢付与の比較 [*Stress awareness vs. stress production : Comparison of primary stress assignment to English words between Japanese and Korean Univ. students*] [第 347 回日本音声学会研究例会 [347th regular meeting of the Phonetic Society of Japan]]. http://www.psj.gr.jp/jpn/regular-meeting/347th.html

Sugahara, M., Coulange, S., & Kato, T. (2024). English Lexical Stress in Awareness and Production : Native and Non-native Speakers. *The 19th LabPhon Conference, 27-29 June 2024, Seoul, Korea*.

Suzuki, S., & Kormos, J. (2020). linguistic dimensions of comprehensibility and perceived fluency : an investigation of complexity, accuracy, and fluency in second language argumentative speech. *Studies in Second Language Acquisition*, *42*(1), 143-167. https://doi.org/10.1017/S0272263119000421

Suzuki, S., & Kormos, J. (2023). The multidimensionality of second language oral fluency : Interfacing cognitive fluency and utterance fluency. *Studies in Second Language Acquisition*, *45*(1), 38-64. https://doi.org/10.1017/S0272263121000899

Suzuki, S., Kormos, J., & Uchihara, T. (2021). The Relationship Between Utterance and Perceived Fluency : A Meta-Analysis of Correlational Studies. *The Modern Language Journal*, *105*(2), 435-463. https://doi.org/10.1111/modl.12706

Tajima, K., Port, R., & Dalby, J. (1997). Effects of temporal correction on intelligibility of foreign-accented English. *Journal of Phonetics*, *25*(1), 1-24. https://doi.org/10.1006/jpho.1996.0031

Tan, T. P. (2008). *Reconnaissance automatique de la parole non-native* [thèse de doct., Univ. Grenoble 1]. http://www.theses.fr/2008GRE10096

Tauberer, J. (2008). Predicting intrasentential pauses : is syntactic structure useful ? *Proc. Speech Prosody 2008*, 405-408.

Tavakoli, P. (2010). Pausing patterns : differences between L2 learners and native speakers. *ELT Journal*, *65*(1), 71-79. https://doi.org/10.1093/elt/ccq020

Tepperman, J., & Narayanan, S. (2005). Automatic syllable stress detection using prosodic features for pronunciation evaluation of language learners. *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005*, *1*, 937-940.

Thomson, R. (2015). Fluency. In *The Handbook of English Pronunciation* (p. 209-226). John Wiley ; Sons, Ltd. https://doi.org/10.1002/9781118346952.ch12

Thomson, R. (2017). Measurement of accentedness, intelligibility, and comprehensibility. In *Assessment in Second Language Pronunciation* (p. 11-29). Routledge.

Tomita, Y., Gao, Y., Minematsu, N., Nakanishi, N., & Saito, D. (2024). Analysis and Visualization of Directional Diversity in Listening Fluency of World Englishes Speakers in the Framework of Mutual Shadowing. *Interspeech 2024*, 4024-4028. https://doi.org/10.21437/Interspeech.2024-1373

Tortel, A. (2021). Le rythme en anglais oral : considérations théoriques et illustrations sur corpus. *Recherche et pratiques pédagogiques en langues - Cahiers de l'APLIUT*, (Vol. 40 N°1). https://doi.org/10.4000/apliut.8857

Tortel, A., & Hirst, D. (2010). Rhythm metrics and the production of English L1/L2. *Proc. Speech Prosody 2010*, paper 959.

Trofimovich, P., Isaacs, T., Kennedy, S., Saito, K., & Crowther, D. (2016). Flawed self-assessment : Investigating self- and other-perception of second language speech. *Bilingualism : Language and Cognition*, *19*(1), 122-140. https://doi.org/10.1017/S1366728914000832

Trofimovich, P., Nagle, C. L., O'Brien, M. G., Kennedy, S., Taylor Reid, K., & Strachan, L. (2020). Second language comprehensibility as a dynamic construct. *Journal of Second Language Pronunciation*, *6*(3), 430-457.

Trofimovich, P., Tekin, O., & Lindberg, R. (2024). Listening and comprehensibility. In E. Wagner, A. O. Batty & E. Galaczi (Éd.), *The Routledge handbook of second language acquisition and listening* (p. 201-211). Routledge. https://doi.org/10.4324/9781003219552-17

Trouvain, J., Bonneau, A., Colotte, V., Fauth, C., Fohr, D., Jouvet, D., Jügler, J., Laprie, Y., Mella, O., Möbius, B., & Zimmerer, F. (2016). The IFCASL Corpus of French and German Non-native and Native Read Speech. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk & S. Piperidis (Éd.), *Proc. of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (p. 1333-1338). European Language Resources Association (ELRA). https://aclanthology.org/L16-1212

Trouvain, J. (2004). *Tempo Variation in Speech Production : Implications for Speech Synthesis* [thèse de doct., Saarland Univ.].

Truong, Q.-T., Kato, T., & Yamamoto, S. (2018). Automatic assessment of L2 English word prosody using weighted distances of F0 and intensity contours. *Interspeech 2018*.

Tsunemoto, A., McAndrews, M., Trofimovich, P., & Friginal, E. (2023). Listener perceptions of customer service agents' performance. *Journal of Second Language Pronunciation*, *9*(2), 234-262. https://doi.org/10.1075/jslp.21027.tsu

Tsunemoto, A., Trofimovich, P., Blanchet, J., Bertrand, J., & Kennedy, S. (2022). Effects of benchmarking and peer-assessment on French learners' self-assessments of accentedness, comprehensibility, and fluency. *Foreign Language Annals*, *55*(1), 135-154. https://doi.org/10.1111/flan.12571

Vaissière, J. (1983). Language-independent prosodic features. In *Prosody : Models and Measurements* (p. 53-65). Springer Verlag. https://shs.hal.science/halshs-00703571

Vaissière, J. (1991). Rhythm, accentuation and final lengthening in French. In J. Sundberg, L. Nord & R. Carlson (Éd.), *Music, Language, Speech and Brain : Proc. of an International Symposium at the Wenner-Gren Center, Stockholm, 5–8 September 1990* (p. 108-120). Macmillan Education UK. https://doi.org/10.1007/978-1-349-12670-5_10

Vaissière, J., & Michaud, A. (2006). Prosodic constituents in French : a data-driven approach. In Y. K. I. Fónagy & T. Moriguchi (Éd.), *Prosody and syntax* (p. 47-64). John Benjamins. https://hal.science/hal-00130794

van Leyden, K., & van Heuven, V. J. (1996). Lexical stress and spoken word recognition. *Linguistics in the Netherlands*, *13*, 159-170. https://doi.org/10.1075/avt.13.16ley

Walker, R., Low, E.-L., & Setter, J. (2021). *English pronunciation for a global world*. Oxford Univ. Press.

WANG, X., ZHANG, J., NISHIDA, M., & YAMAMOTO, S. (2015). Phoneme Set Design for Speech Recognition of English by Japanese. *IEICE Transactions on Information and Systems*, *E98.D*(1), 148-156. https://doi.org/10.1587/transinf.2014EDP7168

Weischedel, R., Palmer, M., Marcus, M., Hovy, E., Pradhan, S., Ramshaw, L., Xue, N., Taylor, A., Kaufman, J., Franchini, M., El-Bachouti, M.,

Belvin, R., & Houston, A. (2013). OntoNotes Release 5.0. https://doi.org/10.35111/xmhb-2b84

Wenk, B., & Wioland, F. (1982). Is French really syllable-timed ? *Journal of Phonetics*, *2*(10), 193-216.

White, S. (1989). Backchannels across Cultures : A Study of Americans and Japanese. *Language in Society*, *18*(1), 59-76. http://www.jstor.org/stable/4168001

Wilkes, A. L., & Kennedy, R. A. (1969). Relationship between pausing and retrieval latency in sentences of varying grammatical form. *Journal of Experimental Psychology*, *79*(2, Pt.1), 241-245.

Witt, S. (1999). *Use of Speech Recognition in Computer-assisted Language Learning* [thèse de doct., Department of Engineering, Univ. of Cambridge].

Witt, S. (2012). Automatic Error Detection in Pronunciation Training : Where we are and where we need to go. *International Symposium on audiovisual detection on errors in pronunciation training*, 1-8.

Witton-Davies, G. (2018). Pauses, Pause Position, and Fluency. In *Reconceptualizing English Language Teaching and Learning in the 21st Century A Special Monograph in Memory of Professor Kai-chong Cheung* (p. 122-133). Crane.

Xu, J., Jones, E., Laxton, V., & Galaczi, E. (2021). Assessing L2 English speaking using automated scoring technology : examining automarker reliability. *Assessment in Education : Principles, Policy & Practice*, *28*, 1-26. https://doi.org/10.1080/0969594X.2021.1979467

Yoon, S. Y., Bhat, S., & Zechner, K. (2012). Vocabulary profile as a measure of vocabulary sophistication. In *Proc. of the Seventh Workshop on Building Educational Applications* (p. 180-189). Association for Computational Linguistics.

Zellner, B. (1994). Pauses and the temporal structure of speech. In E. Keller (Éd.), *Fundamentals of speech synthesis and speech recognition* (p. 41-62). John Wiley. http://cogprints.org/884/

Zielinski, B. (2006). The Intelligibility cocktail : an interaction between speaker and listener ingredients. *Prospect*, *21*(1), 22-45.

Annexes

A Grilles d'évaluation de la prononciation

A.1 Descripteurs du CECRL (2001)

	MAÎTRISE DU SYSTÈME PHONOLOGIQUE
C2	Comme C1
C1	Peut varier l'intonation et placer l'accent phrastique correctement afin d'exprimer de fines nuances de sens.
B2	A acquis une prononciation et une intonation claires et naturelles.
B1	La prononciation est clairement intelligible même si un accent étranger est quelquefois perceptible et si des erreurs de prononciation proviennent occasionnellement.
A2	La prononciation est en général suffisamment claire pour être comprise malgré un net accent étranger mais l'interlocuteur devra parfois faire répéter.
A1	La prononciation d'un répertoire très limité d'expressions et de mots mémorisés est compréhensible avec quelque effort pour un locuteur natif habitué aux locuteurs du groupe linguistique de l'apprenant/utilisateur.

FIG. 11.3 : Échelle « Maîtrise du système phonologique » du CECRL édition 2001, p. 92

A.2 Descripteurs du CECRL (2018)

MAÎTRISE PHONOLOGIQUE			
	MAÎTRISE GENERALE DU SYSTEME PHONOLOGIQUE	**ARTICULATION DES SONS**	**TRAITS PROSODIQUES**
C2	Peut utiliser tout l'éventail des traits phonologiques de la langue cible avec un haut degré de maîtrise – y compris les traits prosodiques tels que l'accent tonique et phrastique, le rythme et l'intonation-, de façon à ce que les moindres détails de son message soient clairs et précis. La présence d'un accent venant d'autres langues n'affecte aucunement ni la compréhension ni l'efficacité de la transmission et de la mise en valeur du sens.	Peut en principe articuler tous les sons de la langue cible avec clarté et précision.	Peut utiliser correctement et de façon efficace les traits prosodiques (par ex. l'accent, le rythme et l'intonation) afin de transmettre de fines nuances de sens (par ex. pour différencier et mettre en valeur).
C1	Peut utiliser avec une assez bonne maîtrise tout l'éventail des traits phonologiques de la langue cible, de façon à être toujours intelligible. Peut articuler pratiquement tous les sons de la langue cible ; on peut noter la présence d'un accent venant d'autre(s) langue(s) mais cela n'affecte en rien la compréhension.	Peut articuler pratiquement tous les sons de la langue cible avec un haut degré de maitrise. Peut en général s'auto corriger quand il/elle a manifestement mal prononcé un son.	Peut prononcer un discours fluide et intelligible en ne faisant que de rares erreurs d'accent, de rythme et/ou d'intonation qui n'affectent ni la compréhension ni l'efficacité. Peut varier l'intonation et placer correctement l'accent pour exprimer exactement ce qu'il souhaite dire.
B2	Peut en général utiliser la bonne intonation, placer correctement l'accent et articuler clairement les sons isolés ; l'accent a tendance à subir l'influence de l'une ou l'autre des langues qu'il/elle parle, mais l'impact sur la compréhension est négligeable ou nul.	Peut, dans de longues parties d'énoncés, articuler clairement une grande quantité des sons de la langue cible ; le tout est intelligible malgré quelques erreurs systématiques de prononciation. Peut, à partir de son répertoire, prédire avec une certaine précision les traits phonologiques de la plupart des mots non familiers (par ex. l'accent tonique en lisant).	Peut utiliser des traits prosodiques (par ex. l'accent, l'intonation, le rythme,) pour faire passer le message qu'il a l'intention de transmettre, mais l'influence des autres langues qu'il/elle parle est notable.
B1	La prononciation est en général intelligible ; l'intonation et l'accentuation des énoncés et des mots sont presque corrects. L'une ou l'autre des langues qu'il/elle parle a en général une influence sur l'accent et la compréhension peut en être affectée.	Est en général totalement intelligible, bien qu'il/elle fasse régulièrement des erreurs de prononciation de sons et de mots isolés qui ne lui sont pas familiers.	Peut transmettre son message de façon intelligible malgré une forte influence de l'une ou l'autre des langues qu'il/elle parle sur l'accent, l'intonation et/ou le rythme.
A2	La prononciation est en général suffisamment claire pour être comprise mais l'interlocuteur devra parfois faire répéter. Une forte influence de l'une ou l'autre des langues parlées sur l'accent, le rythme et l'intonation peut affecter la compréhension et requiert la participation des interlocuteurs. La prononciation des mots familiers est cependant claire.	La prononciation est en général intelligible dans des situations d'échanges quotidiens simples, pourvu que l'interlocuteur fasse l'effort de comprendre certains sons spécifiques. Une mauvaise prononciation systématique des phonèmes n'affecte pas la compréhension, pourvu que l'interlocuteur fasse l'effort de reconnaitre l'influence de la langue du locuteur sur la prononciation et s'y adapte.	Peut utiliser de façon intelligible les traits prosodiques des mots et expressions quotidiens, malgré une forte influence de l'une ou l'autre des langues qu'il/elle parle sur l'accent, l'intonation et/ou le rythme. Les traits prosodiques (par ex. l'accent tonique) des mots familiers et quotidiens et des énoncés simples sont convenables.
A1	La prononciation d'un répertoire très limité d'expressions et de mots mémorisés est compréhensible avec quelque effort pour des interlocuteurs habitués aux locuteurs de son groupe linguistique. Peut reproduire correctement un nombre limité de sons ainsi que d'accents sur des mots et des expressions simples et familiers.	Peut, s'il/elle est guidé de manière précise, reproduire correctement des sons dans la langue cible. Peut articuler un nombre tellement limité de sons que l'interlocuteur doit proposer de l'aide pour que les paroles soient intelligibles (par ex. répéter correctement et demander la répétition de nouveaux sons).	Peut utiliser de façon intelligible les traits prosodiques d'un répertoire limité de mots et d'expressions simples, malgré une très forte influence de l'accent, du rythme, et/ou de l'intonation de l'une ou l'autre des langues qu'il parle ; son interlocuteur doit se montrer coopératif.

A.3 Grille d'évaluation CLES B2 Interaction Orale

Grille d'évaluation de l'interaction orale CLES B2

Critères		Validé		Non validé
		B2	B1	
Pragmatiques	**1 .Situation :** dans le cadre de la situation donnée : - Négocie afin de mener à bien la mission - Sait se positionner en fonction de son rôle dans l'échange			
	2. Contenu : - Utilise des arguments variés et pertinents issus des documents en ajoutant éventuellement des idées personnelles - mobilise au moins deux arguments tirés du dossier relatifs à son rôle			
	3. Interaction : - Sait interagir : prend son tour et l'initiative de la parole quand il convient, sait relancer l'échange si nécessaire - réagit aux sollicitations sans relancer			
	4. Aisance : - Exprime ses idées avec fluidité sans faire de longues pauses (hésitations tolérées) - Exprime ses idées malgré des pauses pour chercher ses mots			
Linguistiques	**5. Phonologie :** - prononciation et intonation suffisamment claires pour être aisément compris(e), même si un accent subsiste -globalement compréhensible malgré l'accent étranger et/ou des erreurs de prononciation			
	6. Cohérence : - dispose d'outils linguistiques variés pour lier, nuancer et adapter son discours. - dispose d'outils linguistiques simples (conjonctions, adverbes) pour lier les informations et présenter ses arguments			
	7. Correction grammaticale : - fait preuve d'un contrôle grammatical assez élevé qui lui permet d'éviter les malentendus (erreurs non systématiques tolérées) - utilise de façon assez exacte un répertoire de structures et « schémas » fréquents			
	8. Lexique : - utilise un lexique juste et varié (quelques lacunes tolérées) - utilise un lexique limité mais approprié à la tâche, éventuellement tiré des documents			

Résultat : B2[1] ☐ B1 ☐ insuffisant ☐ cas de jury[2] ☐

[1] B2 est validé si et seulement si chacun des critères est validé au niveau B2.En cas de non validation d'un des critères, le jury reste souverain et décidera ou non d'attribuer la compétence de production orale.
[2] Se référer à la liste des cas de jury

A.4 Descripteurs du TOEFL iBT

TOEFL iBT®
Independent Speaking Rubric

SCORE	GENERAL DESCRIPTION	DELIVERY	LANGUAGE USE	TOPIC DEVELOPMENT
4	The response fulfills the demands of the task, with at most minor lapses in completeness. It is highly intelligible and exhibits sustained, coherent discourse. A response at this level is characterized by all of the following:	Generally well-paced flow (fluid expression). Speech is clear. It may include minor lapses, or minor difficulties with pronunciation or intonation patterns, which do not affect overall intelligibility.	The response demonstrates effective use of grammar and vocabulary. It exhibits a fairly high degree of automaticity with good control of basic and complex structures (as appropriate). Some minor (or systematic) errors are noticeable but do not obscure meaning.	Response is sustained and sufficient to the task. It is generally well developed and coherent; relationships between ideas are clear (or there is a clear progression of ideas).
3	The response addresses the task appropriately but may fall short of being fully developed. It is generally intelligible and coherent, with some fluidity of expression, though it exhibits some noticeable lapses in the expression of ideas. A response at this level is characterized by at least two of the following:	Speech is generally clear, with some fluidity of expression, though minor difficulties with pronunciation, intonation, or pacing are noticeable and may require listener effort at times (though overall intelligibility is not significantly affected).	The response demonstrates fairly automatic and effective use of grammar and vocabulary, and fairly coherent expression of relevant ideas. Response may exhibit some imprecise or inaccurate use of vocabulary or grammatical structures or be somewhat limited in the range of structures used. This may affect overall fluency, but it does not seriously interfere with the communication of the message.	Response is mostly coherent and sustained and conveys relevant ideas/information. Overall development is somewhat limited, usually lacks elaboration or specificity. Relationships between ideas may at times not be immediately clear.
2	The response addresses the task, but development of the topic is limited. It contains intelligible speech, although problems with delivery and/or overall coherence occur; meaning may be obscured in places. A response at this level is characterized by at least two of the following:	Speech is basically intelligible, though listener effort is needed because of unclear articulation, awkward intonation, or choppy rhythm/pace; meaning may be obscured in places.	The response demonstrates limited range and control of grammar and vocabulary. These limitations often prevent full expression of ideas. For the most part, only basic sentence structures are used successfully and spoken with fluidity. Structures and vocabulary may express mainly simple (short) and/or general propositions, with simple or unclear connections made among them (serial listing, conjunction, juxtaposition).	The response is connected to the task, though the number of ideas presented or the development of ideas is limited. Mostly basic ideas are expressed with limited elaboration (details and support). At times relevant substance may be vaguely expressed or repetitious. Connections of ideas may be unclear.
1	The response is very limited in content and/or coherence or is only minimally connected to the task, or speech is largely unintelligible. A response at this level is characterized by at least two of the following:	Consistent pronunciation, stress and intonation difficulties cause considerable listener effort; delivery is choppy, fragmented, or telegraphic; frequent pauses and hesitations.	Range and control of grammar and vocabulary severely limit or prevent expression of ideas and connections among ideas. Some low-level responses may rely heavily on practiced or formulaic expressions.	Limited relevant content is expressed. The response generally lacks substance beyond expression of very basic ideas. Speaker may be unable to sustain speech to complete the task and may rely heavily on repetition of the prompt.
0	Speaker makes no attempt to respond OR response is unrelated to the topic.			

toefl ibt®

TOEFL iBT®
Integrated Speaking Rubric

SCORE	GENERAL DESCRIPTION	DELIVERY	LANGUAGE USE	TOPIC DEVELOPMENT
4	The response fulfills the demands of the task, with at most minor lapses in completeness. It is highly intelligible and exhibits sustained, coherent discourse. A response at this level is characterized by all of the following:	Speech is generally clear, fluid and sustained. It may include minor lapses or minor difficulties with pronunciation or intonation. Pace may vary at times as the speaker attempts to recall information. Overall intelligibility remains high.	The response demonstrates good control of basic and complex grammatical structures that allow for coherent, efficient (automatic) expression of relevant ideas. Contains generally effective word choice. Though some minor (or systematic) errors or imprecise use may be noticeable, they do not require listener effort (or obscure meaning).	The response presents a clear progression of ideas and conveys the relevant information required by the task. It includes appropriate detail, though it may have minor errors or minor omissions.
3	The response addresses the task appropriately, but may fall short of being fully developed. It is generally intelligible and coherent, with some fluidity of expression, though it exhibits some noticeable lapses in the expression of ideas. A response at this level is characterized by at least two of the following:	Speech is generally clear, with some fluidity of expression, but it exhibits minor difficulties with pronunciation, intonation, or pacing and may require some listener effort at times. Overall intelligibility remains good, however.	The response demonstrates fairly automatic and effective use of grammar and vocabulary, and fairly coherent expression of relevant ideas. Response may exhibit some imprecise or inaccurate use of vocabulary or grammatical structures or be somewhat limited in the range of structures used. Such limitations do not seriously interfere with the communication of the message.	The response is sustained and conveys relevant information required by the task. However, it exhibits some incompleteness, inaccuracy, lack of specificity with respect to content, or choppiness in the progression of ideas.
2	The response is connected to the task, though it may be missing some relevant information or contain inaccuracies. It contains some intelligible speech, but at times problems with intelligibility and/or overall coherence may obscure meaning. A response at this level is characterized by at least two of the following:	Speech is clear at times, though it exhibits problems with pronunciation, intonation, or pacing and so may require significant listener effort. Speech may not be sustained at a consistent level throughout. Problems with intelligibility may obscure meaning in places (but not throughout).	The response is limited in the range and control of vocabulary and grammar demonstrated (some complex structures may be used, but typically contain errors). This results in limited or vague expression of relevant ideas and imprecise or inaccurate connections. Automaticity of expression may only be evident at the phrasal level.	The response conveys some relevant information but is clearly incomplete or inaccurate. It is incomplete if it omits key ideas, makes vague reference to key ideas, or demonstrates limited development of important information. An inaccurate response demonstrates misunderstanding of key ideas from the stimulus. Typically, ideas expressed may not be well-connected or cohesive so that familiarity with the stimulus is necessary to follow what is being discussed.
1	The response is very limited in content or coherence or is only minimally connected to the task. Speech may be largely unintelligible. A response at this level is characterized by at least two of the following:	Consistent pronunciation and intonation problems cause considerable listener effort and frequently obscure meaning. Delivery is choppy, fragmented, or telegraphic. Speech contains frequent pauses and hesitations.	Range and control of grammar and vocabulary severely limit or prevent expression of ideas and connections among ideas. Some low-level responses may rely heavily on practiced or formulaic expressions.	The response fails to provide much relevant content. Ideas that are expressed are often inaccurate, limited to vague utterances, or repetitions (including repetition of prompt).
0	Speaker makes no attempt to respond OR response is unrelated to the topic.			

toefl ibt®

A.5 Descripteurs du TOEFL iTP

ETS TOEFL ITP

Speaking Test Score Descriptors

Score Range	CEFR Level	Proficiency Descriptors
64–68	C1	**Test takers at this level are typically able to:** • express themselves fluently with very little effort or hesitation • produce speech that is clear and well-paced • use stress and intonation effectively to support the meaning of what is being said • use a broad range of grammatical structures and vocabulary to express themselves with precision on most topics
58–63	B2	**Test takers at this level are typically able to:** • produce stretches of mostly well-paced and fluent speech; however, they may hesitate at times as they try to recall certain expressions • use stress and intonation to convey meaning, though there may be some errors or native language influence • use a sufficient range of grammar and vocabulary to give clear descriptions and to express opinions comfortably on most topics
48–57	B1	**Test takers at this level are typically able to:** • produce intelligible speech, although certain unfamiliar words are mispronounced, and pausing for planning and repair is evident • use stress, intonation and rhythm somewhat effectively to convey a message, although these may be influenced by their native language • use a good range of vocabulary related to familiar, everyday topics • express themselves on familiar subjects using basic grammatical structures but as topics become more unfamiliar and/or more complex, errors are more common and cause listener effort
41–47	A2	**Test takers at this level are typically able to:** • speak clearly enough to be understood with some listener effort when talking about familiar, everyday topics; however, pronunciation and word stress errors are noticeable and highly influenced by the speaker's native language • produce choppy speech, with frequent pauses and false starts • use a limited range of grammar and vocabulary • speak in short, memorized phrases to produce brief stretches of speech

Note: Test takers who achieve a Speaking score below 41 have not met the benchmark proficiency for A2 level.

A.6 TOEIC Speaking Test

Descripteurs de niveaux

Level	
8 **Scale Score** **190 – 200**	Typically, test takers at level 8 can create connected, sustained discourse appropriate to the typical workplace. When they express opinions or respond to complicated requests, their speech is highly intelligible. Their use of basic and complex grammar is good and their use of vocabulary is accurate and precise. Test takers at level 8 can also use spoken language to answer questions and give basic information. Their pronunciation and intonation and stress are at all times highly intelligible.
7 **Scale Score** **160 – 180**	Typically, test takers at level 7 can create connected, sustained discourse appropriate to the typical workplace. They can express opinions or respond to complicated requests effectively. In extended responses, some of the following weaknesses may sometimes occur, but they do not interfere with the message: • minor difficulties with pronunciation, intonation or hesitation when creating language • some errors when using complex grammatical structures • some imprecise vocabulary Test takers at level 7 can also use spoken language to answer questions and give basic information. When reading aloud, test takers at level 7 are highly intelligible.
6 **Scale Score** **130 – 150**	Typically, test takers at level 6 are able to create a relevant response when asked to express an opinion or respond to a complicated request. However, at least part of the time, the reasons for, or explanations of, the opinion are unclear to a listener. This may be because of the following: • unclear pronunciation or inappropriate intonation or stress when the speaker must create language • mistakes in grammar • a limited range of vocabulary Most of the time, test takers at level 6 can answer questions and give basic information. However, sometimes their responses are difficult to understand or interpret. When reading aloud, test takers at level 6 are intelligible.
5 **Scale Score** **110 – 120**	Typically, test takers at level 5 have limited success at expressing an opinion or responding to a complicated request. Responses include problems such as: • language that is inaccurate, vague or repetitive • minimal or no awareness of audience • long pauses and frequent hesitations • limited expression of ideas and connections between ideas • limited vocabulary Most of the time, test takers at level 5 can answer questions and give basic information. However, sometimes their responses are difficult to understand or interpret. When reading aloud, test takers at level 5 are generally intelligible. However, when creating language, their pronunciation, intonation and stress may be inconsistent.
4 **Scale Score** **80 – 100**	Typically, test takers at level 4 are unsuccessful when attempting to explain an opinion or respond to a complicated request. The response may be limited to a single sentence or part of a sentence. Other problems may include: • severely limited language use • minimal or no audience awareness • consistent pronunciation, stress and intonation difficulties • long pauses and frequent hesitations • severely limited vocabulary Most of the time, test takers at level 4 cannot answer questions or give basic information. When reading aloud, test takers at level 4 vary in intelligibility. However, when they are creating language, speakers at level 4 usually have problems with pronunciation and intonation and stress. For more information, check the "Read Aloud Pronunciation and Intonation and Stress Ratings."
3 **Scale Score** **60 – 70**	Typically, test takers at level 3 can, with some difficulty, state an opinion, but they cannot support the opinion. Any response to a complicated request is severely limited. Most of the time, test takers at level 3 cannot answer questions and give basic information. Typically, test takers at level 3 have insufficient vocabulary or grammar to create simple descriptions. When reading aloud, speakers at level 3 may be difficult to understand. For more information, check the "Read Aloud Pronunciation and Intonation and Stress ratings."
2 **Scale Score** **40 – 50**	Typically, test takers at level 2 cannot state an opinion or support it. They either do not respond to complicated requests or the response is not at all relevant. In routine social and occupational interactions such as answering questions and giving basic information, test takers at level 2 are difficult to understand. When reading aloud, speakers at level 2 may be difficult to understand. For more information, check the "Read Aloud Pronunciation and Intonation and Stress Ratings."
1 **Scale Score** **0 – 30**	Test takers at level 1 left a significant part of the TOEIC Speaking Test unanswered. Test takers at level 1 may not have the listening or reading skills in English necessary to understand the test directions or the content of the test questions.

Grilles d'évaluation

Scoring Guide for the Read a Text Aloud Task:

Pronunciation

Score	Response Description
3	Pronunciation is highly intelligible, though the response may include minor lapses and/or other language influence.
2	Pronunciation is generally intelligible, though it includes some lapses and/or other language influence.
1	Pronunciation may be intelligible at times, but significant other language influence interferes with appropriate delivery of the text.
0	No response OR no English in the response OR response is completely unrelated to the test.

Scoring Guide for the Read a Text Aloud Task:

Intonation and Stress

Score	Response Description
3	Use of emphases, pauses, and rising and falling pitch is appropriate to the text.
2	Use of emphases, pauses, and rising and falling pitch is generally appropriate to the text, though the response includes some lapses and/or moderate other language influence.
1	Use of emphases, pauses, and rising and falling pitch is not appropriate, and the response includes significant other language influence.
0	No response OR no English in the response OR the response is completely unrelated to the test.

Scoring Guide for the Describe a Picture Task:

Score	Response Description
3	The response describes the main features of the picture. • The delivery may require some listener effort, but it is generally intelligible. • The choice of vocabulary and use of structures allows coherent expression of ideas.
2	The response is connected to the picture, but meaning may be obscured in places. • The delivery requires some listener effort. • The choice of vocabulary and use of structures may be limited and may interfere with overall comprehensibility.
1	The response may be connected to the picture, but the speaker's ability to produce intelligible language is severely limited. • The delivery may require significant listener effort. • The choice of vocabulary and use of structures is severely limited OR significantly interferes with comprehensibility.
0	No response OR no English in the response OR the response is completely unrelated to the test.

Scoring Guide for Respond to Questions (Market Survey) and Respond to Questions Using Information Provided (Agenda) Tasks:

Score	Response Description
3	The response is a full, relevant, socially appropriate reply to the question. In the case of the Agenda questions, information from the prompt is accurate. • The delivery requires little listener effort. • The choice of vocabulary is appropriate. • The use of structures fulfills the demands of the task.
2	The response is a partially effective reply to the question, but is not complete, fully appropriate, or in the case of the Agenda questions, fully accurate. • The delivery may require some listener effort but is mostly intelligible. • The choice of vocabulary may be limited or somewhat inexact, although overall meaning is clear. • The use of structures may require some listener effort for interpretation. • In the case of the Agenda questions, the speaker may locate the relevant information in the prompt but fail to distinguish it from irrelevant information or fail to transform the written language so a listener can easily understand it.
1	The response does not answer the question effectively. Relevant information is not conveyed successfully. • The delivery may impede or prevent listener comprehension. • The choice of vocabulary may be inaccurate or rely on repetition of the prompt. • The use of structures may interfere with comprehensibility.
0	No response OR no English in the response OR the response is completely unrelated to the test.

Scoring Guide for the Express an Opinion Task:

Score	Response Description
5	The response clearly indicates the speaker's choice or opinion, and support of the choice or opinion is readily intelligible, sustained, and coherent. The response is characterized by ALL of the following: • The speaker's choice or opinion is supported with reason(s), details, arguments, or exemplifications; relationships between ideas are clear. • The speech is clear with generally well-paced flow. It may include minor lapses or minor difficulties with pronunciation or intonation patterns that do not affect overall intelligibility. • Good control of basic and complex structures, as appropriate, is exhibited. Some minor errors may be noticeable but they do not obscure meaning. • The use of vocabulary is effective, with allowance for occasional minor inaccuracy.
4	The response clearly indicates the speaker's choice or opinion and adequately supports or develops the choice or opinion. • The response explains the reason(s) for the speaker's choice or opinion, although the explanation may not be fully developed; relationships between ideas are mostly clear, with occasional lapses. • Minor difficulties with pronunciation, intonation, or pacing are noticeable and may require listener effort at times, although overall intelligibility is not significantly affected. • The response demonstrates fairly automatic and effective use of grammar but may be somewhat limited in the range of structures used. • The use of vocabulary is fairly effective. Some vocabulary may be inaccurate or imprecise.
3	The response expresses a choice, preference, or opinion, but development and support of the choice or opinion is limited. • The response provides at least one reason supporting the choice, preference, or opinion. However, it provides little or no elaboration of the reason, repeats itself with no new information, is vague, or is unclear. • The speech is basically intelligible, though listener effort may be needed because of unclear articulation, awkward intonation, or choppy rhythm/pace; meaning may be obscured in places. • The response demonstrates limited control of grammar; for the most part, only basic sentence structures are used successfully. • The use of vocabulary is limited.
2	The response states a choice, preference, or opinion relevant to the prompt, but support for the choice, preference, or opinion is missing, unintelligible, or incoherent. • Consistent difficulties with pronunciation, stress, and intonation cause considerable listener effort; delivery is choppy, fragmented, or telegraphic; there may be long pauses and frequent hesitations. • Control of grammar severely limits expression of ideas and clarity of connections among ideas. • The use of vocabulary is severely limited or highly repetitious.
1	The response is limited to reading the prompt or the directions aloud OR the response fails to state an intelligible choice, preference, or opinion as required by the prompt OR the response consists of isolated words or phrases, or mixtures of the first language and English.
0	No response OR no English in the response OR the response is completely unrelated to the test.

A.7 Descripteurs du IELTS

IELTS Speaking Band Descriptors (public version)

Band	Fluency and coherence	Lexical resource	Grammatical range and accuracy	Pronunciation
9	• speaks fluently with only rare repetition or self-correction; any hesitation is content-related rather than to find words or grammar • speaks coherently with fully appropriate cohesive features • develops topics fully and appropriately	• uses vocabulary with full flexibility and precision in all topics • uses idiomatic language naturally and accurately	• uses a full range of structures naturally and appropriately • produces consistently accurate structures apart from 'slips' characteristic of native speaker speech	• uses a full range of pronunciation features with precision and subtlety • sustains flexible use of features throughout • is effortless to understand
8	• speaks fluently with only occasional repetition or self-correction; hesitation is usually content-related and only rarely to search for language • develops topics coherently and appropriately	• uses a wide vocabulary resource readily and flexibly to convey precise meaning • uses less common and idiomatic vocabulary skilfully, with occasional inaccuracies • uses paraphrase effectively as required	• uses a wide range of structures flexibly • produces a majority of error-free sentences with only very occasional inappropriacies or basic/non-systematic errors	• uses a wide range of pronunciation features • sustains flexible use of features, with only occasional lapses • is easy to understand throughout; L1 accent has minimal effect on intelligibility
7	• speaks at length without noticeable effort or loss of coherence • may demonstrate language-related hesitation at times, or some repetition and/or self-correction • uses a range of connectives and discourse markers with some flexibility	• uses vocabulary resource flexibly to discuss a variety of topics • uses some less common and idiomatic vocabulary and shows some awareness of style and collocation, with some inappropriate choices • uses paraphrase effectively	• uses a range of complex structures with some flexibility • frequently produces error-free sentences, though some grammatical mistakes persist	• shows all the positive features of Band 6 and some, but not all, of the positive features of Band 8
6	• is willing to speak at length, though may lose coherence at times due to occasional repetition, self-correction or hesitation • uses a range of connectives and discourse markers but not always appropriately	• has a wide enough vocabulary to discuss topics at length and make meaning clear in spite of inappropriacies • generally paraphrases successfully	• uses a mix of simple and complex structures, but with limited flexibility • may make frequent mistakes with complex structures, though these rarely cause comprehension problems	• uses a range of pronunciation features with mixed control • shows some effective use of features but this is not sustained • can generally be understood throughout, though mispronunciation of individual words or sounds reduces clarity at times
5	• usually maintains flow of speech but uses repetition, self-correction and/or slow speech to keep going • may over-use certain connectives and discourse markers • produces simple speech fluently, but more complex communication causes fluency problems	• manages to talk about familiar and unfamiliar topics but uses vocabulary with limited flexibility • attempts to use paraphrase but with mixed success	• produces basic sentence forms with reasonable accuracy • uses a limited range of more complex structures, but these usually contain errors and may cause some comprehension problems	• shows all the positive features of Band 4 and some, but not all, of the positive features of Band 6
4	• cannot respond without noticeable pauses and may speak slowly, with frequent repetition and self-correction • links basic sentences but with repetitious use of simple connectives and some breakdowns in coherence	• is able to talk about familiar topics but can only convey basic meaning on unfamiliar topics and makes frequent errors in word choice • rarely attempts paraphrase	• produces basic sentence forms and some correct simple sentences but subordinate structures are rare • errors are frequent and may lead to misunderstanding	• uses a limited range of pronunciation features • attempts to control features but lapses are frequent • mispronunciations are frequent and cause some difficulty for the listener
3	• speaks with long pauses • has limited ability to link simple sentences • gives only simple responses and is frequently unable to convey basic message	• uses simple vocabulary to convey personal information • has insufficient vocabulary for less familiar topics	• attempts basic sentence forms but with limited success, or relies on apparently memorised utterances • makes numerous errors except in memorised expressions	• shows some of the features of Band 2 and some, but not all, of the positive features of Band 4
2	• pauses lengthily before most words • little communication possible	• only produces isolated words or memorised utterances	• cannot produce basic sentence forms	• speech is often unintelligible
1	• no communication possible • no rateable language			
0	• does not attend			

B Sujets utilisés pour CLES-JP et CLES-EN

B.1 Sujet sur l'intelligence artificielle générative

英語ディスカッション・役割分担シート

You will participate in a conversation with another candidate. You will be asked to defend a point of view, negotiate with your partner to reach a compromise according to your role below.

You are working as an English teacher at a university in Japan. You are discussing whether your department should allow your students to use generative AI for the English courses.

ROLE A

You believe that it is almost impossible to prohibit the use of AI, and that proper use of it should be included in the teaching contents.

ROLE B

You are worrying about potential risks of AI, and you believe that students should be discouraged from relying on it.

KEYWORDS

- Academic integrity / ethics
- Accuracy / potential bias
- Availability & accessibility (anytime, anywhere)
- Cheating
- Data privacy
- Environmental impact
- Learning efficiency
- Less fear for making mistakes
- Over-reliance
- Personalized learning
- Translation assistance

B.2 Sujet sur le travail en parallèle des étude

英語ディスカッション・役割分担シート
English discussion – Role play sheet

You will participate in a conversation with another candidate. You will be asked to defend a point of view, negotiate with your partner to reach a compromise according to your role below.

You are a member of the student council at your university. You are going to make a short video to inform the newly enterd students of advantages and disadvantages of part-time jobs. In the video, you are having a discussion with another member of the council by sharing your own experience.

Role A
You are working part-time yourself and you believe it has been mostly beneficial to you. Therefore, you are recommending other students to work part-time.

Role B
You are worrying about negative aspects of working part-time as a university student. You do not recommend other students to work part-time, and warn them to restrict the amount of work if they do.

Some keywords if you need:

- Academic overload, fatigue, stress, burnout
- Aligning work with academic goals
- Career relevance of part-time jobs
- Budgeting skills
- Building a professional network
- Enhanced resume/CV
- Employer expectations
- Employer flexibility and flexible work schedules
- Financial independence
- Gaining industry insights
- Internships vs. part-time jobs
- On-campus vs. off-campus jobs
- Limited time for studies and for extracurricular activities
- Potential impact on grades
- Real-world experience
- Reduced student debt
- Remote work opportunities
- Skill development

C Comparaison des systèmes d'ASR

Avant la sortie du système de reconnaissance Whisper (Radford et al., 2022), que nous avons finalement choisi d'implémenter dans PLSPP, nous avons comparé les performances obtenues par différents systèmes de reconnaissance automatique de la parole (ASR), sur des extraits d'enregistrements issus du corpus CLES-FR.

Dix-sept extraits de parole ont été extraits manuellement des premiers enregistrements effectués dans le cadre de la conception du corpus. Ces extraits sont d'une durée relativement courte (de 15 à 45 s chacun). Nous les avons transcrits manuellement, puis automatiquement à l'aide de six systèmes d'ASR :

- Google Speech Cloud API[8] (v2.29),
- EML Transcription[9] (v1.19),
- Amberscript[10] (v1.3),
- Fraunhofer Speech Recognition[11] (v2.13),
- Radboud University LST[12] (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[13] et l'autre sur LibriSpeech[14]. Ainsi, 10 *settings* différents ont été comparés à la transcription manuelle des 17 extraits de parole. Le calcul du WER est effectué à l'aide de la librairie Python Jiwer v2.5[15].

Les résultats obtenus sont présentés dans le tableau 11.1. On peut constater que les meilleures performances sont obtenues par le système Amberscript, avec toutefois un WER moyen de 25 %. Cependant, c'est également le système le plus cher de tous (10 € par heure de transcription en 2022). Le deuxième système le plus performant, Fraunhoher Speech Recognition, obtient un WER moyen de 45 %, ce qui est considérablement plus élevé. Nous considérons son utilisation avant la parution de Whisper et de son WER moyen autour de 19 %.

[8] Google Speech Cloud (2022) : https://cloud.google.com/speech-to-text/
[9] EML Transcription (2022) : https://www.eml.org/
[10] Amberscript (2022) : https://www.amberscript.com/
[11] Fraunhofer Speech Recognition (2022) : https://www.idmt.fraunhofer.de
[12] LST (2022) : https://webservices.cls.ru.nl/
[13] https://huggingface.co/speechbrain/asr-wav2vec2-commonvoice-en (consulté avril 2022)
[14] https://huggingface.co/speechbrain/asr-crdnn-rnnlm-librispeech (consulté avril 2022)
[15] https://github.com/jitsi/jiwer

ASR_SYSTEM	MEAN	SD	MIN	MAX	File1	File2	File3	File4	File5	File6	File7	File8	File9	File10	File11	File12	File13	File14	File15	File16	File17
ref_manual	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
amberscript_gb	0,25	0,08	0,14	0,44	0,19	0,15	0,16	0,33	0,32	0,25	0,25	0,23	0,14	0,44	0,22	0,26	0,27	0,31	0,34	0,14	0,29
fraunhofer_us	0,45	0,11	0,23	0,63	0,49	0,23	0,31	0,48	0,39	0,45	0,56	0,43	0,34	0,60	0,37	0,58	0,50	0,49	0,63	0,41	0,44
google_us	0,50	0,12	0,32	0,80	0,53	0,50	0,80	0,67	0,32	0,36	0,42	0,42	0,54	0,60	0,49	0,42	0,59	0,40	0,60	0,41	0,41
google_gb	0,55	0,15	0,34	0,96	0,64	0,39	0,67	0,55	0,61	0,36	0,61	0,65	0,56	0,96	0,39	0,46	0,46	0,54	0,60	0,34	0,48
lstenglish_us	0,66	0,11	0,42	0,88	0,57	0,65	0,78	0,88	0,71	0,77	0,78	0,61	0,63	0,64	0,66	0,58	0,71	0,42	0,60	0,71	0,59
lstenglish_gb	0,67	0,11	0,42	0,88	0,55	0,65	0,78	0,88	0,71	0,77	0,78	0,61	0,63	0,64	0,66	0,60	0,73	0,42	0,60	0,71	0,59
eml_gb	0,72	0,10	0,47	0,92	0,70	0,59	0,92	0,73	0,68	0,73	0,89	0,67	0,60	0,80	0,47	0,77	0,71	0,73	0,77	0,71	0,70
eml_us	0,72	0,10	0,47	0,89	0,74	0,61	0,78	0,73	0,79	0,75	0,89	0,67	0,63	0,88	0,47	0,77	0,75	0,72	0,67	0,71	0,67
speechbrain_wav2vec	0,73	0,26	0,28	1,00	0,28	1,00	0,29	0,77	0,36	0,34	0,81	0,64	0,78	0,96	0,76	0,95	0,73	1,00	0,91	0,83	1,00
speechbrain_crdnn	0,75	0,13	0,47	0,99	0,94	0,68	0,82	0,83	0,64	0,68	0,47	0,59	0,66	0,88	0,68	0,74	0,77	0,83	0,99	0,88	0,68

Tab. 11.1 : *Taux d'erreur de mots (WER) obtenus par 10 settings différents de reconnaissance automatique de parole sur 17 extraits du corpus CLES-FR*

D Penn Treebank II Constituent Tags

Source : https://surdeanu.cs.arizona.edu//mihai/teaching/ista555-fall13/readings/PennTreebankConstituents.html (consultée le 3 novembre 2024)

D.1 Clause Level

- S - simple declarative clause, i.e. one that is not introduced by a (possible empty) subordinating conjunction or a wh-word and that does not exhibit subject-verb inversion.
- SBAR - Clause introduced by a (possibly empty) subordinating conjunction.
- SBARQ - Direct question introduced by a wh-word or a wh-phrase. Indirect questions and relative clauses should be bracketed as SBAR, not SBARQ.
- SINV - Inverted declarative sentence, i.e. one in which the subject follows the tensed verb or modal.
- SQ - Inverted yes/no question, or main clause of a wh-question, following the wh-phrase in SBARQ.

D.2 Phrase Level

- ADJP - Adjective Phrase.
- ADVP - Adverb Phrase.
- CONJP - Conjunction Phrase.
- FRAG - Fragment.
- INTJ - Interjection. Corresponds approximately to the part-of-speech tag UH.
- LST - List marker. Includes surrounding punctuation.
- NAC - Not a Constituent ; used to show the scope of certain prenominal modifiers within an NP.
- NP - Noun Phrase.
- NX - Used within certain complex NPs to mark the head of the NP. Corresponds very roughly to N-bar level but used quite differently.
- PP - Prepositional Phrase.
- PRN - Parenthetical.

- PRT - Particle. Category for words that should be tagged RP.
- QP - Quantifier Phrase (i.e. complex measure/amount phrase) ; used within NP.
- RRC - Reduced Relative Clause.
- UCP - Unlike Coordinated Phrase.
- VP - Vereb Phrase.
- WHADJP - Wh-adjective Phrase. Adjectival phrase containing a wh-adverb, as in how hot.
- WHAVP - Wh-adverb Phrase. Introduces a clause with an NP gap. May be null (containing the 0 complementizer) or lexical, containing a wh-adverb such as how or why.
- WHNP - Wh-noun Phrase. Introduces a clause with an NP gap. May be null (containing the 0 complementizer) or lexical, containing some wh-word, e.g. who, which book, whose daughter, none of which, or how many leopards.
- WHPP - Wh-prepositional Phrase. Prepositional phrase containing a wh-noun phrase (such as of which or by whose authority) that either introduces a PP gap or is contained by a WHNP.
- X - Unknown, uncertain, or unbracketable. X is often used for bracketing typos and in bracketing the...the-constructions.

D.3 Word level

- CC - Coordinating conjunction
- CD - Cardinal number
- DT - Determiner
- EX - Existential there
- FW - Foreign word
- IN - Preposition or subordinating conjunction
- JJ - Adjective
- JJR - Adjective, comparative
- JJS - Adjective, superlative
- LS - List item marker
- MD - Modal

- NN - Noun, singular or mass
- NNS - Noun, plural
- NNP - Proper noun, singular
- NNPS - Proper noun, plural
- PDT - Predeterminer
- POS - Possessive ending
- PRP - Personal pronoun
- PRP$ - Possessive pronoun (prolog version PRP-S)
- RB - Adverb
- RBR - Adverb, comparative
- RBS - Adverb, superlative
- RP - Particle
- SYM - Symbol
- TO - to
- UH - Interjection
- VB - Verb, base form
- VBD - Verb, past tense
- VBG - Verb, gerund or present participle
- VBN - Verb, past participle
- VBP - Verb, non-3rd person singular present
- VBZ - Verb, 3rd person singular present
- WDT - Wh-determiner
- WP - Wh-pronoun
- WP$ - Possessive wh-pronoun (prolog version WP-S)
- WRB - Wh-adverb

E Indice d'interférence par locuteur sur le corpus Gold

	Speaker	I_L	Duration (s)
1	dec2022-002_002-003_SPEAKER_00	0,88 %	277
2	dec2022-002_002-003_SPEAKER_01	0,50 %	228
3	dec2022-002_032-044_SPEAKER_00	0,48 %	294
4	dec2022-002_032-044_SPEAKER_01	0,20 %	256
5	dec2022-002_038-016_SPEAKER_00	1,86 %	288
6	dec2022-002_038-016_SPEAKER_01	2,66 %	201
7	dec2022-003_005-025_SPEAKER_00	1,65 %	223
8	dec2022-003_005-025_SPEAKER_01	2,72 %	271
9	dec2022-003_014-006_SPEAKER_00	0,59 %	276
10	dec2022-003_014-006_SPEAKER_01	1,16 %	314
11	dec2022-003_017-029_SPEAKER_01	4,43 %	191
12	dec2022-003_017-029_SPEAKER_02	8,56 %	176
13	dec2022-003_022-010_SPEAKER_00	2,00 %	165
14	dec2022-003_022-010_SPEAKER_01	2,21 %	100
15	dec2022-003_035-026_SPEAKER_01	2,75 %	395
16	dec2022-003_035-026_SPEAKER_02	6,33 %	106
17	dec2022-003_039-040_SPEAKER_00	1,12 %	306
18	dec2022-003_039-040_SPEAKER_01	1,71 %	394
19	dec2022-004_012-021_SPEAKER_00	3,05 %	209
20	dec2022-004_012-021_SPEAKER_01	1,27 %	279
21	dec2022-004_023-009_SPEAKER_00	2,89 %	215
22	dec2022-004_023-009_SPEAKER_01	2,23 %	269
23	dec2022-004_034-042_SPEAKER_00	5,59 %	235
24	dec2022-004_034-042_SPEAKER_01	4,40 %	217
25	dec2022-004_037-018_SPEAKER_00	1,13 %	128
26	dec2022-004_037-018_SPEAKER_01	1,49 %	266
27	dec2022-004_041-027_SPEAKER_00	0,71 %	228
28	dec2022-004_041-027_SPEAKER_01	0,13 %	129
29	dec2022-005_036-047_SPEAKER_00	1,93 %	280
30	dec2022-005_036-047_SPEAKER_01	0,12 %	225
31	dec2022-202_129-077_SPEAKER_00	1,09 %	266
32	dec2022-202_129-077_SPEAKER_01	0,06 %	176
33	dec2022-203_093-115_SPEAKER_00	4,27 %	73
34	dec2022-203_093-115_SPEAKER_01	0,27 %	64
35	jan2023-201_141-155_SPEAKER_00	1,14 %	223
36	jan2023-201_141-155_SPEAKER_01	0,73 %	205
37	jan2023-401_120-034_SPEAKER_00	29,26 %	232
38	jan2023-401_120-034_SPEAKER_01	15,36 %	345
39	jan2023-402_146-048_SPEAKER_00	0,26 %	268
40	jan2023-402_146-048_SPEAKER_01	0,00 %	83

Tab. 11.2 : *Indice d'interférence par locuteur et durée totale de parole (segments de durée supérieure ou égale à* 8 s*)*

F Taux d'erreur de mots sur le corpus Gold

	Speaker	WER	SR	DR	IR	Nb of Words
1	dec2022-002_002-003_SPEAKER_00	11	4,55	2,48	4,14	483
2	dec2022-002_002-003_SPEAKER_01	10	4,74	2,91	2,73	549
3	dec2022-002_032-044_SPEAKER_00	15	6,91	2,56	5,37	391
4	dec2022-002_032-044_SPEAKER_01	11	5,51	2,90	2,90	345
5	dec2022-002_038-016_SPEAKER_00	11	6,27	1,76	3,33	510
6	dec2022-002_038-016_SPEAKER_01	10	3,56	4,35	2,37	253
7	dec2022-003_005-025_SPEAKER_00	13	3,74	6,98	2,49	401
8	dec2022-003_005-025_SPEAKER_01	19	3,32	4,34	11,48	392
9	dec2022-003_014-006_SPEAKER_00	6	1,35	2,03	2,36	592
10	dec2022-003_014-006_SPEAKER_01	8	3,21	2,92	2,34	685
11	dec2022-003_017-029_SPEAKER_01	19	5,84	3,11	9,73	257
12	dec2022-003_017-029_SPEAKER_02	20	7,10	2,78	9,88	324
13	dec2022-003_022-010_SPEAKER_00	20	5,15	12,02	3,00	233
14	dec2022-003_022-010_SPEAKER_01	15	7,98	3,07	3,68	163
15	dec2022-003_035-026_SPEAKER_01	6	4,10	1,28	0,77	781
16	dec2022-003_035-026_SPEAKER_02	21	6,94	5,56	8,33	216
17	dec2022-003_039-040_SPEAKER_00	14	3,96	3,79	6,54	581
18	dec2022-003_039-040_SPEAKER_01	6	1,24	3,19	1,44	971
19	dec2022-004_012-021_SPEAKER_00	7	2,56	2,96	0,99	507
20	dec2022-004_012-021_SPEAKER_01	15	4,97	5,33	4,97	563
21	dec2022-004_023-009_SPEAKER_00	15	7,50	3,61	3,61	360
22	dec2022-004_023-009_SPEAKER_01	15	5,86	3,70	4,94	324
23	dec2022-004_034-042_SPEAKER_00	6	1,81	3,62	0,36	276
24	dec2022-004_034-042_SPEAKER_01	11	4,10	4,10	2,52	317
25	dec2022-004_037-018_SPEAKER_00	6	2,15	1,72	2,58	233
26	dec2022-004_037-018_SPEAKER_01	11	5,37	3,74	2,10	428
27	dec2022-004_041-027_SPEAKER_00	20	10,29	3,22	6,75	311
28	dec2022-004_041-027_SPEAKER_01	7	3,47	0,99	2,97	202
29	dec2022-005_036-047_SPEAKER_00	7	4,42	1,52	1,37	656
30	dec2022-005_036-047_SPEAKER_01	9	4,99	2,49	1,59	441
31	dec2022-202_129-077_SPEAKER_00	11	6,38	2,26	2,67	486
32	dec2022-202_129-077_SPEAKER_01	27	8,53	5,81	12,40	258
33	dec2022-203_093-115_SPEAKER_00	45	29,59	8,16	7,14	98
34	dec2022-203_093-115_SPEAKER_01	45	9,78	7,61	27,17	92
35	jan2023-201_141-155_SPEAKER_00	28	12,61	6,08	9,23	444
36	jan2023-201_141-155_SPEAKER_01	17	9,46	5,68	1,62	370
37	jan2023-401_120-034_SPEAKER_00	63	22,85	3,76	36,29	372
38	jan2023-401_120-034_SPEAKER_01	45	26,32	8,30	10,32	494
39	jan2023-402_146-048_SPEAKER_00	16	7,22	5,35	3,48	374
40	jan2023-402_146-048_SPEAKER_01	13	9,62	1,92	1,92	156
	Mean :	**16,85**	**7,13**	**4,00**	**5,75**	

TAB. 11.3 : Taux d'erreur de mots (WER), de substitutions (SR), de délétions (DR), et d'insertions (IR), et nombre total de mots par locuteur

G Captures d'écran de Dynamic Rater

L'application web Dynamic Rater[16] a été développée pour les besoins de cette étude. Elle se compose de 4 vues principales : une page d'accueil avec la présentation du déroulement de l'expérimentation, une page de questionnaire linguistique, la page d'expérimentation, et la page de fin d'expérimentation.

La page d'accueil a deux objectifs : vérifier que l'utilisateur est correctement identifié avant de commencer l'expérience, et lui expliquer le contexte et le déroulé de celle-ci. Les participants accèdent à Dynamic Rater directement depuis la plateforme Prolific. Lorsqu'ils arrivent sur la page d'accueil, Prolific envoie un token d'identification qui permet de faire le lien avec leur profil Prolific, de les contacter si besoin et de leur attribuer la rétribution financière. Si aucun token n'est détecté, il leur est possible de le saisir manuellement. La description de l'expérience présente dans les grandes lignes ce qu'ils vont devoir faire, les contraintes qu'ils auront, et le temps imparti. Les conditions techniques nécessaires sont également indiquées. Le démarrage de l'expérience est conditionné à la bonne identification du participant. La figure 11.4 montre à quoi ressemble l'écran d'accueil lorsque le participant y arrive depuis la plateforme Prolific.

Le rôle du questionnaire linguistique est de demander directement aux participants d'indiquer leur(s) langue(s) maternelle(s), leur pays de résidence, et les langues étrangères qu'ils ont apprises, pendant combien de temps et dans quels contextes. Ces informations sont déjà données par Prolific, mais poser les questions ici permet d'obtenir des réponses plus à jour et précises, notamment pour les langues apprises. La page du questionnaire est visible figure 11.5.

La phase d'entraînement, figure 11.6, est en tout point identique à celle de l'expérimentation réelle, à la différence qu'il est mentionné qu'il s'agit d'un entraînement, et que les résultats ne sont pas analysés. Un segment audio ne présentant pas de spécificité particulière a été sélectionné pour cette phase. À la fin de la lecture audio s'affichent les curseurs d'évaluation globale, comme pour les stimuli de l'expérience réelle (cf. figure 11.7).

Une fois les 16 segments présentés aléatoirement, une écran de fin d'expérience s'affiche pour remercier le participant et lui laisser la possibilité d'écrire un commentaire global s'il le souhaite (cf. figure 11.8). En cliquant sur le bouton *validate the survey*, il est redirigé vers Prolific, qui est alors informé de la fin de passation.

[16] Code source : https://gricad-gitlab.univ-grenoble-alpes.fr/lidilem/dynamic-rater

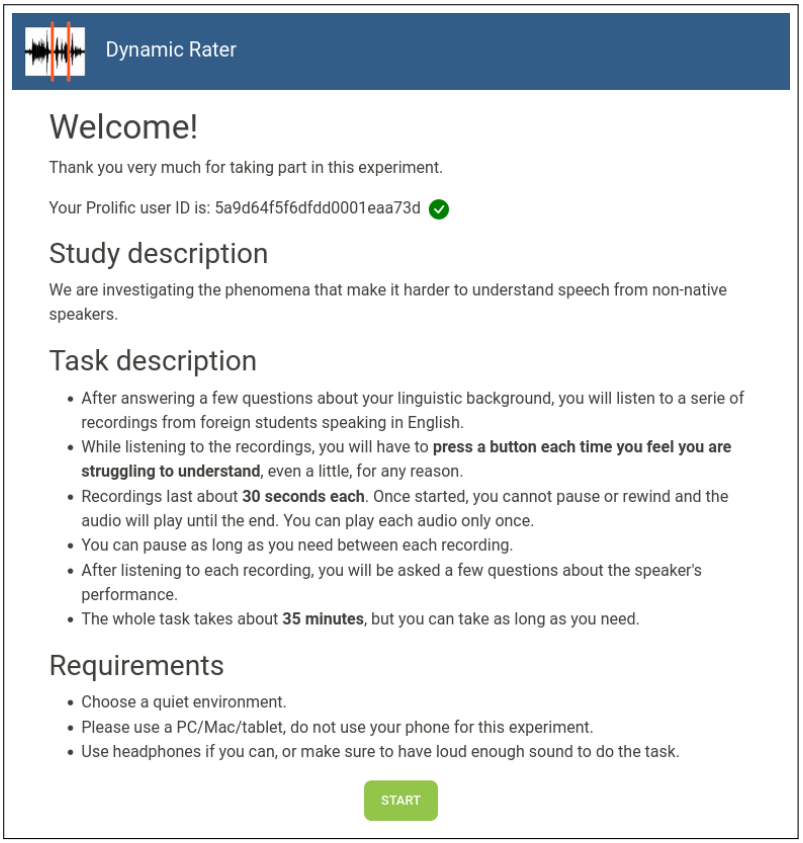


FIG. 11.4 : Page d'accueil de Dynamic Rater

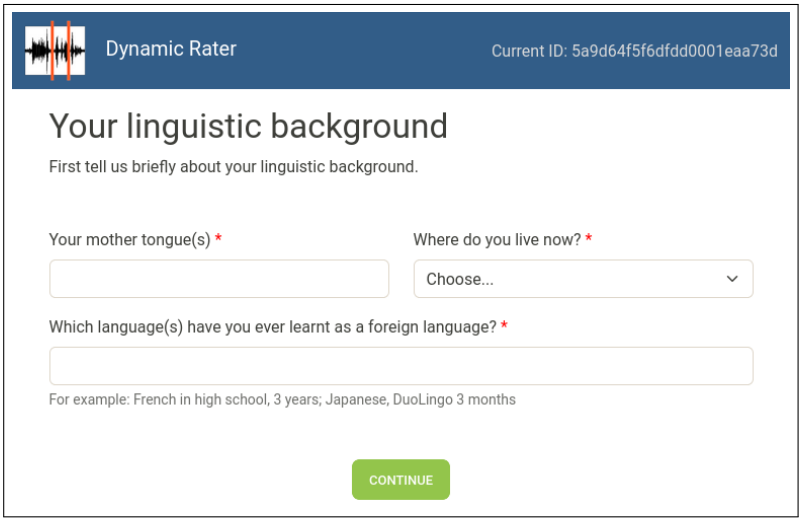


FIG. 11.5 : Questionnaire linguistique

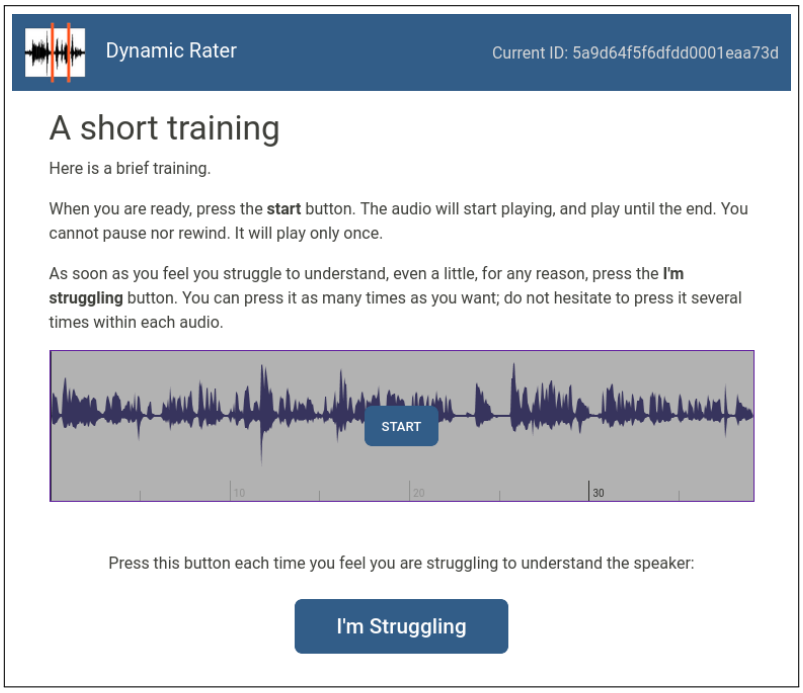


FIG. 11.6 : Phase d'entraînement

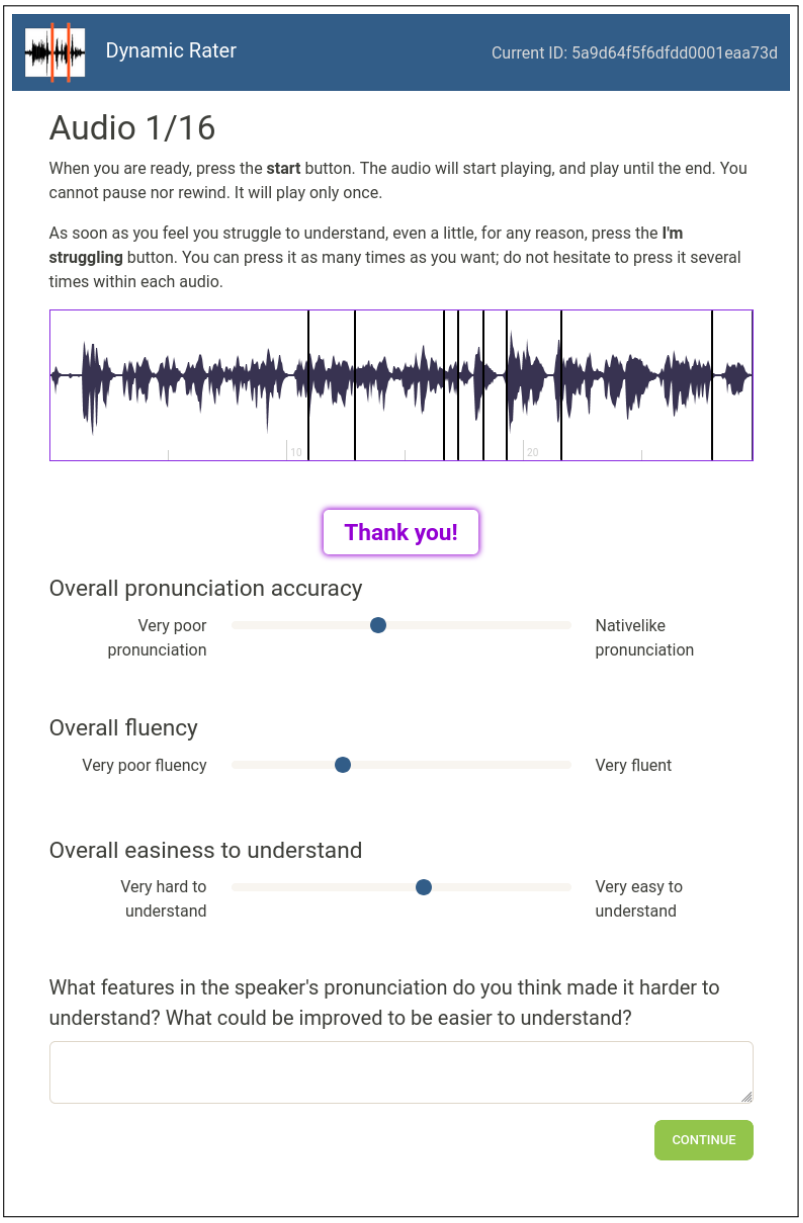


FIG. 11.7 : Évaluation globale à la suite de l'évaluation dynamique d'un enregistrement

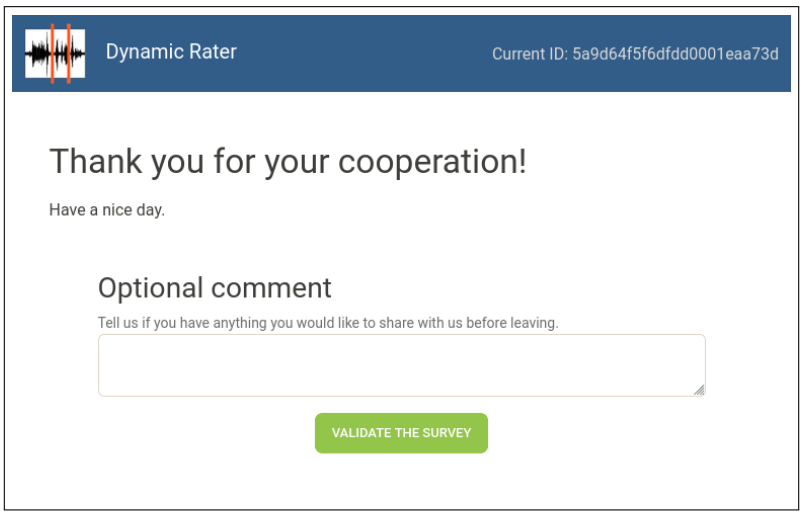


FIG. 11.8 : Écran de fin d'expérience

H Communications & publications

H.1 Communications directement en lien avec cette thèse

- Nakanishi, N., **Coulange, S.** (2025). Beyond Intuition : Identifying Key Factors Affecting L2 Speech Comprehensibility. The 11th International Symposium on the Acquisition of Second Language Speech, Apr. 23-25, Toronto, Canada. [ABSTRACT]

- **Coulange, S.**, Kato, T., Rossato, S., Masperi, M. (2024). Exploring Impact of Pausing and Lexical Stress Patterns on L2 English Comprehensibility in Real Time. Proceedings of Interspeech 2024, Sep. 1-5, Kos, Greece. [PAPER]

- Nakanishi, M., **Coulange, S.** (2024). Measuring speech rhythm through automated analysis of syllabic prominences. "Prosodic features of language learners' fluency" Satellite Workshop of Speech Prosody, Jul. 1, Leiden, Netherlands. [ABSTRACT]

- Sugahara, M., **Coulange, S.**, Kato, T. (2024). English Lexical Stress in Awareness and Production : Native and Non-native Speakers. The 19th Conference on Laboratory Phonology, Jun. 27-29, Seoul, Korea. [ABSTRACT]

- **Coulange, S.**, Fries, M.-H., Masperi, M., Rossato, R. (2024). A corpus of spontaneous L2 English speech for real-situation speaking assessment. Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), May 20-25, Torino, Italy. [PAPER] [POSTER]

- **Coulange, S.**, Kato, T., Rossato, R., Masperi, M. (2024). An automated pipeline for preprocessing spontaneous L2 English prosody. 13th International Seminar on Speech Production (ISSP 2024), May 13-17, Autrans, France. [ABSTRACT] [POSTER]

- **Coulange, S.**, Kato, T., Rossato, R., Masperi, M. (2024). Dynamic Approach to Comprehensibility Assessment in Foreign Language Pronunciation Training. 8th International Conference on English Pronunciation : Issues and Practices, May 8-10, Santander, Spain. [ABSTRACT]

- **Coulange, S.**, Kato, T., Rossato, R., Masperi, M. (2024). Enhancing Language Learners' Comprehensibility through Automated Analysis of Pause Positions and Syllable Prominence. Languages 9, no. 3 : 78. [PAPER]

- Kimura, T., **Coulange, S.**, Kato, T. (2024). 日本人小学生による英語暗唱音声における語彙強勢位置の自動推定と母語話者評価 [Automatic estimation and native speakers' evaluation of lexical stress positions in English recitation speech produced by Japanese elementary school children]. 日本音響学会第 151 回研究発表会 [Spring Meeting of the Acoustic Society of Japan], Mar. 6-8, Tokyo, Japan. [PAPER]

- **Coulange, S.**, Konishi, T., Kato, T., Sugahara, M., Rossato, R., Masperi, M. (2024). A corpus of spontaneous dialogues in L2 English by French and Japanese L1 speakers for automated assessment of fluency. 6th International Symposium on Learner Corpus Studies in Asia and the World (LCSAW6), Feb. 3, Kobe, Japan. [POSTER]

- **Coulange, S.**, Kato, T., Rossato, R., Masperi, M. (2023). Comprehensibility diagnosis of spontaneous L2 English : Automated analysis of pausing and lexical stress patterns. Invited presentation for the workshop "Tools in L2 research", Nov. 24-25, Zurich, Switzerland. [SLIDES]

- **Coulange, S.**, Kato, T., Rossato, R., Masperi, M. (2023). フランス人学習者による自発 L2 英語発話における語彙アクセント自動測定 [Automatic Measurement of Lexical Stress in Spontaneous L2 English Speech of French Learners]. 第 37 回日本音声学会全国大会 [The 37th General Meeting of the Phonetic Society of Japan], Sep. 16-17, Sapporo, Japan. pp. 126-131 [PAPER]

- **Coulange, S.**, Kato, T. (2023). Pause position analysis in spontaneous speech for L2 English fluency assessment. 日本音響学会第 150 回研究発表会 [Autumn Meeting of the Acoustic Society of Japan], Sep. 26-28, Nagoya, Japan. pp. 991-994 [PAPER] [POSTER]

- **Coulange, S.** (2023). Computer-aided pronunciation training in 2022 : When pedagogy struggles to catch up. Proc. 7th International Conference on English Pronunciation : Issues and Practices (EPIP7), May 18-20, Grenoble, France. [PAPER]

- **Coulange, S.** (2023). Conception d'un module de diagnostic automatique de la prononciation. In Soubre, V. (eds.) L'évaluation en tant que soutien d'apprentissage, Synergies France n°17, Gerflint. [PAPER]

- **Coulange, S.** (2022). Conception d'un système automatique de diagnostic de la prononciation en anglais spontané L2. Séminaire LIDILEM Didactique de l'Oral en Langue Étrangère, Nov., Grenoble.

- Frost, D., **Coulange, S.** (2022). Évaluations qualitative et quantitative de la prononciation. Séminaire LIDILEM Didactique de l'Oral en Langue Étrangère, Jun., Grenoble.

- **Coulange, S.** (2022). Perspectives de diagnostic automatique de la prononciation en anglais pour SELF. Webinaire "Language assessment at the crossroads", LIDILEM ILCEA4, May, Grenoble.

H.2 Communications indirectement en lien avec cette thèse

La liste suivante présente un certain nombre de communications ayant utilisé une version de PLSPP, ou dans laquelle nous avons adapté une méthodologie d'analyse ou de mesure mise au point dans cette thèse.

- Erickson, D., Raso, T., Lundmark, M. S., Frid, J., **Coulange, S.** (submitted). The Many Colors of Prominence : A Pilot Study of Topic Prosodic Forms. Journal of Speech Sciences.

- Nishioka, M., Kato, T., Sugahara, M., **Coulange, S.** (2025). 日本人小学生による英語復唱音声における語彙強勢の分析 [Analysis of Lexical Stress in English Repetition Speech Produced by Japanese Primary School Children]. 日本音響学会第 153 回研究発表会 [Spring Meeting of the Acoustic Society of Japan], Mar. 17-19, Saitama, Japan.

- Frost, D., Skarnitzl, R., **Coulange, S.**, Hosseini, H. (2024). Perceived ease of understanding in French-accented academic discourse : and the chief culprits are...? The 17th International Conference on Native and Non-Native Accents of English, Dec. 12-14, Łódź, Poland.

- Raso, T., Erickson, D., **Coulange, S.**, Lundmark, M. S., Frid, J. (2024). Acoustic, articulatory and perceptual characteristics of Topic Prosodic Forms in English utterances. Seminário Internacional de Fonologia, Nov. 6-7, Rio de Janeiro, Brasil.

- Sugahara, M., **Coulange, S.**, Kato, T. (2023). 意識されている強勢 vs. 発話における強勢 — 日本人と韓国人の大学生による英単語への主強勢付与の比較 [Stress awareness vs. stress production : Comparison of primary stress assignment to English words between Japanese and Korean university students]. 第 347 回日本音声学会研究例会 [The 347th regular meeting of the Phonetic Society of Japan], Nov. 25, online.

Résumés

A Résumé français

La compétence de production orale en langue étrangère est communément évaluée à travers la capacité du locuteur à se faire comprendre. Le terme de compréhensibilité est souvent utilisé dans le domaine de l'acquisition des langues pour désigner le degré d'effort requis par un auditeur pour comprendre le locuteur. Celle-ci dépend de nombreux facteurs, côté locuteur comme auditeur, parmi lesquels la position des marqueurs d'hésitation et le rythme de la parole sont souvent décrits comme facteurs clés dans les grilles d'évaluation. En revanche, les approches automatiques reposent encore majoritairement sur des comparaisons à un modèle, souvent à partir de parole lue, sans porter une attention particulière aux phénomènes linguistiques susceptibles de perturber la compréhension.

Dans cette thèse, nous proposons un outil d'évaluation automatique de la production orale spontanée en anglais, ciblant spécifiquement deux phénomènes linguistiques susceptibles d'impacter la compréhensibilité du locuteur : la distribution syntaxique des pauses et l'accentuation lexicale. Cet outil a été utilisé pour analyser la variation de ces phénomènes chez des locuteurs de niveau B1 et B2 francophones et japonophones (L2), ainsi que chez des locuteurs anglophones natifs (L1). Nous avons ensuite mesuré leur impact sur la perception de l'effort de compréhension en temps réel chez des auditeurs anglophones natifs.

Les analyses révèlent qu'une différence majeure entre les locuteurs B1 et B2 d'une part, et entre les L1 et L2 d'autre part, réside dans la fréquence des pauses de bas niveau syntaxique : plus le niveau de compétence augmente, plus les pauses tendent à se concentrer aux frontières syntaxiques de haut niveau (par exemple, entre les propositions). À partir de ces observations, nous avons défini un score de distribution syntaxique des pauses pour caractériser cette concentration.

Concernant l'accentuation lexicale, les mesures font ressortir une influence marquée des patterns accentuels de la langue maternelle des locuteurs. Les francophones

augmentent généralement la hauteur (f_0) et allongent la dernière syllabe des mots, tandis que l'intensité reste stable sur l'ensemble des syllabes. Bien que les locuteurs B1 et B2 se chevauchent largement, une progression significative est observée entre ces deux niveaux : les locuteurs B2 accentuent plus fréquemment la syllabe attendue et produisent un contraste acoustique plus marqué en termes de f_0 et d'intensité. En revanche, les locuteurs japonais, dont la langue maternelle intègre un accent lexical, positionnent l'accent avec davantage de précision et produisent des contrastes prosodiques plus importants que les francophones.

Enfin, pour mesurer l'impact des différents types de pauses et de patterns d'accentuation lexicale sur la perception de la compréhensibilité, nous avons recruté 60 auditeurs afin d'évaluer dynamiquement 16 extraits de paroles issus du corpus des locuteurs francophones mentionné ci-dessus. Lors de l'écoute, il leur était demandé de cliquer sur un bouton dès qu'ils ressentaient un effort pour comprendre le locuteur, quelle qu'en soit la raison. L'analyse des patterns de clics, une fois normalisés, indique une augmentation significative de la difficulté perçue entre 0 et 3 secondes après une pause située à l'intérieur d'un syntagme, ainsi qu'entre 2 et 3 secondes après un mot dont la syllabe accentuée n'était pas celle attendue. À l'inverse, la difficulté perçue diminue significativement après une pause située entre deux propositions ou après un mot correctement accentué. Ces observations renforcent l'idée que les pauses de bas niveau syntaxique et les mots accentués de manière inattendue impactent directement la perception de difficulté de compréhension. Notre expérience démontre que cet impact est direct et mesurable.

B English abstract

Automated Assessment of Spontaneous Speech in English as a Foreign Language : the Role of Pauses and Lexical Stress in Speaker Comprehensibility

Oral production proficiency in a foreign language is commonly assessed through the speaker's ability to be understood. The term comprehensibility is frequently used in the field of second language acquisition to refer to the level of effort required by a listener to understand the speaker. Comprehensibility depends on numerous factors, both speaker- and listener-related, among which the position of hesitation markers and speech rhythm are often highlighted as key factors in evaluation scales. In contrast, automatic assessment approaches still predominantly rely on comparisons with a model, often based on read speech, without specific attention to linguistic phenomena likely to disrupt comprehension.

In this thesis, we propose a tool for the automatic evaluation of spontaneous oral production in English, specifically targeting two linguistic phenomena likely to

impact speaker comprehensibility : the syntactic distribution of pauses and lexical stress. This tool was used to analyze variation of these phenomena among French- and Japanese-speaking learners of English at B1 and B2 proficiency levels (L2), as well as among native English speakers (L1). We then measured their impact on the perception of real-time comprehension effort in native English-speaking listeners.

The analyses reveal that a major difference between B1 and B2 speakers, as well as between L1 and L2 speakers, lies in the frequency of low-syntactic-level pauses : as proficiency increases, pauses tend to concentrate more at high-syntactic-boundary locations (e.g., between clauses). Based on these observations, a syntactic pause distribution score was developed to quantify this concentration.

Regarding lexical stress, the findings show a strong influence of speakers' native language stress patterns. French speakers typically increase pitch (f_0) and lengthen the final syllable of words, while intensity remains stable across syllables. Although B1 and B2 speakers largely overlap, significant improvement is observed between these levels : B2 speakers more frequently stress the expected syllable and produce stronger acoustic contrasts in terms of f_0 and intensity. On the other hand, Japanese speakers, whose native language incorporates lexical stress, position stress more accurately and produce greater prosodic contrasts than French speakers.

Finally, to measure the impact of different types of pauses and lexical stress patterns on perceived comprehensibility, we recruited 60 listeners to dynamically evaluate 16 speech excerpts from the aforementioned corpus of French speakers. While listening, participants were asked to click a button whenever they experienced difficulty understanding the speaker, for whatever reason. An analysis of the normalized click patterns indicates a significant increase in perceived difficulty between 0 and 3 seconds following a pause within a phrase, and between 2 and 3 seconds after a word with an unexpected stressed syllable. Conversely, perceived difficulty significantly decreased after a pause between clauses or following a correctly stressed word. These findings support the idea that low-syntactic-level pauses and unexpected lexical stress directly impact the perception of comprehensibility. Our experiment shows that this impact is direct and measurable.