Sommaire

Introduction

Introduction générale...

partie I

Contexte théorique

Commentaires : Introduction de partie...

Chapitre 1

Évaluation de la prononciation

L'évaluation de la production orale touche de nombreux aspects de la langue, de la cohérence syntaxique à la précision du vocabulaire, en passant par la pragmatique ou la cohésion du discours. S'il y a pourtant un aspect qui influence globalement l'évaluation, c'est la qualité de la prononciation (Pennington, 1999). Par prononciation, nous entendons ici tous les aspects de la parole relatifs à la phonétique et à la prosodie.

1.1 Évaluation humaine

L'évaluation de la prononciation, lorsqu'elle est explicite dans les grilles d'évaluation, se résume souvent à un score global établi par l'enseignant sur la précision phonétique ou la fluence de parole (Piccardo, 2016). Toutefois, les enseignants ont rarement les outils et la formation nécessaires pour évaluer ces aspects de manière précise et systématique, et il en résulte souvent un jugement basé sur l'intuition, avec une certaine variabilité d'un évaluateur à l'autre ou chez un même évaluateur. Gilquin et al., 2022 notent par exemple que les enseignants non natifs ont tendance à être plus sévères que les enseignants natifs lorsqu'ils évaluent la prononciation, et qu'ils ne se basent pas sur les mêmes critères pour évaluer leurs étudiants. Par ailleurs, un évaluateur ne sera pas sensible de la même manière à la prononciation des apprenants selon son degré de familiarité avec leur accent (Didelot et al., 2019). De manière générale, un manque de formation des enseignants pour enseigner et évaluer la prononciation est clairement ressenti (Baker, 2011 ; Breitkreutz et al., 2001 ; Burgess & Spencer, 2000 ; Derwing & Munro, 2015 ; Frost & O'Donnell, 2018 ; Piccardo, 2016). Il est ainsi apparu nécessaire de développer des critères d'évaluation précis. "He's got something in his mouth" : expression d'un des profs participants à la mise au point de l'échelle de compréhensibilité d'Isaacs et al., 2018, qui ne trouve pas les mots pour décrire ce qui

se passe ; comme la plupart des profs participants : trop peu de connaissances sur la prononciation, la phonologie, la prosodie... (cf. podcast de Glenn Fulcher)

Dans cette section, nous présentons différentes grilles d'évaluation de la production orale en L2 destinées à des évaluateurs et/ou des enseignants, afin d'observer la place réservée au rythme et étudier les descripteurs de compétence par niveau qui lui sont associés.

1.1.1 Évaluation holistique

Le CECRL

Le Cadre Européen Commun de Référence pour les Langues (CECRL) est une initiative du Conseil de l'Europe de définir des descripteurs de compétence détaillés pour faciliter l'enseignement et l'évaluation des langues étrangères. La première édition de l'ouvrage (Conseil de l'Europe, 2001) proposait une échelle de "Maîtrise du système phonologique" qui a largement été critiquée pour son manque de précision (Piccardo, 2016). Celle-ci s'appuyait en effet sur l'intuition de l'évaluateur (ex. "prononciation et intonation claires et naturelles" au niveau B2, "prononciation clairement intelligible" en B1, "net accent étranger" en A2) et prenait pour modèle la prononciation d'un locuteur natif sans pour autant la définir.

Lors de la mise au point d'une nouvelle édition des descripteurs en 2018, ces limitations ont été reconnues par les concepteurs des nouveaux descripteurs (Piccardo, 2016) qui qualifiaient les descripteurs phonologiques d'"échelle la moins réussie" (p. 133) et de seule échelle avec une norme native. Les nouveaux descripteurs quant à eux abandonnent la comparaison au modèle natif et se focalisent sur l'intelligibilité comme base théorique principale du contrôle phonologique. On accepte maintenant un accent qui n'affecte pas la compréhension en C2, ainsi que l'influence d'autre langues connues par l'apprenant.

Des précisions sont données quant à la maîtrise des traits prosodiques (au niveau B2 : « 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 » (Conseil de l'Europe, 2018 : 142)), l'accent et le rythme sont également 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."

Les descripteurs du TOEFL

Le TOEFL est un test certificatif pour évaluer l'anglais langue étrangère et développé par l'entreprise 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.

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 par la suite par une combinaison d'évaluations automatiques et d'évaluateurs humains certifiés. De cette manière le TOEFL garantit la qualité de l'évaluation en permettant aux évaluateurs humains de se concentrer sur le contenu de la production sans être biaisé 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[1]).

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 s à 30 s, puis doit enregistrer sa réponse au microphone pendant 45 s à 60 s.

- **Question 1** "Independent Speaking Task" il y est demandé au candidat de 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 réponse. Exemple d'énoncé tiré d'une vidéo tutoriel (ETS.org[2]) : *"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."*

[1] https://www.ets.org/toefl/test-takers/ibt/scores.html, consulté le 22/07/2024
[2] https://www.ets.org/toefl/test-takers/ibt/about/content/speaking.html, consulté le 22/07/2024

- **Questions 2 à 3** "Integrated Speaking Tasks" combinent la production orale avec la compréhension orale et écrite :
  - **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 présenter 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.

Chaque question est évaluée sur quatre critères de manière holistique sur une échelle de 0 à 4. Les critères sont les suivants : **Delivery** clear and fluid speech, good pronunciation, natural pace, good intonation ; **Language use** use of grammar and vocabulary ; et **Topic development** how fully the candidates answer, how clearly they express their ideas, how they connect ideas.

Les points importants mis en avant sont : 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 d'évaluation de la production orale du TOEFL iBT est disponible en ligne[3]. Elle est divisée en deux rubriques : "Independent Speaking" et "Integrated Speaking", chacune décrivant 5 niveaux (de 0 à 4) sur 3 critères : Delivery, Language use et Topic Development. Le critère qui nous intéresse est Delivery. Ses descripteurs prennent bien en compte l'effort perçu par l'auditeur et l'intelligibilité du locuteur. Au niveau 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'intelligilité). Au niveau 4, le discours est généralement fluide avec des difficultés mineures qui n'affectent pas l'intelligibilité.

[3] https://www.ets.org/pdfs/toefl/toefl-ibt-speaking-rubrics.pdf

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"

Integrated Speaking Rubric (delivery) pratiquement mot pour mot les descripteurs de Independent 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 test TOEFL ITP Speaking test. 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 compteur de temps de préparation s'active, puis le compteur de l'enregistreur. Il est possible de faire une pause entre les questions. Seule la conversation de la question 4 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. Une version démo du TOEFL ITP Speaking test est disponible en ligne[4].

La grille d'évaluation de la production orale pour le TOEFL ITP est accessible en ligne[5]. Elle décrit 4 niveaux de compétences de A2 à C1. Comme pour les descripteurs du TOEFL iBT, il est fait mention du degré d'effort requis par l'auditeur et

[4] https://www.ets.org/toefl/itp/prepare.html
[5] https://www.ets.org/pdfs/toefl/toefl-itp-speaking-descriptors.pdf

d'intelligibilité de la parole.

- ==A2== : speak clearly enough to be uderstood, with some listener effort (familiar, everyday topics) ; Pronunciation and word stress errors are noticeable and highly influenced by the speaker's native language. Choppy speech, frequent pauses and false starts. Brief stretches of speech. - ==B1== : intelligible speech, although certain unfamiliar words are mispronounced ; 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 (familiar subjects ; as topics become more unfamiliar and/or more complex, errors are more common and cause listener effort) - ==B2== : mostly well-paced and fluent speech, however may hesitate at times ; use stress and intonation to convey meaning, though there may be some errors or native language influence (confortable on most topics) - ==C1== : express themselves fluently with very little effort or hesitation, speech is clear and well-paced ; use stress and intonation effectively (express themselves with precision on most topics)

=> Transition importante entre B1 et B2 (comme pour le cadre) : B2 = "well-paced and fluent", plus qu'une légère influence de la L1. On ne parle plus d'effort côté auditeur, alors que c'est le cas en B1.

Les descripteurs du test IELTS

Présentation générale du test

La version publique de la grille d'évaluation de la production orale du test IELTS est accessible en ligne[6]. Elle décrit 9 niveaux selon 4 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 (2 : pause longuement avant la plupart des mots, 3 : longues pauses, 4 : pauses notables), à partir de 5 : fluide dans les contextes simples. Répétitions et auto-corrections mentionnées de 4 à 9 (4 : répétition et auto-réparation fréquentes, 6 : répétitions, auto-corrections et hésitations occasionnelles, 9 : rares répétitions et auto-corrections, toute hésitation est liée au contenu plutôt qu'à la recherche de mots ou à la grammaire).

Les descripteurs de prononciation sont très limités et difficilement interprétables. Il est fait mention de "pronunciation features" et "mispronunciations" sans plus de détails. La perception côté auditeur apparaît brièvement (4 : "les erreurs de prononciation sont fréquente et occasionnent des difficultés chez l'auditeur", 9 : "compréhensible sans effort"), de même pour l'intelligibilité du locuteur (8 : "L'accent de

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

la L1 a un impact négligeable sur l'intelligibilité").

Il n'est fait nulle part mention d'accentuation lexicale, ni de paramètres prosodiques en général.

Les descripteurs du test TOEIC

TOEIC tests vs. TOEIC Bridge tests : les deux sont sur ordi.

TOEIC tests : measure a wider range of English-language proficiency for non-native speakers and place more emphasis on English communication in the workplace.

TOEIC Speaking Test Format : https ://www.ets.org/toeic/test-takers/about/speaking-writing.html 11 questions, environ 20 min. 2 read aloud (45 sec prep, 45 sec rép) 2 descriptions d'images (45 prep, 30 rép) 3 répondre à une question (3 sec de préparation, 15 ou 30 sec de réponse) 3 répondre à une question en utilisant les infos données (45 sec de lecture, 3 sec de préparation par question, 15 et 30 secondes de réponse). La dernière question (celle à 30 sec de réponse) est posée deux fois. 1 expression de son opinion (45 sec prép, 60 sec rép)

Grille générale : https ://www.ets.org/pdfs/toeic/toeic-speaking-writing-score-descriptors.pdf

TOEIC Speaking Proficiency Level Descriptors 8 niveaux Plus centrés sur la réussite de la tâche (capacité de répondre à la question et se faire comprendre) plutôt que description détaillée des problèmes potentiels. "consistent pronunciation, stress, and intonation difficulties" (4), "unclear pronunciation or inappropriate intonation or stress" (6), "long pauses and frequent hesitations" (4, 5) La bascule B2 en termes de prononciation, telle que décrite par le CECRL, semble se jouer au niveau 7 : "minor difficulties with pronunciation, intonation, or hesitation" (7) Pour la lecture à voix haute, on pourra noter la mention "difficult to understand" (niveaux 2, 3) puis "vary in intelligibility" (4), "generally intelligible" (5), "intelligible" (6), "highly intelligible" (7).

Grille détaillée TOEIC Speaking : https ://www.ets.org/pdfs/toeic/toeic-speaking-writing-examinee-handbook.pdf Scoring Guide for the Read aloud Task (p15) : 4 niveaux (0-3) 2 critères : Pronunciation 1 : Pronunciation may be intelligible at times, but significant other language influence interferes with appropriate delivery of the text 2 : Pronunciation is generally intelligible, though it includes some lapses and/or other language influence. 3 : Pronunciation is highly intelligible, though the response may include minor lapses and/or other language influence.

Intonation and Stress 1 : Use of emphases, pauses, and rising and falling pitch is

not appropriate, and the response includes significant other language influence. 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. 3 : Use of emphases, pauses, and rising and falling pitch is appropriate to the text

Scoring Guide for the Describe a Picture Task p17 : 4 levels, plus sur l'adéquation de la réponse en termes de contenu, choix du vocabulaire et des structures. Mention de l'intelligibilité du locuteur et de l'effort demandé à l'auditeur.

Scoring Guide for Respond to Questions (Market Survey) and Respond to Questions Using Information Provided (Agenda) Tasks p20 : idem.

Scoring Guide for the Express an Opinion Task p22 : 6 niveaux plus détaillés, couvre contenu, grammaire voc mais aussi prononciation et fluence Niveau 2, on peut noter notamment : "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." Niveau 3 : 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. Niveau 4 : Minor difficulties with pronunciation, intonation, or pacing are noticeable and may require listener effort at times, although overall intelligibility is not significantly affected. Niveau 5 : 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.

TOEIC Bridge Speaking and Writing tests : sur ordinateur, focus on basic to intermediate level English communication skills for everyday life. Test de production orale : évalue comp anglais sur tâches de communication orale en milieu privé, public et professionnel familier (familiar workplace)/ →ability to communicate immediate needs, provide basic info, interact on topics of perso interest with people who are speaking clearly. Les candidats répondent à des question simples sur des sujets familiers et utilisent des phrases pour décrire des événements de la vie quotidienne. Ils peuvent être amené à expliquer brièvement leur opinion ou leurs projets, et raconter des histoires simples. Les candidats doivent pouvoir : "pronounce words in a way that can be understood by proficient users of English, using intonation, stress, and pauses to pace speech and contribute to comprehensibility" Format du test TOEIC Bridge Speaking : 8 questions (environ 15 min) composé de tâches de lecture à voix haute, description d'image, listen and retell, interaction courte (voice mail après lecture d'un petit texte genre mémo), raconter une histoire (sur la base d'une suite d'images) et make and support a recommendation (sur la base d'un texte court et d'un contexte écrit (il s'agit de lister les options proposées dans le texte, en recommander une et expliquer pourquoi)). Temps de préparation : - 25 secondes pour les deux tâches de lecture, suivie de 30 secondes d'enregistrement. - deux tâches de description d'image :

30 sec prép, 30 sec rép. - Listen and retell : 15 sec prép, 30 sec rép. - short interaction : 30 sec prép, 30 sec rép. - raconter une histoire : 45 sec prép, 60 sec rép. - make and support a recommendation : 60 sec prép, 60 sec rép.

Critères d'évaluation : Globalement : se concentre sur le fullfilment de la réponse (offtopic ? lecture complète ? réponse adéquate ?), la pertinence du vocabulaire et des structures utilisées et l'intelligibilité et la compréhensibilité générale du locuteur sans donner plus de détails. Tâches de lecture (p13) : 4 niveaux (0 à 3) : ==1== : "Speech is mostly unintelligible and requires significant listener effort to understand", ==2== : "At the word level, pronunciation is sometimes unintelligible and requires some listener effort", "intonation and stress are somewhat appropriate", "lapses and/or other language influence are present", ==3== : "other-language influence does not affect overall intelligibility", "At the word level, pronunciation is mostly intelligible, but there may be some minor lapses", "At the phrase and sentence level, intonation and stress are mostly appropriate" Pas plus d'info niveau accentuation et fluence.

Tâches de description d'image (p15) : 4 niveaux (0 à 3) : "mostly unintelligible" (1), "sometimes unintelligible and requires listener effort" "errors interfere with comprehensibility" (2), "generally intelligible but may require some listener effort" "minor errors that do not affect meaning" (3) Pas plus d'info niveau accentuation et fluence.

Listen and Retell (p17) toujours 4 niv "mostly unintelligible and/or requires significant listener effort" (1) "sometimes unintelligible and/or sometimes requires listener effort" (2) "generally intelligible but may require some listener effort" (3)

Short interaction (p19) idem Raconter une histoire (p21) : ==5 niveaux==, similaire Make and support a recommendation : ==5 niveaux==, similaire

Cas du CLES

Le Certificat de Compétences en Langues de l'Enseignement Supérieur (CLES) est un test certificatif universitaire établi 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). Chaque niveau du CECRL y est évalué indépendamment : le candidat doit donc 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 compétences : compréhension écrite, compréhension orale, expression écrite et expression orale, en monologue pour le niveau B1, et en interaction pour les niveaux B2 et C1.

Sur le site ils disent "Le niveau B2 est comme une évolution du niveau B1" mais

on attend que le candidat soit "significativement plus fluide"[7]. Au niveau B2 est attendu que le candidat "S'exprime avec fluidité (contrairement au B1, le candidat ne doit pas faire de longues pauses)" et "le candidat doit être aisément compréhensible".

Au niveau B2, l'examen consiste en une interaction orale sous 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 (exemple : l'usage de la cigarette électronique, des caméras du surveillance ou les tests cliniques sur les animaux). 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 la limite de dix minutes. Ils sont évalués en direct par un ou deux évaluateurs certifiés CLES présents dans la salle.

L'évaluation est faite sur 8 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). Pour chacun des critères, l'évaluateur peut attribuer le niveau B2, B1, ou "non validé". Le niveau B2 en interaction orale n'est validé que si l'ensemble des critères sont validés au niveau B2. Parmi ces huit critères, deux nous intéressent particulièrement : 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 claire pour être aisément compris(e), même si un accent subsiste" et "globalement compréhensible malgré l'accent étranger et/ou des erreurs de prononciation". Le niveau B2 semble donc ici aussi caractérisé par une certaine fluidité de parole et d'aisance de compréhension côté auditeur.

Au niveau C1, "il est attendu que les candidats fassent preuve de fluidité, d'aisance et de spontanéité".

1.1.2 Titre à redéfinir

Descripteurs prosodiques de Frost et O'Donnell, 2018

Frost et O'Donnell, 2018 proposent une une grille d'évaluation conçue pour évaluer la prononciation des locuteurs anglophones non-natifs, et plus spécifiquement les apprenants francophones. Elle combine des éléments segmentaux et suprasegmentaux, en donnant une grande importance aux derniers —d'où son nom de des-

[7] https://www.certification-cles.fr/se-preparer/grilles-d-evaluation/grilles-d-evaluation-1196363.kjsp

cripteurs prosodiques. Les descripteurs sont divisés en 5 catégories : rythme et accent lexical, intonation, réduction syllabique, voyelles pleines, et phonotactique (comprenant entre autres les phénomènes de liaisons, d'élision, ou d'assimilation). Le focus porte principalement sur la qualité et la position de l'accent lexical (comme garant du rythme), la variation de l'intonation et son adéquation aux intentions du locuteur, ou encore la qualité des voyelles accentuées ou réduites.

L'échelle de compréhensibilité d'Isaacs et al., 2018

L'échelle de compréhensibilité d'Isaacs et al., 2018[8] a été spécifiquement conçue pour l'évaluation formative. Comme pour les descripteurs prosodiques de Frost et O'Donnell, 2018, son objectif est d'aider les enseignants et les apprenants à prendre conscience des caractéristiques de la prononciation qui rendent la parole difficile à comprendre, et de mettre des mots et des descripteurs derrière les jugements holistiques généralement effectués intuitivement avec des échelles de Likert. L'échelle de compréhensibilité décrit 6 niveaux de compétence, avec une échelle globale et 4 sous-échelles ciblant respectivement la phonétique, la fluence, le vocabulaire et la grammaire – les deux derniers ayant une importance moindre d'après les auteurs. La sous-échelle de phonétique fait ressortir l'importance de la position de l'accent lexical ; celle de fluence la localisation des marqueurs d'hésitation, sans pour autant indiquer leurs types ni ce qu'est une position inappropriée. Les auteurs précisent que leur échelle n'est pas adaptée aux tâches de production de parole trop prédictible, comme la lecture, les phrases indépendantes ou les dialogues, mais plutôt à l'évaluation de la parole spontanée.

1.2 Évaluation automatique

1.2.1 Prédiction de score

1.2.2 Évaluation diagnostique

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 (désormais ASR). 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

[8] https://www.iris-database.org/details/N1U17-87aaM

identifie alors quelle option 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 et débit de parole). Les premiers systèmes évaluant la parole spontanée sont arrivés dans les années 2000, en se focalisant sur la fluence de la parole (Cucchiarini et al., 2002), mais le spontané reste très marginal encore aujourd'hui par rapport à la parole lue, bien plus prédictible.

Après un fort intérêt pour l'évaluation automatique de la prononciation à la fin des années 90, la discipline s'est essoufflée à cause d'une mauvaise fiabilité des systèmes alors commercialisés (Witt, 2012). Elle est toutefois rapidement revenue à la charge avec la généralisation des smartphones à la fin des années 2000, l'augmentation de la puissance de calcul et l'amélioration des systèmes d'ASR. Notons la création du groupe SLaTE (Speech and Language Technology for Education) au sein de l'ISCA (International Speech Communication) en 2007, qui travaille spécifiquement sur l'apprentissage des langues et les technologies de la parole (Ellis, Bogart, 2007). Pour une revue détaillée des applications antérieures à 2011, le lecteur peut consulter (Witt, 1999, O'Brien, 2006, Levis, 2007, Eskenazi, 2009 ou encore Delmonte, 2011).

Mais à quoi ressemblent les systèmes d'aujourd'hui et comment évaluent-ils la prononciation ?

Commençons par distinguer deux types d'évaluation. L'évaluation formative d'une part (low stake assessment), qui a pour but d'accompagner l'apprenant dans son parcours et cibler ses acquis et ses difficultés ; et l'évaluation certificative d'autre part (high stake assessment), dont les enjeux sont plus importants, puisqu'ils peuvent être la condition d'obtention d'un diplôme, d'un emploi, voire d'un visa.

Commençons par l'évaluation formative. La majorité des applications récentes d'apprentissage des langues proposent des fonctionnalités d'évaluation de la prononciation. Citons seulement les plus connues : Duolingo, Memrise, Babbel, Busuu, Rosetta Stone, ou encore LingoChamp. Elles fonctionnent toutes sur le même principe, à savoir la lecture d'un mot ou d'une phrase à voix haute, et fournissent un feedback binaire (type bon/mauvais, Memrise, Duolingo, Busuu) ou un score de confiance de reconnaissance (en pourcentage, LingoChamp, Rosetta Stone). Elles proposent toutes d'entendre le mot ou l'énoncé en parallèle de la lecture, d'office (Memrise, Busuu, Babbel) ou à la demande (Duolingo, Rosetta Stone). D'autres applications sont dédiées à la prononciation, comme ELSA Speak ou IELTS Speaking Practice par exemple. Il s'agit toujours de lecture et/ou de répétition, mais les rétroactions proposées sont plus détaillées : les mots ou les phonèmes sont colorés en fonction de la confiance de l'ASR : le système indique quels phonèmes ont été reconnus (parmi

les phonèmes de l'anglais uniquement). Enfin, des scores spécifiques à telle ou telle catégorie de phonèmes sont proposés (ex. voyelles réduites, consonnes occlusives, phonèmes en fin de mot sur ELSA) ou des scores globaux (nombre moyen de mots correct par minute sur IELTS Speaking Practice).

On constate que la majorité de ces applications évaluent la parole lue (ou répétée) et proposent donc des protocoles d'élicitation très contraints. Quant aux rétroactions fournies à l'apprenant, si elles ne sont pas seulement binaires, elles se concentrent généralement sur les aspects segmentaux de la parole, sur la base de la reconnaissance de chaque mot ou phonème attendu, et sont présentées à l'utilisateur sans filtrage ni hiérarchisation. On peut alors se demander à juste titre si ces systèmes ont vraiment un impact positif sur la compétence des appre- nants. Plusieurs études proposent une évaluation de ces outils, mais la plupart sont financées par l'entreprise qui développe l'application en question et sont bien peu critiques, peu sont publiées dans des revues avec comités de lecture et le financement ou l'appartenance des auteurs ne sont pas toujours précisés. On notera tout de même quelques rares études moins enthousiastes quant à l'efficacité de ces outils, comme Becker et Edalatishams, 2018, Krashen, 2013, 2014, ou encore Nielson, 2011, mais celles-ci sont peu nombreuses.

Du côté de la recherche pourtant, l'évaluation automatique de la prononciation est toujours assez tendance, et un grand nombre d'études sont publiées chaque année dans les revues phare comme InterSpeech ou LREC.

La majorité des études publiées depuis les années 90, et encore aujourd'hui en 2022, présente des systèmes qui basent leurs scores sur une comparaison des mots ou des phonèmes reconnus par un système d'ASR avec un texte cible. Ce type de scores est appelé Goodness of pronunciation et a été introduit par Witt et Young (2000). La reconnaissance est plus ou moins contrainte et plus ou moins adaptée à la parole non standard de l'apprenant, et se décline dans une formidable diversité. Il peut s'agir d'un ASR classique et d'un taux de reconnaissance du texte cible (c'est le cas des applications commerciales présentées plus haut) ; certaines études proposent d'adapter le système de reconnaissance à la parole non native, en augmentant son lexique phonétisé avec des erreurs phonologiques types (Extended Recognition Network, Bada et al., 2020, Lee, Glass 2015), ou en adaptant les modèles acoustiques à partir d'enregistrements de la langue maternelle des apprenants (Tan, 2008, Goronzy et al., 2004), ou directement avec de la parole L2 (Duan et al. 2017, Li et al., 2016). D'autres encore proposent de comparer la reconnaissance d'un ASR standard avec celle d'un ASR entraîné sur 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 (Ding et al., 2020,

Arias et al., 2010), ou un modèle appris sur un ensemble d'enregistrements pour permettre plus de variabilité (Truong et al., 2018, Wang et al. 2015). On constate que dans la grande majorité de ces systèmes, l'évaluation de la prononciation se traduit sous la forme d'une distance par rapport à un modèle natif. Une grande partie d'entre eux se fonde sur la reconnaissance de la parole pour établir le score, qu'il s'agisse d'un taux de reconnaissance de mots ou de phonèmes, ou bien d'un score de proximité au modèle, plus difficile à interpréter. Certains combinent ces valeurs avec des mesures de débit de parole ou de fréquence des pauses pour ajouter un aspect prosodique à l'évaluation, mais peu d'innovation a été faite sur ce plan ces 20 dernières années. Nous rejoignons le constat de Detey et ses collègues (2016) quant au fait que la prosodie n'est pas suffisamment prise en compte dans l'évaluation. Rappelons enfin que presque tous les outils évaluent de la parole lue, souvent hors contexte, qu'il s'agisse d'applications commercialisées ou d'études plus exploratoires. Il s'agit même encore parfois de lecture de mots isolés (Kato et al., 2019, Truong et al. 2018). Si le défi technique d'évaluer la parole spontanée est clair, c'est davantage une sous-estimation de son importance dans la compétence en langue qui ressort de notre état de l'art. Nous faisons le même constat pour la prosodie.

Du côté certificatif, le problème est pris différemment. Plutôt que de comparer une production d'apprenant à un modèle natif, ce sont des modèles statistiques qui sont entrainés à prédire le score d'un évaluateur humain. Ces modèles de prédiction sont appris sur la base d'un grand nombre de productions d'apprenants évaluées par des experts sur différents critères (souvent globaux, du type fluence, degré d'accent, intelligibilité ou compréhensibilité), ainsi que sur des mesures automatiques diverses (avant tout des paramètres suprasegmentaux, mais de plus en plus souvent combinés avec des paramètres segmentaux (Fontan et al., 2018, Evanini, Wang, 2013), lexicaux (Yoon et al., 2012) ou syntaxiques (Loukina et al., 2015, Bhat, Yoon 2015, Chen, Zechner, 2011). Parmi les paramètres supraseg- mentaux 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 un très grand nombre de paramètres (77 pour Coutinho et al., 2016, 75 pour Loukina et al., 2015), mais ce sont souvent les mêmes qui sont les plus significatifs : le débit de parole et d'articulation, la proportion de pauses et leur durée moyenne. Une fois le modèle entrainé sur le corpus, le système est capable de prédire le score d'un nouvel enregistrement qui n'a pas été évalué manuellement. 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 avec de la parole spontanée : 0,823 pour Saito et al., 2022, 0,8-0,9 pour Shen et al., 2021, 0,799 pour Fu et al., 2020. Ce sont ces sys-

tèmes qui sont généralement exploités dans les tests certificatifs automatisés, comme le Versant English Test (Pearson Education, Inc., 2022), ou le TOEFL iBT (Zechner et al., 2019).

Si ces systèmes se révèlent performants pour prédire un niveau global sur une production donnée, ils sont toutefois difficilement exploitables en contexte diagnostique. Il est en effet peu pertinent d'indiquer à l'apprenant qu'il parle trop lentement ou trop vite, ou bien qu'il fait trop de pauses, et que c'est à cause de cela que son niveau est plus faible. Ces phénomènes sont une conséquence de difficultés en amont, mais pas nécessairement un problème en soi. Ce qui paraît plus pertinent en revanche, c'est d'identifier les éléments les plus susceptibles de dégrader la compréhensibilité en situation d'interaction orale, de les localiser et de les mesurer.

Chapitre 2

Compréhensibilité du locuteur

2.1 Définitions

2.2 Facteurs impactant la compréhensibilité du locuteur

2.2.1 Côté locuteur

2.2.2 Côté auditeur

2.3 Évaluation de 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 fluence et le rythme de la parole sont deux éléments qui reviennent régulièrement dans les grilles d'évaluation de la production orale. Dans ce chapitre, nous proposons d'approfondir ces deux notions, et présenter en détails deux phénomènes linguistiques qui y sont étroitement liés : les pauses et l'accent lexicale.

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 fluence. 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 plus spécifiquement la production de parole (*speech fluency* ou *oral fluency*). Cette 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 3 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 à la fluidité de la parole produite, la troisième enfin à la perception de

fluidité par l'auditeur. Lickley (2015) reprend les mêmes catégories mais les appelle les deux premières 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 que remédier au problème 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, mais certains auteurs tentent de mesurer directement la fluence de surface, à partir de l'analyse du 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 minute, avec ou sans pauses), le ratio de phonation (temps de parole sans pause divisé par temps de parole total), le nombre moyen de mots ou de syllabes par segments 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 présence 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 le rythme. Là encore, de nombreuses définitions co-existent, mais la plupart des auteurs semblent s'accorder sur le fait qu'il 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[1]. Di Cristo et Hirst (1997) le définissent comme l'organisation temporelle des proéminences. On parle aussi parfois de tempo (Frost & Picavet, 2014). Maintenant, il est légitime de discuter l'applicabilité de ce concept à une langue humaine naturelle, a priori beaucoup moins prédictible et régulière que ne peut l'être la musique par exemple (encore que). Toujours est-il que ce terme est largement utilisé pour qualifier un on-ne-sait-trop-quoi qui rend la parole plus « naturelle », mais surtout et comme nous allons le voir, plus facile à comprendre. Si la notion de fluence de la parole renvoie souvent aux patterns de pauses, celle du rythme renvoie principalement à l'accentuation des syllabes, et en particulier au phénomène d'accentuation lexicale.

[1] "*Rhythm is the recurrence of a perceivable temporal patterning of strongly marked (focal) values and weakly marked (non-focal) values of some parameter as constituents of a tendentially constant temporal domain (environment).*" (Gibbon & Gut, 2001, p. 95)

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 plus 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 6 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 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 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).

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 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 des pauses des segments de mots assez courts entre lesquels les interlocuteurs ont la possibilité de réagir, sans pour autant prendre le tour de parole *(backchannel*[2]*)*. Maynard (1989) appelle ces segments courts entre pauses des *Pause-bounded Phrasal Units* (PPU), caractéristiques par leur courte durée (2,36 mots en moyenne en japonais d'après son étude). Les pauses séparant les PPU servent au locuteur à 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 ; au contraire, il va plutôt avoir tendance à attendre que le locuteur apporte des informations complémentaires par lui-même. Ce type de pauses en japonais a donc, à

[2]Le backchannel, ou 相槌 *aizuchi* en japonais, correspond à l'utilisation fréquente d'interjections dans une conversation pour indiquer que le locuteur est écouté. C'est un phénomène particulièrement courant en japonais, mais qui existe aussi dans une moindre mesure en anglais et en français (White, 1989).

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 japonaise, 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. 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. 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 silencieuse, mais la valeur de ce seuil est très variable d'une étude à l'autre, comme le suggère le tableau ??. La revue d'une quarantaine d'études analysant les phénomènes de pauses dans la parole non pathologique nous montre qu'il 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, d'organisation 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).

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

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)*

travers une analyse de corpus écrit multilingue[3], Compione et Véronis observent que la distribution des durées de pauses suit une distribution logarithmique multimodale et non une loi arithmétique et normale comme il est couramment admis jusqu'alors. Ils identifient deux gaussiennes autour de 150 ms et 500 ms, quelque 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. 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[4]. Enfin, Goldman et al. (2010) analysent un corpus de 40 min de français de différentes situations de communication[5], 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), accompagnée 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 fixe et unique.

Si l'on part du principe qu'une pause est un phénomène perceptif, il semble peu pertinent de déterminer un seuil absolu de durée qui ne tienne pas compte du contexte (présence d'hésitations, longueur des segments) ou du débit de parole du locuteur. Certaines études choisissent ainsi 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 (Fauth & Trouvain, 2018). De rares études font état d'un seuil relatif au débit de parole du locuteur, variant entre 180 ms et 250 ms chez Duez (1982, 1991) (calculé à partir de la durée moyenne des occlusives intervocaliques), de 98 ms à 490 ms chez Kirsner et al. (2003) (calculé à partir de la distribution des durées de pauses par locuteur), ou de 138 ms à 384 ms chez de Jong et Bosker (2013) (calculé à partir du débit d'articulation de chaque enregistrement). De son côté, 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 "pause subjectives".

[3] Campione et Véronis (2002) analysent l'anglais, le français, l'allemand, l'italien et l'espagnol dans le corpus *Eurom*.

[4] Demol et al. (2007) analysent l'anglais, le français, l'italien, l'espagnol, le roumain et le néerlandais.

[5] Lecture, narration conversationnelle, journal télévisé et conférence scientifique.

Grosman et al. (2018) identifient également 2 gaussiennes dans la distribution des durées de pauses du corpus LOCAS-F[6], 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 locuteurs ; 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 ms 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 ne préconisent pas de définir de seuil de durée fixe pour exclure certaines données, et exclure seulement les mesures aberrantes (ils n'ont ainsi supprimé que 4 % des pauses de leur corpus).

- durée arithmétique ou logarithmique ? GrosmanAl2018 utilisent $\log_{10}$ ms. Kahng2018 log-transforme les durée moyennes de pauses, mais aussi les fréquences de pauses intra-proposition et inter-proposition pour approximer une distribution normale. SheaLeonard2019 transforment toutes leurs distributions qui ne sont pas normales avec log et square root transformations. Mais ChristodoulidesAl2017 critique l'utilisation des transformations logarithmiques → est-ce que ça change vraiment quelque chose ?

Les pauses ne se limitent toutefois pas aux phénomènes d'interruption de phonation. On parle de « pauses pleines » lorsque qu'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 répétions ou les reformulations comme pauses pleines : tout ce qui, en somme, interrompt le flux du discours.

3.2.3 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 (Candea, 2000), de pause agrammaticale ou encore disfluente (Fauth & Trouvain, 2018).

[6] *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 multilogues. 14 situations de communication différentes comprenant conférences scientifique, débats et discours politiques et académiques, interactions formelles et informelles, interviews, journaux radiophoniques, lectures radiophonique.

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[7] 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 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 préparée de ceux qu'ils appellent des "*successful speakers*" : 15 locuteurs anglophones natifs et non-natifs enregistrés lors de discours politiques, de Ted Talks, ou dans des vidéos à succès sur les réseaux sociaux. 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 proposition (*« 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 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

[7] 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.

et des mots grammaticaux qui en dépendent (Mertens, 2008), 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). Par ailleurs, plus la frontière syntaxique est grande, plus la pause est longue. 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[8]. 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 grande, plus la pause est longue (intra-ga $<$ inter-ga $<$ inter-séquence syntaxique $<$ inter-unité de rection), quelque soit la situation de parole. Ajoutons que, bien que significativement corrélée, la durée de la pause ne dépend pas que de sa position, mais peut être également influencée par la longueur des constituants la précédant ou la suivant par exemple (Krivokapić, 2007).

Grosjean et Deschamps (1975) comparent quant à eux 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 7 positions de pauses possibles : soit en fin de proposition (qu'ils appellent phrases, combinant un syntagme nominal (SN) et un syntagme verbal (SV), éventuellement accompagné de compléments), soit à l'intérieur d'une proposition, 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'intérieur 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 pour les locuteurs non-natifs ? Dans une analyse de la distribution des pauses en parole lue en français, Fauth et Trouvain (2018) observent que les lecteurs non-natifs font plus de pauses à l'intérieur des énoncés que les lecteurs natifs, et les débutants plus que les avancés. Le premier groupe est constitué de 20 lecteurs germanophones lisant à haute voix un texte en français des Trois Petits Cochons issu du

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

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. Ils 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 (dans son cas, anglophones et turcophones) ont tendance à faire plus de pauses à l'intérieur des énoncés[9] 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. Des résultats similaires sont observés par Kahng (2014) et Shea et Leonard (2019).

3.2.4 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 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)

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

constatent par exemple 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.

3.2.5 Pauses et évaluation de la fluence

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)*, ou plus largement de l'énoncé (proposition principale avec ses relatives). Il peut s'agir de fréquence de pause par type (par exemple, inter- ou intra-proposition), ou une durée moyenne, ou encore le nombre d'énoncés suivis d'une pause par exemple.

On peut donc constater que 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 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 ou de compréhensibilité. 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 univer-

sité 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[10]. La distribution syntaxique des pauses semble donc jouer un rôle important dans la perception de la fluence.

Dans une seconde expérimentation, Kahng (2018) 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 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

[10] 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 %.

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 vis-à-vis 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 ($R^2 = -4{,}96$ et $R^2 = -4{,}33$, $p < 0{,}001$) et la perception de fluence ($R^2 = -6{,}93$ et $R^2 = -5{,}33$, $p < 0{,}001$). La fréquence des pauses intra-mot est également un indicateur fort, suivi par celle des pauses inter-proposition ($R^2 = 2{,}23$, $p < 0{,}05$ pour la fluence, mais non significatif pour le niveau global).

3.3 L'accent lexical

L'accent tonique *(stress)* 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 ***diffe**r* et *de**fe**r* en anglais, ou *a**u**n* et *a**ú**n* 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 (qualité vocalique). Toutefois ces 4 niveaux ne

sont pas nécessairement exploités 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. 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 adopter une approche intégrée de l'accentuation (Vaissière, 1983).

Comme pour le phénomène de pause 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, qui sera à la fois influencé par l'accent phonétique et l'accent « linguistique » (théorique, plus ou moins conscientisé par l'auditeur).

Nous proposons de présenter en détails les caractéristiques fonctionnelles et acoustiques de l'accent lexical en anglais, puis celles de l'accent en français et en japonais. Nous nous intéresserons ensuite 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 aux différents moyens existants pour mesurer automatiquement 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, plus fortes, plus hautes, et présentent un mouvement de la F_0 plus important, avec une qualité vocalique dite « pleine », comparativement aux syllabes non accentuées qui auront tendance à être « réduites » (Cutler, 2015). Ainsi, 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 très accentué (quand se superposent l'accent lexical, l'accent de phrase et l'accent contrastif) à complètement réduit, voire supprimé. On distingue ainsi jusqu'à 7 niveaux d'accentuation, mais la plupart des auteurs s'accordent à dire que 4 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

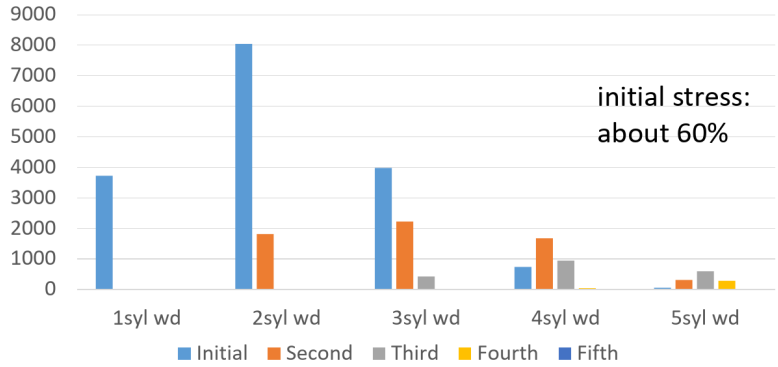


FIG. 3.1 : Distribution de la position de l'accent lexical en anglais dans les mots de 1 à 5 syllabes, à partir de la base de données CELEX (Sugahara, 2020, p. 8)

désambiguïsation lexicale. En anglais, les mots pleins (noms, verbes, adjectifs, adverbes etc.) sont généralement accentués, tandis que les mots grammaticaux (prépositions, déterminants, particules etc.) sont généralement réduits (Tortel, 2021). L'accent joue également un rôle important dans la morphologie dérivationnelle, car il change fréquemment selon la catégorie du mot (**per**son vs. per**so**nifier) et aide à distinguer des mots au sein de la même catégorie (**pho**tograph vs. pho**to**grapher). Ainsi, 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).

Si la position de l'accent lexical est variable en anglais, il est toutefois important de noter que la majorité des mots se trouveront accentués sur la première syllabe. 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 mot dans le flux de parole.

3.3.2 L'accent lexical en français et en japonais

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.

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 naturel 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). Comme pour les autres dimensions, il n'est pas évident de distinguer la variation due à la position de la syllabe (on observe souvent une baisse de l'intensité en fin de groupe de souffle,

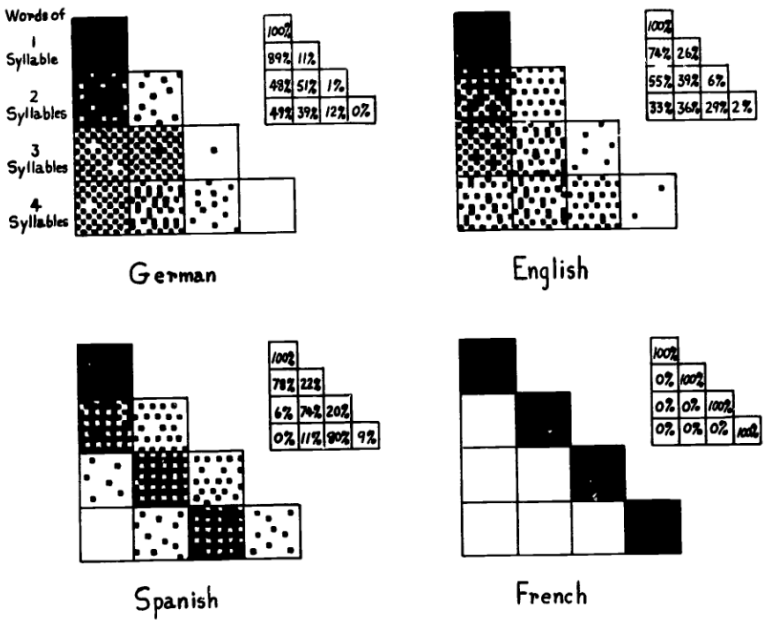


Fig. 3.2 : Comparaison de la position de l'accent primaire dans quatre langues (Delattre, 1963, p. 200)

Di Cristo, 1998), de celle qui serait due spécifiquement à la position de l'accent.

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.

Selon Sugahara (2020), les deux dialectes principaux du japonais, à savoir le dialecte de Tōkyō et celui du Kansai, ont tous les deux une 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 *hasi* (baguettes de table), *hasi* (pont) ou *hasi* (limite). Toutefois la position de l'accent varie en fonction du dialecte. On pourra ainsi avoir *kokoro* à Tōkyō et *kokoro* dans le Kansai (cœur, esprit), ou encore *kamakiri* à Tōkyō et *kamakiri* 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

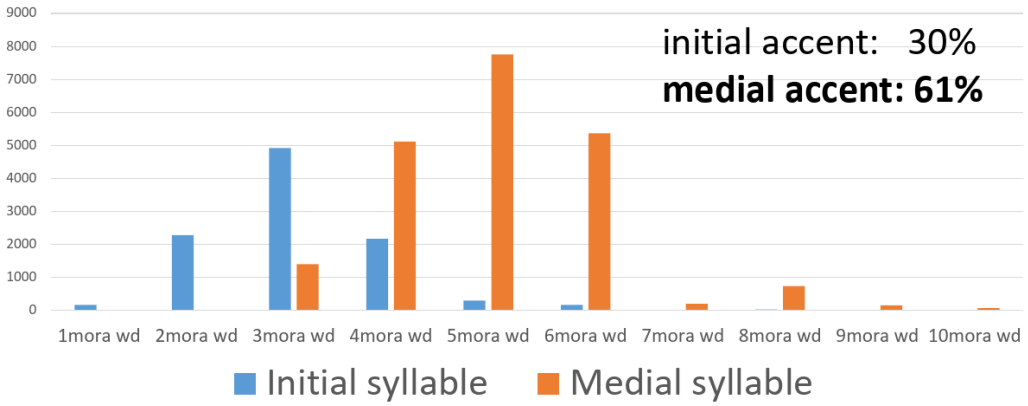


FIG. 3.3 : Distribution de la position de l'accent lexical dans les mots de 1 à 10 mores du Ōsaka-Tōkyō Accent Dictionary (Sugahara, 2020, p. 15)

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 l'accent lexical dans les mots de 1 à 10 mores dans le Ōsaka-Tōkyō Accent Dictionary REFERENCE. 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 ***kyooto*** (Kyōto) devient *kyooto-si* (la ville de Kyōto), mais *kyooto-**daigaku*** (l'université de Kyōto). Il serait intéressant de connaître la distribution de le position de l'accent sur un corpus de mots courants, mais nous n'avons malheureusement pas pu trouver cette information.

Ainsi il est généralement admis que le japonais est une langue à accent lexical, majoritairement en position médiale, et caractérisé par une variation de la F_0 sans modification significative des autres dimensions prosodiques ou segmentales (Sugahara, 2020).

3.3.3 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 mater-

nelle, et cela peut poser plus ou moins de problème selon que ces règles ou tendances diffèrent de la langue cible (Cutler, 2015). Par exemple, le locuteur francophone, habitué à un accent fixe sur le syllabe finale et une qualité et une quantité vocalique constante dans les voyelles non accentuées, aura tendance à accentuer la dernière syllabe des mots en anglais et à ne pas réduire les syllabes précédentes (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. De plus, puisque l'accentuation en français ne joue pas de rôle de désambiguïsation lexicale comme en anglais, les locuteurs francophones ont souvent des difficulté 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).

3.3.4 Accent lexical et évaluation de la 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 2.2.1). Ils calculent un *word stress error ratio* 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 ($0{,}74$, $p < 0{,}01$). Contrai-

rement aux paramètres de fluence qui apparaissent plus discriminants dans les petits niveaux, 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[11]. Il s'agit cette fois de voir sur quelles dimensions les évaluateurs s'appuient explicitement pour juger les locuteurs, et observer quels scores obtenus par dimension sont les plus corrélés 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 apparaissent fortement corrélées avec les annotations effectuées par Isaacs et Trofimovich (2012) ($r = 0{,}76$, $p < 0{,}01$ entre le critère « rythme » et la proportion de réduction vocalique – c'est la deuxième corrélation la plus haute parmi les 11 critères, après celle calculée entre le débit de parole et la longueur moyenne des énoncés, $0{,}78$ $p < 0{,}01$ – et $r = 0{,}70$, $p < 0{,}01$ entre le critère « accentuation lexicale » et le *word stress error ratio*). Ensuite, le critère « rythme » apparaît le plus corrélé avec le jugement de compréhensibilité parmi les 5 critères de prononciation ($0{,}79$), alors que le critère « accentuation lexicale » n'est qu'en quatrième position ($0{,}62$) après les erreurs segmentales ($0{,}75$), le débit de parole ($0{,}66$), et avant l'intonation ($0{,}54$).

Une autre étude intéressante est celle de Field (2005). Il s'intéresse à l'impact du déplacement de l'accent lexical sur la reconnaissance de mots isolés. Il constate qu'un déplacement de l'accent vers la droite impacte plus l'intelligibilité du mot qu'un déplacement vers la gauche, ce qui semble cohérent avec la tendance majoritaire en anglais à l'accent en initiale. Par ailleurs, le même déficit d'intelligibilité est observé quelque soit la langue maternelle de l'auditeur[12].

3.3.5 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). Certains

[11] 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).

[12] Field a mené son expérimentation sur 82 auditeurs natifs et 76 locuteurs non-natifs, dont les langues maternelles sont le coréen (16), le japonais (15), le mandarin (10), l'espagnol (9), le portugais (6) et l'italien (6), ainsi que d'autres langues avec moins de 5 locuteurs respectifs.

systèmes intègrent également des informations segmentales, telles que les coefficients cepstraux (Ferrer et al., 2015 ; C. Li et al., 2007). Toutes ces études ont entraîné leurs systèmes sur des corpus de parole lue, voire de mots isolés, parfois générés artificiellement à l'aide de systèmes de synthèse de parole. C'est le cas par exemple de Korzekwa et al. (2021), qui ont entraîné un réseau de neurones à attention pour détecter les mots dont l'accent primaire n'est pas placé correctement. À partir des paramètres acoustiques classiques (F_0, intensité et durée) extraits sur chaque phonème d'un corpus de mots isolés partiellement générés par synthèse vocale, ils mettent au point un modèle capable de déterminer si un mot est bien accentué avec une précision de 94,8 %. Toutefois, la moitié des erreurs présentes dans le corpus ne sont pas détectée (rappel de 49,2 %). Les auteurs ajoutent que leur modèle n'est pas adapté à l'analyse de mots en contexte, les résultats étant biaisés par les phénomènes de liaison et de coarticulation. Des performances similaires étaient obtenues quelques années avant en combinant coefficients cepstraux et paramètres prosodiques dans un modèle à mélange de gaussiennes (Ferrer et al., 2015).

On constate que les études qui tentent de mesurer les patterns accentuels des mots, malgré leurs infrastructures complexes et leur grand nombre de paramètres (jusqu'à 39 lorsque les coefficients cepstraux sont utilisés par Ferrer et al., 2015), sont limitées et semblent souvent déconnectées de la réalité de la parole et des enjeux pédagogiques. Saito et al. (2022) se rendent à l'évidence : on ne sait pas encore analyser automatiquement la précision de l'accentuation lexicale. Par ailleurs, nous n'avons trouvé à ce jour aucun système proposant de mesurer le degré de contraste prosodique entre les syllabes.

3.4 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 discours – qu'il s'agisse d'hésitations, de faux départ, de répétitions etc. –, participent à structurer le message et aident l'auditeur à traiter l'information. Leur présence peut aussi perturber la compréhension, en particulier lorsqu'elles interviennent à l'intérieur d'un constituant syntaxique. Le lien entre les pauses et la syntaxe est clair, et il semble difficilement concevable d'effectuer des mesures sur les pauses sans considérer leur position dans l'énoncé. En outre, nous retenons que la notion de pause n'est pas absolue. On peut percevoir une pause purement par cohérence syntaxique (sans aucun indice acoustique), ou au contraire ne pas la percevoir malgré la présence d'un silence ou d'une hésitation prolongée. Toujours est-il que la présence d'une pause n'est pas anodine : perçue ou non, elle semble avoir un impact sur la qualité de la transmission du message et sur la perception de fluence

qu'en aura l'auditeur.

La qualité de réalisation de l'accent lexical semble aussi, du moins en anglais, impacter la perception de compréhensibilité du locuteur. Le fait que les patterns accentuels de la L1 transparaissent souvent dans la L2 peut rendre la segmentation ou la reconnaissance du mot difficile, surtout quand les tendances sont opposées, comme en français (accent en finale) et en anglais (tendance à l'initiale). Quatre dimensions sont en jeu en anglais pour réaliser cette accentuation : une variation de hauteur, d'intensité, de durée et de qualité vocalique. Ces 4 dimensions sont intriquées et peuvent être plus ou moins exploitées selon les contextes et les locuteurs. Selon les langues, l'accent n'est pas toujours réalisé à l'aide de ces 4 paramètres : le français aura tendance à privilégier la variation de durée, là où le japonais s'appuiera sur la hauteur ; et l'utilisation des autres dimensions peut s'avérer difficile. Ce qui ressort clairement est la tendance des syllabes de l'anglais à se réduire ou s'accentuer de manière marquée, tandis que ce phénomène n'est pas présent, ou beaucoup moins marqué, en français et en japonais.

Chapitre 4

Outils de traitement automatique de la parole

Chapitre 5

Questions de recherche

L'objectif de cette thèse est de répondre à trois questions : Peut-on concevoir un outil d'annotation automatique de la parole spontanée L2 capable d'identifier des phénomènes impactant la compréhensibilité ? Observe-t-on des différences significatives entre les locuteurs de niveaux B1 et B2 en termes de segmentation et de rythme de la parole ? Et enfin, peut-on mesurer précisément l'impact de ces facteurs sur la perception de la compréhensibilité par l'auditeur ?

partie II

Méthodes

Chapitre 6

Collecte de données de parole

Notre objectif est donc d'observer les patterns de pauses et d'accentuation lexicale chez des locuteurs francophones et japonophones de l'anglais, à la frontière entre un niveau « intelligible mais difficile à comprendre » (équivalent B1) et « intelligible est facile à comprendre » (équivalent B2). Par ailleurs, il est important que la parole soit autant écologique que possible : éviter la parole lue, car elle ne fait pas ressortir les mêmes difficultés, ou la production monologale qui est rapidement limitée et peut paraître décontextualisée. Il est toutefois important que le cadre de production soit suffisamment contrôlé pour limiter l'hétérogénéité des données, comme la durée de parole, la participation des locuteurs, leurs niveaux ou encore la qualité audio. Des locuteurs issus d'un contexte universitaire permettrait probablement de limiter l'influence de l'âge, du statut, de l'expérience sociale et professionnelle ou de la motivation d'apprentissage.

La session d'interaction orale du CLES B2 est apparue comme une opportunité intéressante pour constituer ce type de corpus. Cet examen propose une mise en situation originale sous la forme d'un jeu de rôles d'une dizaine de minutes entre deux ou trois candidats, qui sont évalués *in situ* par un ou deux évaluateurs accrédités.

Grâce au concours 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, dont une partie des enregistrements constitue notre corpus principal. Par ailleurs, afin d'observer comment se comportent des locuteurs de langue maternelle japonaise et des locuteurs anglophones natifs dans une situation de communication similaire, nous avons constitué un second corpus grâce au concours des universités Dōshisha à Kyōto et Waseda à Tōkyō, au Japon.

Ce chapitre présente le contexte dans lequel ont été effectués les enregistrements, et le type de parole qui a été ainsi collecté. La description des données utilisées pour

nos analyses est présentée chapitre 9.

6.1 Contextes d'enregistrement

La certification CLES évalue les compétences des candidats indépendamment pour chaque niveau et sur 4 compétences langagières. Chaque compétence est évaluée dans une session d'examen spécifique, et c'est la session d'interaction orale du CLES B2 qui nous intéresse ici. Dans cette session, deux ou trois candidats s'engagent dans un jeu de rôles lors duquel il leur est demandé de présenter et défendre un point de vue sur un sujet polémique d'actualité, et d'aboutir à un compromis en une dizaine de minutes. Leur point de vue est déterminé par le rôle qui leur est attribué aléatoirement. Le sujet de la discussion est en lien direct avec les autres sessions du CLES effectuées en amont ; les candidats ont donc déjà eu l'occasion d'écouter, de lire et d'écrire à propos de la thématique en question.

Les candidats disposent de 2 minutes de préparation avant de commencer la discussion. Ils ont la possibilité de prendre des notes mais ne sont pas autorisés à lire pendant le débat. La discussion prend fin au bout de 10 minutes, ou lorsque les participants arrivent à un consensus et que les examinateurs jugent avoir suffisamment de matière pour évaluer le niveau des candidats.

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'examen du CLES. Toutefois, nous sommes conscients que le fait qu'il s'agisse d'un 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.

Le CLES n'est utilisé que dans les universités françaises, il donc est difficile de recueillir une quantité suffisante d'enregistrements de locuteurs de langue maternelle japonaise. Nous avons donc organisé une collecte d'enregistrements au Japon, en reproduisant autant que possible la situation de production orale du CLES B2. Les conditions d'enregistrement sont les mêmes que pour le CLES, à la différence que les locuteurs ont été recrutés et rémunérés, et que l'enregistrement ne s'est pas fait dans le cadre d'un examen réel. Il y a donc inévitablement moins d'enjeux pour les locuteurs, qui ne sont pas là pour passer une épreuve. Par ailleurs, les candidats ne sont pas préparés à la thématique comme ils le sont dans le corpus CLES.

Pour compenser ces différences, le temps de préparation avant le débat était plus flexible, la durée de la discussion n'était pas limitée mais devait durer au moins 10 minutes. Les conditions pour participer à l'enregistrement étaient les suivantes : être de langue maternelle japonaise, avoir un niveau d'anglais au moins équivalent à B1 (TOEFL iBT Speaking 16, IELTS Speaking 5.5, Eiken English Proficiency test pre-1 level), accepter les conditions d'utilisation et de diffusion des enregistrements, et être étudiant à l'université Waseda ou Dōshisha. Un questionnaire linguistique a été rempli par les candidats pour connaître, entre autres, leur niveau d'anglais certifié (TOEFL, TOEIC ou équivalent) et leurs séjours à l'étranger. Les binômes ont été constitués de manière à ce que les participants aient un niveau d'anglais équivalent.

Nous avons constitué un corpus complémentaire de locuteurs anglophones 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.

6.2 Sujets de certification utilisés

De nouveaux sujets de certification sont régulièrement édités par le CLES. Les sujets utilisés cette fois-ci concernent l'usage de la e-cigarette, la généralisation des caméras de surveillance et les tests cliniques sur les animaux. La formulation des sujets varient légèrement lorsque trois candidats participent au débat, de manière à ce que le troisième prenne un rôle de médiateur.

L'ensemble des énoncés du CLES B2 sur le thème de la e-cigarette 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 pas encore être rendus publics.

Dans le cas des enregistrements effectués au Japon, il n'a pas été possible d'utiliser les mêmes sujets. Aussi, deux sujets similaires à ceux du CLES ont été conçus pour l'occasion : l'utilisation des intelligences artificielles génératives en classe de langue et travailler en parallèle de ses études, (cf. annexes XX). Créer ces sujets nous a permis d'être plus près des préoccupations actuelles des étudiants et permettre une discussion plus riche.

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

6.3 Évaluation des locuteurs

Lors de la session d'examen, un ou deux examinateurs accrédités évaluent en direct les candidats sur 8 critères de production orale au niveau B2. Ces critères couvrent autant la qualité du contenu que la forme et le respect des consignes (cf. grille CLES B2, section 1.1.1). Le niveau B2 en interaction orale n'est attribué au candidat que s'il valide l'ensemble des 8 critères. À défaut, il peut valider le niveau B1 si ses compétences sont jugées suffisantes, ou ne rien valider du tout.

Dans le cas des locuteurs japonophones, il n'a pas été possible de bénéficier de la présence d'évaluateurs CLES. Le niveau des participants est donc estimé à partir des scores obtenus via d'autres certifications officielles, sur la base d'une auto-déclaration.

6.4 Caractéristiques techniques des enregistrements

Plusieurs sessions CLES ont été enregistrées sur des périodes différentes entre mai 2020 et janvier 2023. Les microphones utilisés sont variés, l'échantillonnage varie entre 44.1 et 48 kHz, 16 ou 24 bit PCM, stereo ou mono. Tous les enregistrements ont par la suite été rééchantillonnés à 44.1 kHz, 16 bit PCM stereo (pcm_s16le).

L'ensemble des enregistrements effectués au Japon a quant à lui été effectué avec un Zoom Handy Recorder H2n, et échantillonné à 44.1 kHz, 16 bit PCM stereo, 1411 kb/s.

Chapitre 7

Annotations et mesures

À partir des enregistrements de parole spontanée recueillis, nous souhaitons observer où les locuteurs ont tendance à placer leurs pauses et comment ils produisent l'accent lexical. En outre, s'agissant de parole conversationnelle, un certain nombre de traitements sont nécessaires en amont pour pouvoir effectuer ces mesures. Nous tenons également à ce que l'ensemble de ces traitements soit fait de manière automatique, afin de voir si ce type de mesures peut être effectué sans intervention manuelle.

Ce chapitre présente les différents traitements réalisés et les métriques utilisées pour les évaluer, ainsi que les mesures effectuées pour analyser les patterns de pauses et d'accentuation lexicale.

7.1 Identification du locuteur

Tout d'abord, les enregistrements de parole que nous voulons analyser sont des conversations spontanées entre plusieurs locuteurs, il est donc avant tout nécessaire d'identifier qui parle quand, de manière à pouvoir relier chaque phénomène de parole au bon locuteur.

À l'issue de ce module de traitement, nous souhaitons obtenir des segments de parole mono-locuteurs qui correspondent plus ou moins aux tours de parole. La principale contrainte qui se pose ici est de savoir jusqu'où nous tolérons les réactions de l'interlocuteur lorsqu'elles ne coupent pas la parole du locuteur. En effet, si l'outil de segmentation est trop sensible, la moindre réaction de l'interlocuteur risque de mettre fin au segment de parole en cours, résultant en de nombreux segments courts et non-terminés, ou commençant au milieu d'un énoncé. À l'inverse, si l'outil n'est pas assez sensible, les segments risquent de contenir beaucoup de parole de l'interlocuteur, qui

pourraient alors être analysée par erreur comme provenant du locuteur. En outre, la solution de supprimer a posteriori les passages de l'interlocuteur risquent d'affecter également la parole du locuteur, notamment lorsqu'il y a chevauchement de parole. Il faudrait pouvoir isoler la parole de chaque locuteur afin de pouvoir en conserver qu'une, même dans les cas de chevauchements.

Il est important de vérifier la qualité de la segmentation de parole et l'annotation en locuteur pour s'assurer que les mesures effectuées par la suite sont attribuées au bon locuteur. La précision de la reconnaissance étant dépendante du type d'enregistrements utilisés (qualité audio, nombre et voix des locuteurs, organisation des tours de parole etc.), nous proposons de tester le module sur un échantillon du corpus CLES manuellement annoté en locuteurs. Cet échantillon consiste en 20 enregistrements (3 h, 40 locuteurs, 2 locuteurs par enregistrements) segmentés automatiquement avec Whisper puis annotés manuellement en locuteurs.

La comparaison de l'annotation automatique et de l'annotation manuelle sera effectuée indépendamment pour chaque locuteur, en observant ⓐ la proportion de segment correspondant au locuteur cible, ⓑ la proportion de segment qui correspond au mauvais locuteur, et ⓒ la proportion de segment du locuteur cible manquée. À partir de ces valeurs, nous pouvons calculer un score de précision ($\frac{ⓐ}{ⓑ}$) et de rappel ($\frac{ⓐ}{ⓐ+ⓒ}$), et ainsi quantifier la proportion d'erreur par locuteur (qui impactera inéluctablement les résultats individuels), et la proportion de parole correctement annotée.

ILLUSTRATION ICI

7.2 Reconnaissance et alignement de la parole

L'étape suivante consiste à transcrire la parole des locuteurs et de d'aligner la transcription au signal, au moins au niveau du mot, de manière à pouvoir localiser les pauses en contexte, ainsi que les mots sur lesquels nous effectuerons les mesures acoustiques d'accentuation.

La première question qui se pose ici est de savoir s'il vaut mieux transcrire l'ensemble de la conversation puis de segmenter en locuteurs, ou bien de segmenter d'abord, puis de transcrire segment par segment. Dans le premier cas, le système de reconnaissance dispose de l'ensemble de la conversation et donc du contexte global, pouvant améliorer la précision de la reconnaissance ; en contrepartie, la conversation est longue (environ 10 min) et plusieurs locuteurs se partagent la parole avec d'éventuels chevauchements. Il semble de plus en plus envisageable de choisir la première option car les systèmes de reconnaissance de la parole gèrent de mieux en mieux la reconnaissance des dialogues spontanés ; toutefois, au moment de faire ce choix, en

2021, nous optons pour la deuxième option, qui nous semble plus raisonnable : segmenter d'abord puis transcrire chaque segment indépendamment.

Pour évaluer la qualité de la reconnaissance de la parole, nous avons calculé le taux d'erreur mot *(word error rate, WER)* sur des corpus de parole plus ou moins contrôlée. D'abord sur des phrases porteuses élicitées par des locuteurs anglophones natifs, japonophones et coréanophones (5 h22 min, 4 799 phrases, 54 locuteurs) ; puis sur des textes lus par des locuteurs natifs et japonophones (34 h, 954 textes, 57 locuteurs) ainsi que des locuteurs francophones (3 h45 min, 1 texte commun, 148 locuteurs) ; et enfin sur de la parole spontanée transcrite manuellement issue du corpus CLES mentionné dans la section précédente (3 h, 40 locuteurs). Le WER est calculé en faisant le somme des substitutions, délétions et insertions, le tout divisé par le nombre total de mots dans le texte de référence, donnant un pourcentage d'erreur de reconnaissance. En complément du WER, nous analyserons également le nombre de substitutions, délétions et insertions.

Dans le cas des phrases porteuses et des textes lus, nous avons pris pour référence le texte source, en partant du principe que les lecteurs ont lu ce qui leur était demandé de lire. Pour la parole spontanée, les transcriptions ont été obtenues automatiquement avec Whisper puis corrigées manuellement.

Pour évaluer la qualité de l'alignement au niveau du mot, nous proposons d'adapter la méthode employée dans la section précédente pour comparer l'alignement cible à un alignement de référence. Nous disposons de deux enregistrements alignés au niveau du mot, qui pourront servir de référence. Ces enregistrements proviennent de l'étude de Frost et al. (2024), il s'agit de l'enregistrement de deux enseignants francophones faisant leur cours en anglais (3 min48 s et 3 min34 s). Les enregistrements ont d'abord été alignés automatiquement avec WebMAUS (Kisler et al., 2017), puis corrigés manuellement par un phonéticien. Nous utiliserons cet alignement comme référence. Pour chaque enregistrement, nous calculerons la proportion d'alignement correspondant entre l'hypothèse et la référence, ainsi que la proportion d'erreur d'alignement spécifique aux mots cibles sur lesquels porteront les analyses acoustiques.

7.3 Détection des noyaux syllabiques

Pour effectuer les mesures d'accentuation lexicale, il est encore nécessaire de segmenter les mots en syllabes et identifier les noyaux syllabiques sur lesquels porteront les analyses acoustiques. Nous envisageons deux méthodes pour localiser les syllabes : une méthode acoustique et une méthode phonologique. La méthode acoustique consiste à se baser sur les pics d'intensité. Chaque syllabe d'un mot consiste en principe en un pic d'intensité, nous pourrions donc considérer chaque pic à l'intérieur

d'un mot comme un noyau de syllabe. La méthode phonologique consiste quant à elle à utiliser un alignement forcé de chaque phonème correspondant au mot, à partir d'un dictionnaire phonologique, puis de considérer chaque intervalle vocalique comme noyau syllabique. Dans le premier cas de figure, les noyaux sont donc représentés par des points (maximums d'intensité), et dans le second, ce sont des intervalles de durée vocalique.

Si la méthode phonologique permet d'obtenir autant d'intervalles vocaliques que de syllabes attendues pour un mot donné (puisque l'alignement se base sur un dictionnaire phonologique), la méthode acoustique permet de d'estimer le nombre et la position des syllabes quelque soit le mot et la façon dont il est prononcé. Toutefois, nous n'avons pas encore formalisé de méthode pour vérifier si les noyaux acoustiques correspondent effectivement à un noyau vocalique, ou si les noyaux non-détectés le sont effectivement à cause d'une élision de syllabe par le locuteur. Nous nous contenterons dans un premier temps de compter le nombre de mots dont le nombre de syllabes acoustiques correspond au nombre de syllabe attendu dans un dictionnaire phonologique de référence.

7.4 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. Comme Fauth et Trouvain (2018), nous entendrons par « pause » toute interruption de parole, qu'il s'agisse de pause pleine ou silencieuse, faux départ, répétitions ou allongements.

À ce stade, nous disposons de l'alignement des mots au signal de parole. Nous pouvons donc par extension localiser dans la chaîne de texte les silences ou ce qui peut constituer une pause pleine. Chaque intervalle séparant deux mots sera considéré comme pause potentielle. Il sera ensuite possible de définir un seuil de durée minimum pour considérer ou non ces intervalles comme des pauses. À l'avenir, nous souhaitons toutefois moduler ce seuil en fonction du débit de parole (cf. chapitre 14).

Pour étiqueter les pauses en fonction de leur position syntaxique, nous proposons d'effectuer une analyse syntaxique par constituants, pour délimiter et hiérarchiser l'énoncé en propositions et en syntagmes. Chaque intervalle sera ainsi annoté de son contexte gauche et droit : la catégorie du mot qui précède et qui suit, le consti-

tuant le plus grand qui se termine, le nombre de mots qu'il contient et sa profondeur syntaxique (estimée à partir du nombre de constituants en cours), le constituant le plus grand qui commence, son nombre de mots et sa profondeur syntaxique. L'étiquette du constituant pourra ensuite être interprétée en frontière de proposition, de syntagme ou de mot à partir des catégories de constituants de PennTree Bank, donnée en Annexe A.

7.4.1 Analyses

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 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.

Grâce à l'étiquetage présenté ci-dessus, nous pouvons catégoriser chaque frontière de mot en fonction du type de frontière syntaxique : inter-proposition (si une proposition se termine ou commence), inter-syntagme (si un syntagme se termine ou commence), ou à défaut, intra-syntagme (si aucune frontière de constituant n'est présente). Par extension, nous pouvons donc catégoriser les pauses en fonction du type de frontière sur laquelle elles interviennent. On pourra alors comparer les groupes de locuteurs (B1, B2, francophones, japonophones ou anglophones natifs) en fonction de la fréquence des pauses produites et de leur distribution syntaxique.

On trouve différentes mesures de fréquence des pauses dans la littérature : le nombre de pauses par minute, par mot, par syllabe, ou encore par tour de parole. La fréquence des pauses par minute est influencée par le débit de parole : plus le locuteur parle vite, plus le nombre de mots par minute augmente et par conséquent le nombre de pauses potentielles, bien que cela puisse paraître contre-intuitif. Pour neutraliser l'influence du débit de parole, nous choisirons de calculer la fréquence des pauses par mot, ou plus exactement par token issu de la phase de transcription et d'alignement. Pour la fréquence des pauses en fonction de leur position, 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)) :

$$F_{p,i\in\{BC,BP,WP\}} = \frac{N_{p,i}}{N_i} \tag{7.1}$$

avec $N_{p,i}$ le nombre de pauses p par catégorie i, et N_i le nombre de frontières

syntaxique de type i. La valeur obtenue indique par exemple à quelle fréquence deux propositions sont séparées par une pause chez un locuteur donné. Nous compléterons cette mesure par la proportion de pauses P_i de chaque catégorie i :

$$P_{i \in \{BC,BP,WP\}} = \frac{N_{p,i}}{N_p} \tag{7.2}$$

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.

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 baserons sur les résultats de 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 %.

7.4.2 Score de distribution syntaxique

Nous proposons de calculer un score de distribution syntaxique des pauses (SDS) 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 de chaque catégorie (inter-proposition (BC), inter-syntagme (BP) et intra-syntagme (WP)) que l'on pondère 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. Ce calcul peut se noter de la façon suivante :

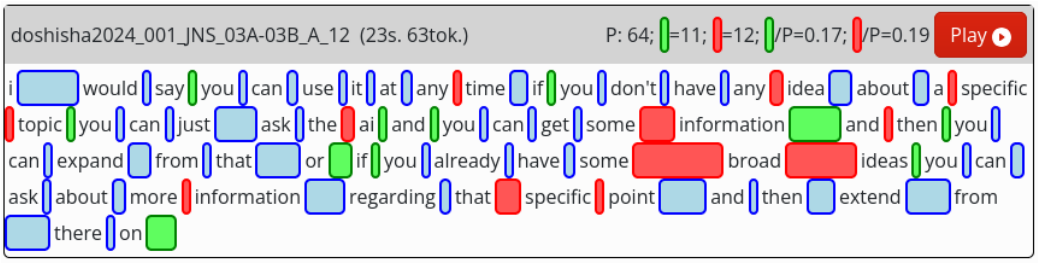


FIG. 7.1 : Transcription d'un segment audio indiquant chaque type de frontière syntaxique par une couleur : vert pour inter-proprosition, bleu pour inter-syntagme et rouge pour intra-syntagme. La longueur des intervalles est proportionnelle à leur durée ; seule les plus longs sont considérés comme des pauses par PLSPP (cliquer ici pour accéder à la visualisation en ligne)

$$SDS = \sum_{i \in BC,BP,WP} (p_i \cdot w_i) = \frac{N_{BC} \cdot w_{BC} + N_{BP} \cdot w_{BP} + N_{WP} \cdot w_{WP}}{N_p} \quad (7.3)$$

Nous proposons de fixer 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 participeront 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, ce qui semble toutefois peu probable.

7.4.3 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 fait que le nombre de frontières intra-syntagmes est assez limité en anglais, et réduit ainsi la probabilité d'y trouver une pause. La figure 7.1 illustre cet état de fait : elle présente un segment du corpus CLES-JP où chaque frontière syntaxique est colorée en fonction de son niveau. On y voit seulement 12 frontières intra-syntagme (en rouge), contre 41 inter-syntagmes (bleues). 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 nouvelle approche pour contourner ces limitations.

Au lieu de considérer des niveaux de frontières syntaxiques en fonction de leur

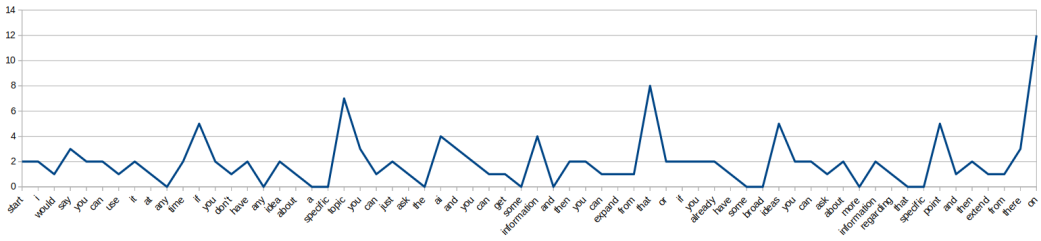


FIG. 7.2 : Nombre de constituants se fermant ou s'ouvrant après chaque mot

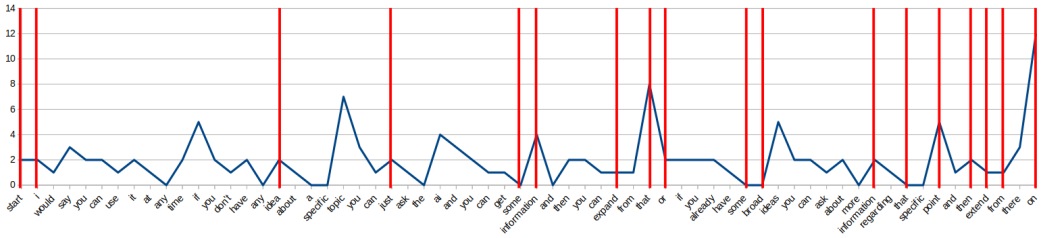


FIG. 7.3 : Même figure avec position des pauses (en rouge, seuil de durée minimale à 180 ms)

type, il s'agit de les considérer relativement aux autres. Plus le nombre de constituants qui s'ouvrent ou se ferment est élevé à un endroit donné, plus la frontière est importante. Quelque soit la nature de la frontière, c'est leur concomitance qui définit leur importance. 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 degré de frontière – par exemple la somme des constituants qui se ferment ou qui s'ouvrent après chaque mot. La figure 7.2 représente cette valeur pour le même segment que la figure précédente. On y distingue des pics qui semblent correspondre à des positions naturelles de pauses dans l'énoncé. Plus la valeur est basse en revanche, moins une pause à cet endroit semble probable.

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 ci-dessus : 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 :

- **Analyse globale :** nombre de pauses par token et durée moyenne des pauses ;
- **Analyse structurelle :** fréquence des pauses par type de frontière syntaxique ($F_{p,i}$), proportion des pauses par catégorie (P_i), score de distribution syntaxique basé sur les propositions et les syntagmes, score de distribution syntaxique basé

sur le niveau d'importance des frontières.

7.4.4 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 Mareková, développé à l'université Constantine le Philosophe à Nitra (Slovaquie), et l'Institut d'Informatique de Bratislava (REFERENCE). Le 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, et est entièrement transcrit est annoté 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, il s'agira de calculer le score de précision et de rappel de l'annotation automatique à partir des annotations de référence, et tenter de quantifier et expliquer les erreurs de détection et d'étiquetage.

Nous n'avons malheureusement pas trouvé de corpus annoté manuellement en syntagmes.

7.5 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). Nous ne disposons pas à ce jour de système d'évaluation de la précision de l'accent lexical adapté à la parole spontanée (Saito et al., 2022). Nous tenterons d'apporter une solution de traitement possible.

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 (plus basse en F_0, en intensité, et plus courte en durée). Au niveau segmental, la voyelle accentuée est pleine et parfois diphtonguée, tandis que la voyelle réduite est centralisée et tend vers le phonème /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.

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 principe en mesure d'identifier le patron accentuel attendu pour chaque mot du corpus, en recourant à un dictionnaire phonologique de référence. Pour chaque mot polysyllabique lexical, nous proposons de mesurer le degrés 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 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.

Deux scores seront ensuite calculés par locuteur : un score de position de l'accent, représentant le pourcentage de mots pour lesquels la syllabe proéminente correspond à la syllabe qui porte l'accent lexical selon le dictionnaire de référence (on pourrait donc dire que l'accent est correctement positionné) ; et un contraste prosodique moyen $\overline{C_w}$ 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 :

$$C_w = P_{s,w} - \overline{P_{u,w}} \tag{7.4}$$

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, +100 indique un contraste maximum positif, -100 indique un contraste maximum négatif, signalant que la proéminence se situe sur une autre syllabe que celle censée être accentuée.

Nous effectuerons les mesures suivantes pour chaque groupe de locuteurs :

- Nombre de mots, nombre de mots polysyllabiques lexicaux, nombre de mots annotés ;

- Proportion de mots par catégorie grammaticale et par nombre de syllabes ;
- Proportion de mots selon la position de l'accent lexical (attendu) et selon la position de la syllabe proéminente produite (réalisé) ;
- Score de position de l'accent par locuteur ;
- Contraste prosodique moyen par locuteur, moyen et par dimension (F_0, intensité et durée) ;

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

a) Demander à des locuteurs 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 ;

d) Comparer les résultats obtenus sur différents types de parole avec 1) la méthode acoustique et 2) la méthode phonologique d'identification des noyaux syllabiques, pour voir si les mesures sont cohérentes et si les mêmes tendances générales sont observées.

a) Évaluation perceptive par des locuteurs natifs

La première approche consiste à vérifier si les annotations automatiques d'accentuation lexicale sont cohérentes avec le jugement d'auditeurs natifs. En d'autres mots : est-ce que les anglophones natifs entendent l'accent au même endroit que la machine ? Nous avons recruté 10 évaluateurs anglophones natifs à qui nous avons fait annoter manuellement 6 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, ils sont donc habitués à l'influence du japonais sur la prononciation de l'anglais, mais aucun d'entre eux n'est enseignant d'anglais. En ce qui concerne les enregistrements, six élèves entre 9 et 11 ans d'une école primaire privée de la préfecture de Kyōto ont été 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 est fournie aux évaluateurs au format papier, en 6 exemplaires, et les évaluateurs doivent noter à la main la position de l'accent qu'ils perçoivent pour chaque enregistrement écouté sans contrainte d'écoute, dans un ordre aléatoire.

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, 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 ailleurs, un score d'accentuation S_w a été calculé pour chaque mot à partir des valeurs prosodiques mesurées par le système, puis comparé à un second score indiquant la moyenne des jugements humains pour le mot en question. L'équation 7.5 détaille le calcul effectué pour obtenir le score d'accentuation :

$$S_w = \frac{P_{s,w}}{P_{s,w} + \overline{P_{u,w}}} \tag{7.5}$$

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 en 0 et 1, 0.5 indiquant aucun contraste entre la syllabe accentuée et les autres syllabes, et 1 indiquant un contraste positif maximal. Les résultats obtenus sont détaillés en section 3.1.

b) Annotation automatique et conscience phonologique

La deuxième approche a consisté à comparer l'annotation automatique et le jugement de position théorique de l'accent par les mêmes locuteurs. Une liste de 57 mots cibles a été enregistrée dans des phrases porteuses par 12 locuteurs anglophones natifs, 14 locuteurs japonophones et 11 locuteurs coréanophones, puis annotée automatiquement avec notre système. En parallèle, ces mêmes locuteurs ont passé un test de conscience phonologique, lors duquel il leur était demandé d'indiquer la voyelle qui, selon eux, porte l'accent primaire, sur les mêmes mots cibles. 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 entre l'accent théorique et l'annotation automatique 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. Les résultats obtenus sont détaillés en section 3.2.

c) Annotation de parole produite par des locuteurs natifs

La troisième approche a consisté à considérer les locuteurs anglophones natifs comme une référence en termes d'accentuation lexicale, en partant du principe qu'ils accentueront systématiquement la syllabe censée porter l'accent primaire. Nous avons commencé avec de la parole lue en studio par des professionnels de la voix, enregistrés dans le cadre de la constitution d'un manuel scolaire d'anglais, et donc avec pour objectif de représenter un modèle d'anglais idéal. Par la suite, nous avons annoté des enregistrements de parole spontanée conversationnelle issus du corpus CLES.

Dans les deux cas, nous avons calculé la proportion de mots accentuée conformément au dictionnaire de référence, et le degré de contraste prosodique mesuré sur chaque dimension entre la syllabe accentuée attendue et les autres syllabes du mot. Les résultats obtenus en parole lue en studio sont présentés en section 3.3 ; ceux obtenus en parole spontanée sont présentés dans le chapitre 9.

d) Comparaison méthode acoustique *vs.* méthode phonologique

Enfin, nous avons souhaité observer comment varient les annotations automatiques selon que les noyaux syllabiques sont extraits de manière acoustique ou phonologique. Pour ce faire, nous avons comparé les résultats obtenus sur plusieurs corpus (parole lue, locuteurs natifs et non-natifs), en termes de proportion de mots correctement accentués par niveau et par locuteur, ainsi que de contraste prosodique observé sur chacune des 3 dimensions. Les résultats sont présentés section 3.4.

Chapitre 8

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

Ce chapitre présente la méthode employée pour mesurer l'impact de la distribution des pauses et de la qualité de l'accent lexical sur la perception de compréhensibilité du locuteur. Nous nous sommes inspirés du protocole expérimental de Nagle et al. (2019) et du logiciel Idiodynamic (MacIntyre, 2012), que nous avons simplifiés et adaptés à une évaluation de plus grande envergure en *crowd-sourcing*. Nous tenterons de répondre aux questions suivantes : Q1) Les auditeurs montrent-ils un comportement cohérent dans l'évaluation dynamique de la compréhensibilité des locuteurs L2, malgré les variations inter-évaluateurs ? Q2) Observons-nous une diminution de la compréhensibilité après des pauses WP et des mots dont le pattern accentuel n'est pas approprié ? Nous nous attendons à constater une augmentation du nombre de clics à la suite des pauses WP et des patterns accentuels incorrects, et inversement une diminution après les pauses BC et les patterns accentuels appropriés. On peut également s'attendre à ce que les enregistrements contenant plus de WP et de patterns accentuels incorrects seront jugés globalement moins compréhensibles.

Nous présentons d'abord notre protocole expérimental, les stimuli sélectionnés pour l'expérimentation, ainsi que les participants à l'expérience. Nous détaillons dans la dernière section les différents traitements effectués sur les données collectées, et les analyses effectuées. La plateforme expérimentale développée dans le cadre de cette étude est présentée avec les résultats, chapitre 13.

8.1 Adaptation du protocole

Nous avons donc 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, et ainsi analyser les fluctuations de ce jugement à travers le temps, pendant l'écoute. Si l'étude de Nagle et al. analyse ces fluctuations de manière globale dans une approche exploratoire, sans cibler de phénomène linguistique précis, nous proposons quant à nous d'exploiter le protocole pour observer comment varie le jugement des participants à la suite de certaines pauses ou patterns accentuels. Plus concrètement, nous souhaitons savoir si le jugement de compréhensibilité a tendance à diminuer à la suite de pauses de bas niveau (intra-syntagme, a priori disfluentes) ou de patterns accentuels inattendus, comparés au jugement mesuré à la suite de pauses de haut niveau (inter-proposition, a priori structurantes) ou 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 suivie d'entretiens rétrospectifs individuels comme c'est le cas dans le protocole d'origine. L'expérimentation a ainsi été repensée pour permettre une passation en complète autonomie : elle a été simplifiée et raccourcie pour ne pas dépasser un temps théorique de 35 minutes. Une rapide explication de la tâche à effectuer est fournie à l'écrit au début de l'expérience, suivie de 3 questions pour vérifier le profil du participant (une phase de filtrage est effectuée en amont, cf. section 8.3), et d'une phase d'entraînement. Par la suite, lorsque l'évaluateur est jugé trop peu actif, une *pop-up* de rappel s'ouvre avant le stimulus suivant. La consigne reste présente pour tous les stimuli.

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 doit faire une effort pour comprendre le locuteur, quelque soit la raison. De plus, il n'y a plus d'incrémentation du jugement comme c'est le cas sur le logiciel Idiodynamic. 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. À la fin de la lecture, il lui est demandé d'évaluer globalement la performance du locuteur en termes de qualité de prononciation, de fluence, 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 du locuteur qui l'ont rendu difficile à comprendre et quels conseils pourraient lui être donnés pour s'améliorer. Après la phase d'entraînement, tous les stimuli sont présentés de manière

aléatoire, et l'ensemble des stimuli est présenté à l'ensemble des évaluateurs.

8.2 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 chacun d'eux 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 caractérisée par une grande proportion de pauses intra-syntagme et de mots dont le pattern accentuel n'est pas celui qui est attendu, 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 la parole en termes de fluence et de qualité d'accentuation, nous avons calculé deux scores pour chaque segment : la proportion de pauses de type intra-syntagme (P_{WP}, présenté dans le chapitre précédent : nombre de pauses intra-syntagme divisé par le nombre de pauses total) et un score accentuel moyen $\overline{S_w}$. Comme le contraste prosodique C_w présenté dans le chapitre précédent, section 7.4, le score accentuel S_w représente le degré de contraste entre la syllabe censée porter l'accent primaire P_s et la moyenne des autres syllabes $\overline{P_u}$, mais normalise la différence des deux valeurs par leur somme pour obtenir une différence relative. La formule est la suivante :

$$S_w = \frac{P_{s,w} - \overline{P_{u,w}}}{P_{s,w} + \overline{P_{u,w}}} \tag{8.1}$$

Ainsi la différence entre 95 et 85 donnera un score S plus petit que la différence entre 15 et 5, tandis que le contraste C est toujours de 10 points. Il s'est avéré par la suite que ce score n'est pas bien approprié pour représenter le degré de contraste entre les syllabes d'un mot, puisqu'il n'y a pas de raison de donner moins de poids aux contrastes entre valeurs prosodiques élevées. Par contrainte de temps au moment de la rédaction de ce manuscrit, nous choisissons cependant de présenter les résultats obtenus en l'état, avec ce score accentuel S_w.

En projetant les segments sur un plan défini par ces deux dimensions, il s'agit alors de choisir les segments parmi ceux situés aux extrémités : P_{WP} élevé et score accentuel bas, et P_{WP} faible et score élevé (cf. figure 8.1). On appellera le groupe contenant les premiers « *low* », et les derniers « *high* ». Le seuil de durée de pause est fixé à 250 ms-2 s et seul les segments de plus de 60 tokens ont été considérés pour éviter

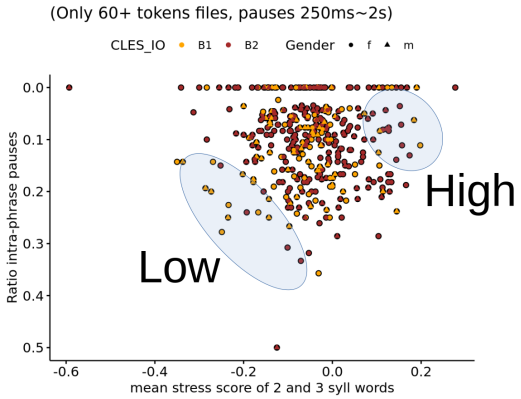


FIG. 8.1 : Choix des stimuli à partir des segments de plus de 60 tokens, projetés en fonction de P_{WP} et du score accentuel moyen

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

les segments trop courts.

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 8.1 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 ($> 0{,}2$), *Stress*Δ pour les mots au contraste peu marqué (score entre $-0{,}2$ et $0{,}2$) et *StressX* pour les mots au contraste négatif fort (score $< -0{,}2$).

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 polysyl-

labiques correctement reconnus et alignés au niveau du mot et des syllabes totalisent 82,0 %. Au risque de perdre un peu en précision, nous avons décidé de conserver les annotations automatiques en l'état, par soucis de cohérence avec notre démarche tout-automatique (aucune modification manuelle n'a été effectuée depuis le début des traitements).

8.3 Sélection des participants

Les participants ont été recrutés sur la plateforme britannique Prolific[1]. 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 fixé 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.[2]

8.4 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, 3 scores globaux numériques et un commentaire textuel optionnel. Étant donné que le comportement lors de l'expérience peut varier largement d'un participant à l'autre, il est nécessaire de vérifier la consistance des évaluateurs et de procéder à une normalisation des données recueillies. La consistance interne et la fiabilité des résultats est estimée à partir de l'alpha de Cronbach (Cronbach, 1951), et l'accord inter-annotateur est calculé à partir du coefficient de corrélation intra-classe[3] (Shrout & Fleiss, 1979),

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

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

[3] Nous indiquerons les valeurs de l'accord absolu et de la consistance moyenne inter-annotateur, respectivement ICC1k et ICC3k du package R psych v2.4.1.

à partir des scores numériques bruts. Les scores sont ensuite normalisés par z-score de manière à les rendre comparables.

Pour les patterns de clics, 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 :

$$M_w = \sum_{r=1}^{R} \left(C_{r,w} - \overline{C_r} \right) \tag{8.2}$$

avec M_w le nombre de m-clics dans la fenêtre w, R est le nombre d'évaluateurs, $C_{r,w}$ est 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, et dans quelle temporalité à la suite de l'événement. Un test de rangs non-paramétrique (Wilcoxon-Mann-Whitney) est 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 à accentuation de type StressO et StressX.

Pour vérifier notre dernière hypothèse, nous proposons de comparer les enregistrements en fonction des scores globaux obtenus (qualité de la prononciation, fluidité et compréhensibilité). Étant donné le nombre limité d'enregistrements, nous proposons de les diviser 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 seconde), des pauses inter-proposition et du score accentuel moyen. Les deux distributions sont ensuite comparées à l'aide d'un test non paramétrique (Wilcoxon-Mann-Whitney) et la taille de l'effet avec le delta de Cliff.

partie III

Résultats

Chapitre 9

Description du corpus de parole

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

Dans ce chapitre, nous présentons les trois corpus de données qui ont vu le jour à partir de ces enregistrements : un corpus CLES-FR de locuteurs francophones, un corpus CLES-JP de locuteurs japonophones, et un corpus CLES-EN de locuteurs anglophones natifs. Nous présenterons ensuite les deux dépôt de données publics associés, et les annotations manuelles qui ont été réalisées sur une partie des données pour confectionner un échantillon *gold standard*.

9.1 Corpus CLES-FR

Nous commencerons par décrire l'ensemble des enregistrements collectés lors des sessions du CLES, puis nous présenterons la sélection des données effectuée pour les analyses de notre travail de recherche.

Les 260 locuteurs enregistrés pendant des sessions CLES sont répartis en 232 binômes, 15 trinômes et 13 monômes. Les 13 monômes sont issus de session CLES B1 et présentent une situation de parole différente des autres groupes. Les 247 autres locuteurs ont été enregistrés lors de session CLES B2. La distribution homme/femme est équilibrée (130 femmes, 130 hommes). Une grande majorité des locuteurs a été certifiée B2 lors de l'examen (n=151, 58%), contre 75 B1 (29%) et 34 non-validés (13%). Quarante-cinq locuteurs (17%) n'ont pas déclarer le français comme langue mater-

nelle, parmi eux 22 locuteurs n'ont déclaré aucune langue, 6 locuteurs ont déclarer une langue arabe, 3 une langue chinoise et on compte encore 11 autres langues déclarées par une ou deux personnes à la fois.

Nous souhaitons observer les patterns de pauses et d'accentuation chez les locuteurs francophones B1 ou B2, aussi nous n'avons conservé dans le corpus final que les candidats ayant déclarer le français comme langue maternelle, ayant validé l'un des deux niveaux à l'examen, et nous avons également mis de côté les 13 monômes du CLES B1 car ils présentent une situation de parole trop différente des autres locuteurs. Le corpus final obtenu compte ainsi 170 locuteurs. Nous ferons dorénavant référence à ce corpus francophone B1/B2 par le nom de « CLES-FR ».

Le corpus CLES-FR comprend 99 locuteurs de niveau global B2 (58%) et 71 de niveau global B1 (42%). Le niveau obtenu spécifiquement pour la compétence d'interaction orale est B2 pour 118 locuteurs (69%) et B1 pour 52 locuteurs (31%). La distribution homme/femme reste relativement équilibrée avec 89 femmes (52%) et 81 hommes (48%). Parmi les 170 locuteurs, 11 sont enregistrés en trinômes (6%).

+ temps total estimé ; durée moyenne, min max ; durée médiane par locuteur, min max IQR

9.2 Corpus CLES-JP

Vingt-neuf étudiants de langue maternelle japonaise ont été enregistrés dans une situation de parole similaire à celle du corpus CLES-FR. Il y a 17 femmes (59%) et 12 hommes (41%). Leur niveau de compétence en anglais est estimé à partir de résultats obtenus à différentes certifications (TOEFL, IELTS, ou Eiken principalement) : 5 d'entre eux sont de niveau équivalent B1 (17%), 15 de niveau équivalent B2 (52%), et 9 de niveau équivalent C1 (31%). Les participants ont été répartis en binômes en faisant en sorte que chaque participant ait un niveau comparable à celui de son interlocuteur. L'un des 15 binômes enregistrés est constitué d'un étudiant de niveau C1 et d'un enseignant d'anglais de langue maternelle japonaise qui a dû participer suite à l'absence d'un candidat. Les tours de parole de l'enseignant ne font pas partie du corpus CLES-JP.

La principale différence avec les locuteurs du CLES-FR, hormis la langue maternelle, est le fait que les participants sont volontaires, rémunérés, et n'ont pas d'enjeu spécifique comme l'obtention d'un diplôme.

+ temps total estimé ; durée moyenne, min max ; durée médiane par locuteur, min max IQR

9.3 Corpus CLES-EN

Le corpus complémentaire de locuteurs anglophones natifs est quant à lui 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.

+ temps total estimé ; durée moyenne, min max ; durée médiane par locuteur, min max IQR

9.4 Publication des données

Les enregistrements des sessions CLES ont pu être organisées dans le cadre de l'examen et utilisées intégralement pour notre recherche. Toutefois, seuls les enregistrements pour lesquels l'ensemble des participants ont donné leur accord pour la publication des données a pu être mis à disposition de la communauté. Parmi les 260 locuteurs enregistrés, 162 ont donné leur accord (62%), et 138 d'entre eux font partie d'un enregistrement où tous les locuteurs ont donné leur accord (et donc diffusable en l'état).

Un corpus public d'une partie des enregistrements CLES a ainsi été mis à disposition sur la plateforme Ortolang[1]. Il réuni 62 enregistrements de 128 locuteurs, totalisant 10 h de parole. Parmi les locuteurs, 119 ont déclaré avoir le français pour langue maternelle (93%). On compte 61 femmes (48%) pour 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 pu être entièrement mis à disposition sur la même plateforme **(Lien pérenne)**.

9.5 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 contenir 20 candidats certifiés B2 et 20 certifiés B1, et un équilibre homme/femme. Le travail d'annotation a été effectué

[1] https://hdl.handle.net/11403/cles-spontaneous-english

par Nathanaël Berthet, stagiaire d'excellence de licence d'informatique à l'université Grenoble Alpes. Une transcription automatique des enregistrements à d'abord été effectuée l'aide du logiciel Whisper (Radford et al., 2022, modèle `base` multilingue), puis manuellement corrigé et segmenté 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.

Chapitre 10

Développement de l'outil PLSPP

Une chaîne de traitements automatisés a été développée dans le cadre de ce travail de recherche afin d'identifier la position syntaxique des pauses et de mesurer le degré de proéminence acoustique des syllabes à partir des enregistrements du CLES. Cet outil, auquel nous ferons référence par l'acronyme PLSPP (Pauses and Lexical Stress Processing Pipeline) se compose d'une suite de modules recourant tantôt à des outils existants, tantôt à des scripts d'analyses originaux. PLSPP est entièrement open-source est disponible sur ce dépôt GitLab[1].

Ce chapitre présente chaque étape de traitement et chaque module qui composent les versions 1 et 2 de PLSPP. Les versions suivantes seront brièvement présentées en fin de chapitre.

La figure 10.1 présente l'architecture générale de PLSPP. Les modules de traitement sont les suivants :

- Identification du locuteur et segmentation de la parole ;
- Reconnaissance automatique de la parole et alignement au mot ;
- Détection des noyaux syllabiques (acoustique ou phonologique) ;
- Étiquetage morphosyntaxique et analyse par constituants ;
- Annotation des pauses ;
- Extraction des paramètres acoustiques, annotation des proéminence syllabiques et comparaison avec le dictionnaire phonologique de référence.

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

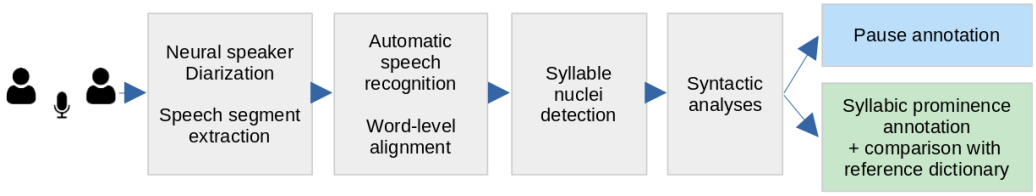


FIG. 10.1 : Architecture générale de PLSPP

10.1 Identification automatique du locuteur

Analyser automatiquement les enregistrements du corpus CLES est un challenge pour plusieurs raisons. Parmi elles, le fait qu'il s'agisse de conversations entre plusieurs locuteurs nécessite de pouvoir identifier qui parle quand. Par ailleurs, il arrive que la parole de plusieurs locuteurs se chevauche par moment, ou que l'un d'eux réagisse brièvement sans pour autant interrompre le tour de parole en cours. Enfin, certains tours peuvent être très courts ou non terminés. La difficulté ici est donc, outre l'identification du locuteur, de segmenter la parole en tours de parole.

L'outil choisi pour effectuer ce travail d'identification du locuteur et de découpage en segments de parole est pyannote.audio (Bredin, 2023 ; Plaquet & Bredin, 2023), combiné à un script de découpage et d'extraction audio paramétrable conçu pour les besoins de la présente pipeline de traitements.

Identification du locuteur Elle est gérée par Pyannote, qui est appelé par le script `diarisationPyannote.py`, qui prend en entrée les enregistrements audio bruts et renvoie un fichier texte par enregistrement listant chaque segment de parole détecté et son locuteur.

Fusion des segments Le script `pyannote2TextGrid.py` convertit ensuite les fichiers Pyannote en format TextGrid, en fusionnant les segments consécutifs du même locuteur. Un seuil paramétrable permet de jouer sur la sensibilité du découpage : plus il est élevé, plus les segments consécutifs ont tendance à être fusionnés, au risque toutefois de contenir des réactions de l'interlocuteur. Plus le seuil est bas, plus on limite la présence de l'interlocuteur dans le segment de parole, mais les segments ont alors tendance à être plus courts. La difficulté ici est de déterminer à partir de quelle durée on considère une prise de parole de l'interlocuteur comme un changement de tour, et jusqu'à quelle durée on la considère comme un simple backchannel qui peut être conservé dans le tour du premier locuteur. Nous avons paramétré ce seuil à 1 s par défaut, ce qui signifie qu'un segment est coupé à partir d'1 s de silence pour le locuteur

courant.

[[ILLUSTRATION AVANT et APRÈS FUSION]]

Extraction Le script `intervalles2wav.praat` extrait enfin chaque segment de parole en fichiers audio indépendants. Les paramètres disponibles sont la durée minimum des segments à extraire (par défaut 8 s) et la marge de découpage avant et après le segment (par défaut 10 ms).

En sortie de ce module, chaque enregistrement se retrouve découpé en autant de fichiers audio qu'il contient de segments de parole de la durée minimum paramétrée. Chaque fichier contient en théorie la parole d'un seul locuteur et peut être analysé indépendamment par les modules suivants.

10.2 Reconnaissance automatique de parole et alignement

Pour l'étape de reconnaissance automatique de la parole, WhisperX (Bain et al., 2023) est apparu comme l'outil le plus adéquat car il combine la haute précision de reconnaissance de Whisper avec une étape supplémentaire d'alignement au mot. Les récentes version de WhisperX intègrent maintenant une étape d'identification du locuteur avec Pyannote.audio, mais pour garantir une meilleure flexibilité notamment dans le découpage des segments, nous avons laissé les deux étapes séparées.

Transcription alignement au mot Le script `myWhisperxTG.py` exécute la reconnaissance de la parole et l'alignement au mot avec WhisperX. Il prend en entrée les fichiers mono-locuteur précédemment créés et renvoie la transcription alignée au mot en format TextGrid pour chaque fichier audio. Le script accepte plusieurs arguments comme le modèle utilisé (par défaut base.en), le type de processeur (par défaut CUDA, CPU et GPU en parallèle) et plusieurs paramètres techniques ajustables en fonction du serveur à disposition. Le modèle d'alignement est Wav2Vec2.0, mais celui-ci n'est pas paramétrable pour le moment.

10.3 Détection des noyaux syllabiques

Dans la première version de PLSPP, les mesures acoustiques pour l'accentuation lexicale sont réalisées au niveau des noyaux syllabiques estimés à partir des pics d'intensité par un script Praat de de Jong et al., 2021. En combinant ces noyaux syllabiques

avec l'alignement au mot de Wav2Vec2.0, il est possible d'identifier chaque syllabe des mots et de comparer leur mesures prosodiques. À partir de la version 2 de PLSPP, les mesures sont effectuées au niveau des intervalles vocaliques, localisées grâce à une couche supplémentaire d'alignement des phonèmes.

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 dans la version 2 de PLSPP recourt au Montreal Forced Aligner (MFA, McAuliffe et al., 2017) pour aligner le texte brut transcrit par WhisperX. L'avantage qu'il présente est qu'il effectue un alignement au mot plus juste et ajoute une couche d'alignement phonémique, permettant d'identifier le noyaux vocalique des syllabes. En contrepartie, MFA est plus sensibles 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.

10.4 Analyses syntaxiques

Deux types d'analyses syntaxiques sont effectuées : 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 consti-

tuants. Il prend en arguments le modèle d'analyse, par défaut `benepar_en3`[2].

10.5 Annotation des pauses

En sortie du module de transcription et d'alignement, nous disposons de la position estimée de chaque mot dans le signal audio. Par la suite, les analyses syntaxiques ont permis d'annoter chaque mot de leur partie du discours et d'obtenir l'arbre syntaxique à partir de la transcription de l'extrait. Dans ce module, nous nous intéressons non plus aux mots mais aux intervalles entre les mots. Dans le cas de l'alignement avec Wav2Vec2.0, tous les mots sont séparés par un intervalle vide étiqueté `<p:>` (parfois d'une durée de seulement quelques millisecondes). Ces intervalles ne sont pas nécessairement vides au sens de silencieux ; ils peuvent contenir des hésitations, des allongements, voire des faux départs – tout ce que WhisperX ne transcrit pas. Il nous a semblé opportun de considérer ces intervalles comme potentielle interruption du flux de parole.

Dans le cas de l'alignement avec MFA, les intervalles vides sont plus rares mais peuvent également être très courts (ex. 30 ms). Toutefois, l'alignement de Wav2Vec2.0 s'étant révélé plus fiable en parole spontanée que MFA, l'annotation des pauses est faite pour l'instant exclusivement à partir de l'alignement de Wav2Vec2.0.

Annotation des pauses 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 inter-mots, 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. À partir de là, l'utilisateur peut définir un seuil à partir duquel considérer un intervalle comme pause, et faire les analyses qu'il souhaite.

10.6 Annotation des proéminences syllabiques

L'objectif de ce module est de mesurer le degré de proéminence acoustique de chacune des syllabes des mots polysyllabiques, d'identifier la syllabe proéminente, et comparer sa position avec celle de l'accent lexical primaire tel qu'il est attesté dans un

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

dictionnaire de référence. Les mesures acoustiques sont réalisées sur trois dimensions : la fréquence fondamentale (F_0), l'intensité et la durée. Selon la version de PLSPP, l'annotation est faite tantôt ponctuellement au niveau des pics d'intensité des syllabes (version 1), tantôt sur toute la durée de la voyelle (versions 2 et suivantes). Afin de filtrer les mots potentiellement mal alignés, seuls les mots dont le nombre de syllabes détectées correspond à une réalisation possible selon le dictionnaire de référence CMU Pronouncing Dictionary[3] sont analysés. Dans les version 2 et suivantes, ce filtrage est optionnel.

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. 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). En contrepartie, il est nécessaire d'avoir suffisamment de mesures pour chaque locuteur, sans quoi plusieurs centiles peuvent renvoyer aux mêmes valeurs absolues. Une méthode de normalisation alternative est à l'étude pour permettre une annotation cohérente lorsque moins de données sont disponibles.

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 noyau syllabique acoustique (pic d'intensité), la F_0 est mesurée à partir du point le plus proche (”Get value at time...”, ”Hertz”, ”Nearest”), 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 en une moyenne des trois dimensions, et le pattern observé sur chacune des trois dimensions acoustiques (*cf.* figure 10.2). Les symboles pour représenter la syllabe proéminente et les autres syllabes sont personnalisables au début du script (par défaut ”O” et ”o” respectivement).

Cette version est actuellement la plus robuste car elle s'appuie sur une combinaison de l'alignement au mot et de la détection acoustique des noyaux syllabiques. Toutefois, les mesures sont effectuées de manière ponctuelle au niveau du maximum

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

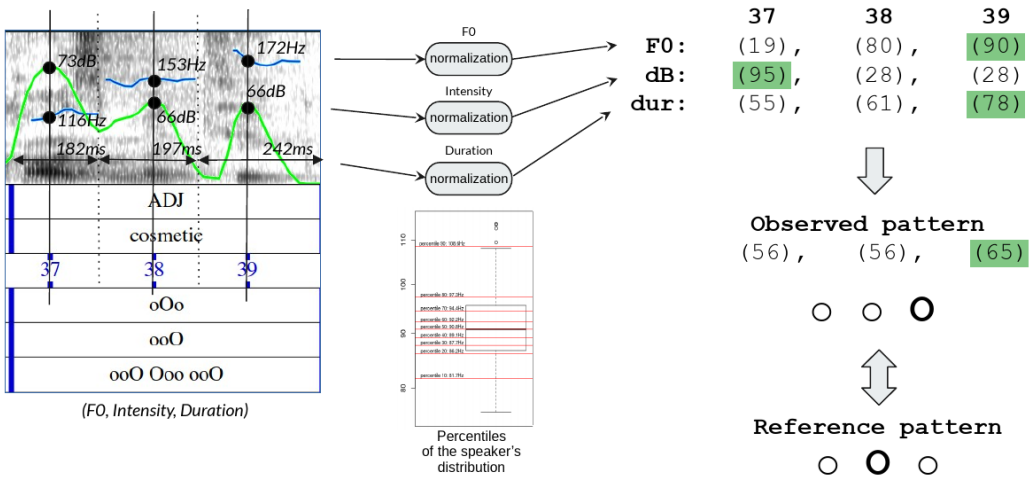


FIG. 10.2 : Extraction des paramètres prosodiques (PLSPP v1). À gauche un aperçu du fichier TextGrid de sortie avec les mesures acoustiques absolues indiquées en surimpression, la courbe bleue indique la F0, la verte l'intensité ; à droite les centiles correspondants au mesures absolue. "Observed pattern" correspond à la position de la syllabe proéminente (moyenne des trois dimensions prosodiques), "Reference pattern" correspond à la position attendue de l'accent primaire pour le mot "cosmetic"

d'intensité de la syllabe, et ne prennent donc pas en compte la variation de F0 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 faites sur une fenêtre glissante de taille paramétrable (par défaut 10 ms, inspiré de Ferrer et al., 2015) et les valeurs moyenne, minimum et maximum ainsi que l'écart type sont enregistrées. De même que pour la v1, un fichier TextGrid est généré avec les mêmes tiers que

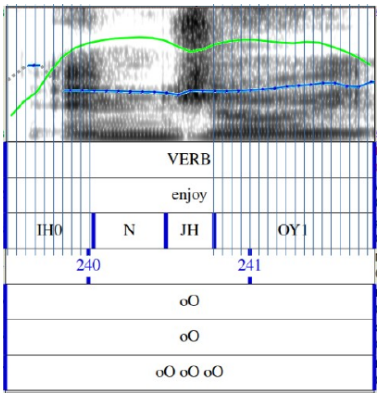


FIG. 10.3 : Extraction des paramètres prosodiques avec PLSPP v2. Les barres bleues ajoutées en surimpression représentent les frames de 10ms pour le calcul de la F0 (courbe bleue) et de l'intensité (courbe verte)

listées précédemment (*cf.* figure 10.3).

10.7 Évolution de PLSPP

PLSPP se décline aujourd'hui en 4 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, Fries et al., 2024 ; Coulange & Kato, 2023 ; Coulange, Kato, Rossato & Masperi, 2024a, 2024b, 2024c ; Coulange et al., 2023) ;
- **PLSPP v2** se base sur une identification phonologique des noyaux syllabiques, les annotations de l'accent sont plus précises mais moins robustes aux disfluences de la parole et donc moins adaptée à la parole spontanée. Elle a été utilisée pour l'analyse de phrases porteuses ou de textes récités par des locuteurs japonophones, coréanophones et anglophones natifs (Kimura et al., 2024 ; Sugahara et al., 2023, 2024) ;
- **PLSPP v3** est une évolution de la v2 permettant l'analyse des mots monosyllabiques. Elle permet de mesurer le contraste accentuel entre les mots lexicaux et les mots grammaticaux, et a été utilisée sur des textes lus par des locuteurs japonophones et anglophones natifs (Nakanishi & Coulange, 2024) ;
- **PLSPP v4** est une évolution de la v3 qui intègre des mesures de qualité vocalique pour analyser le degré de réduction et de diphtongaison des voyelles, combiné avec des mesures physiologiques d'aperture de la mâchoire réalisées avec un articulographe électromagnétique (EMA, Lezcano et al., 2020). Cette version a été utilisée sur de la parole de locuteurs lusophones (Brésil) et anglophones natifs (Raso et al., 2024)).

Le diagramme 10.4 présente chaque version de PLSPP et leurs différences.

10.8 Interface de visualisation des annotations

Présenter ici l'interface de visualisation interactive des annotations : statistiques globales sur la position et la réalisation de l'accent, avec options de filtrage de la population et des mots cibles ; visualisation segment par segment des mots cibles et de

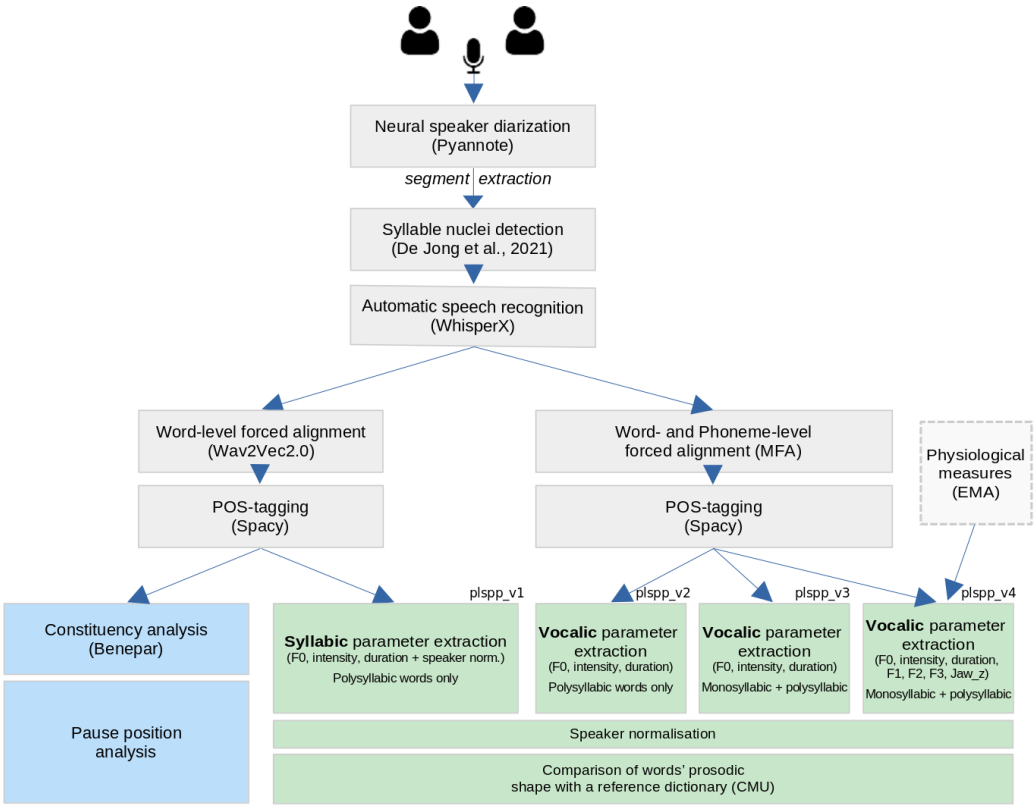


FIG. 10.4 : Architecture des 4 versions actuelles de PLSPP

leur accentuation en contexte, et visualisation segment par segment des pauses par catégorie, en contexte, et paramètres associés.

Chapitre 11

Évaluation du système

11.1 Modules de prétraitements

1.1 Identification automatique du locuteur

1.2 Reconnaissance automatique de la parole

1.3 Alignement mot-signal

1.4 Détection des noyaux syllabiques

11.2 Annotation des pauses

11.3 Annotation de l'accent lexical

3.1 Évaluation perceptive par des locuteurs natifs

Cette étude a été coordonnée par Takuya Kimura, étudiant de licence en informatique à l'université Dōshisha, et a été publiée et présentée en mars 2024 à l'Acoustical Society of Japan (Kimura et al., 2024).

3.2 Annotation automatique et conscience phonologique

Cette étude à été coordonnée par Mariko Sugahara, enseignante chercheuse du département d'anglais de l'université Dōshisha, et a fait l'objet de plusieurs communications (Sugahara et al., 2023, 2024).

3.3 Annotation de parole produite par des locuteurs natifs

3.4 Comparaison méthode acoustique *vs.* méthode phonologique

Chapitre 12

Analyses en parole spontanée

Ce chapitre présente les résultats d'analyse des annotations de pauses et d'accentuation produites par PLSPP sur les trois corpus de parole spontanée conversationnelle. Nous présenterons d'abord les analyses relatives à la distribution des pauses, puis celles concernant les patterns accentuels. Dans chacune des sections, les résultats obtenus avec les locuteurs B1 et B2 du corpus CLES-FR seront mentionnés en premier, et une sous-section finale sera dédiée à la comparaison des groupes de niveau des locuteurs japonophones (CLES-JP) et des locuteurs anglophones natifs (CLES-EN).

12.1 Analyse des patterns de pauses

Sur les 10 h20 min de parole continue extraite du corpus CLES-FR (70 locuteurs B1, 99 locuteurs B2), 72 140 intervalles inter-mots ont été analysés. La figure 12.1 présente la distribution de durée de ces intervalles. Leur durée médiane est de 80 ms avec le premier quartile à 40 ms et le troisième à 261 ms, toutefois on voit que la

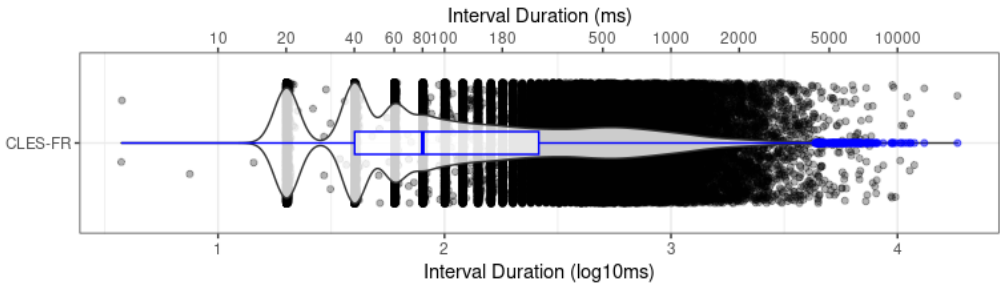


Fig. 12.1 : Distribution de durée des 72 140 intervalles inter-mots du corpus CLES-FR

distribution ne suit pas une loi normale, la grande majorité des intervalles étant de durée inférieure à 200 ms (70,3%), puis un second pic plus étalé autour de 500 ms. On peut voir que l'alignement des mots par Wav2Vec2.0 est effectué sur une fenêtre de 20 ms. La figure ne comprend pas 24 valeurs proches de 0 ($\log_{10}$de durée< 0). Parmi ces intervalles, on en compte 22 796 dont la durée est supérieure à 180 ms (32%), 1 085 supérieures à 2 s (1,5%), et 83 à 5 s (0,1%).

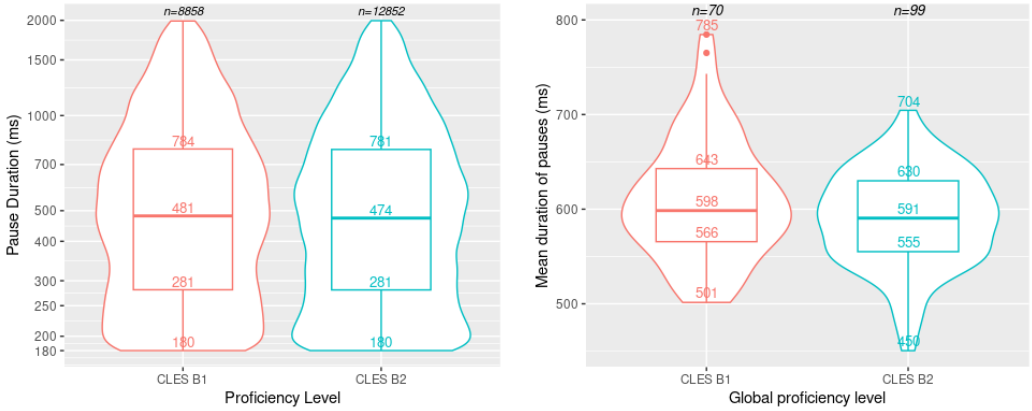
1.1 Durées et fréquences des pauses

Fixons maintenant un seuil minimum de durée à 180 ms pour ne considérer que les intervalles suffisamment longs pour être perceptibles par l'auditeur, et donc susceptibles d'être perçus comme des pauses. Comme indiqué dans le chapitre 7, nous fixons également un seuil maximum à 2 s pour ignorer les intervalles longs pouvant être dus à des erreurs d'alignement. Restent ainsi 21 710 intervalles (30%), que nous considérerons comme pauses dans la suite des analyses. La valeur médiane se situe maintenant à 481 ms, le premier quartile à 281 ms et le troisième à 782 ms.

La différence de durée de pause en fonction du niveau du locuteur est négligeable, bien que significative, entre les locuteurs B1 et B2 ($p < 0{,}05$, médiane à respectivement 481 ms et 474 ms, cf. figure 12.2a). Que se passe-t-il si l'on choisit des seuils de durée de pause différents ? Si l'on fixe le seuil de durée minimum à 250 ms, la différence entre B1 et B2 n'est plus significative (médianes à 581 ms). La prise en compte de pauses plus longues (jusqu'à 5 s) ne semble pas affecter beaucoup les mesures : la différence B1-B2 reste significative avec un seuil de 180 ms-5 s ($p < 0{,}01$, médianes à 501 ms), et reste non significative avec un seuil de 250 ms-5 s. D'après ces résultats, il semble pertinent de considérer les pauses courtes (inférieures à 250 ms) pour distinguer les niveaux B1 et B2, mais les pauses supérieures à 2 s ne semblent pas discriminantes. Dans tous les cas, la différence entre les deux groupes de locuteurs ne semble pas se situer au niveau de la durée des pauses, toutes pauses confondues, comme l'indique le Δ de Cliff toujours proche de 0 (cf. tableau 12.1). Si l'on calcule

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. 12.1 : *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.*



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

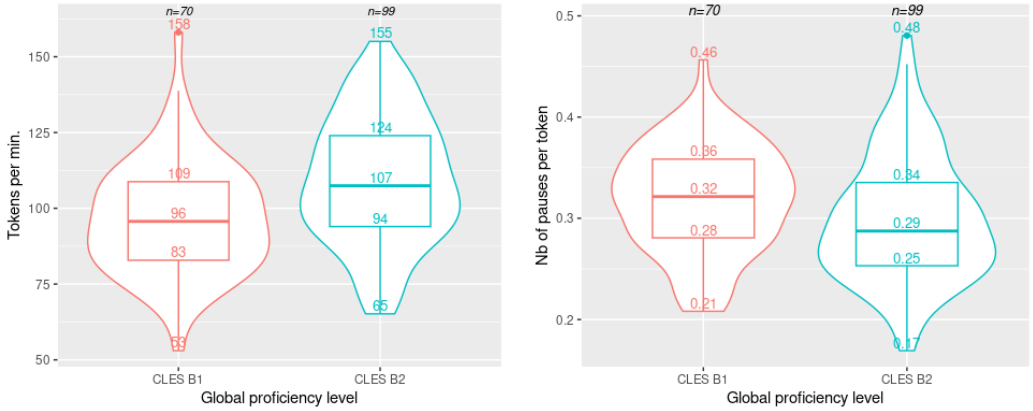
Fig. 12.2 : *Durées des pauses dans le corpus CLES-FR (*180 ms-2 s*)*

la durée moyenne des pauses par locuteur (cf. figure 12.2b), il n'y a toujours pas de différence significative entre les locuteurs B1 et B2, mais on se retrouve avec deux distributions aux formes distinctes. Celle des locuteurs B1 pointe vers le haut, indiquant que certains locuteurs font des pauses de durée moyenne particulièrement longue (notamment 6 locuteurs au-delà de 700 ms), tandis que celle des B2 pointe vers le bas, indiquant le cas inverse (5 locuteurs en dessous de 500 ms).

Le débit de parole des locuteurs B1 étant significativement plus lent que celui des B2 ($p < 0{,}001$, $\Delta = -0{,}35$ (medium), médianes respectives 96 et 107 tokens/minute, cf. figure 12.3a), il ne semble pas pertinent de tenir compte du nombre de pauses par minute par locuteur (différence non significative, médianes à 32 pour B1 et 34 pauses/minute pour B2). Le nombre de pauses par token permet quant à lui de mesurer la fréquence d'occurrence des pauses sans être influencé par la vitesse d'élocution. Pour un même nombre de mots, les locuteurs B1 font plus de pauses que les locuteurs B2 ($p < 0{,}05$, $\Delta = 0{,}154$ (small), médianes respectives à $0{,}32$ et $0{,}29$ pauses par token, cf. figure 12.3b).

1.2 Distribution syntaxique

S'il est possible de distinguer les locuteurs B1 et B2 à partir des mesures de débit de parole ou de fréquence de pauses, ces deux critères ne sont pas pour autant la cause d'une moins bonne compréhensibilité du locuteurs. Le chapitre 3 met en avant l'importance de la distribution syntaxique des pauses, et plusieurs études ont montré que la fréquence des pauses survenant à l'intérieur des groupes syntaxiques a tendance à être négativement corrélée avec la perception de fluence, tandis que celle des pauses



(a) Débit de parole *(b) Fréquence des pause*

Fig. 12.3 : *Débit de parole (gauche) et fréquence des pauses (droite) par locuteur dans le corpus CLES-FR (*180 ms-2 s*)*

survenant entre les groupes semblent moins impacter le jugement de fluence. Inspirés par des études précédentes (Kahng, 2014 ; Kallio et al., 2022 ; Shea & Leonard, 2019 ; Suzuki & Kormos, 2020), nous proposons dans un premier temps de calculer la fréquence des pauses relativement à deux types de groupes syntaxiques : les propositions et les syntagmes. Une pause pourra donc être inter-propositionnelle *(between clauses, BC)* si elle se trouve en frontière de proposition, inter-syntagme *(between phrases, BP)* si elle se trouve en frontière de syntagme, ou à défaut des deux premiers, intra-syntagme *(within phrases, WP)*. Pour une meilleure consistance avec les études précédentes, nous donnerons également les proportions de pauses intra-propositionnelles *(Within clauses, WC)*, qui comprend les pauses inter- et intra-syntagmes.

La figure 12.4 présente la proportion de pauses par type de frontière syntaxique. Pour le même nombre de propositions, il apparaît que les locuteurs B1 ont tendance à faire plus de pauses inter-propositionnelles que les B2 ($p < 0{,}001$, médianes respectives à 47 % et 42 %, $\Delta = 0{,}311$ (small) $IC = [0{,}132; 0{,}47]$), mais pas significativement plus à l'intérieur de celles-ci (*ns.*, médianes à 28 et 25 %, $\Delta = 0{,}172$ (small) $IC = [0; 0{,}334]$). Tous les locuteurs, même ceux à plus faible niveau, semblent donc privilégier les frontières de haut niveau syntaxique (entre les propositions) pour placer leurs pauses. Descendons maintenant au niveau du syntagme. Là encore, les locuteurs B1 font plus de pauses inter-syntagmes que les B2 mais la différence n'est pas significative (*ns.*, médianes à 29 et 26 %, $\Delta = 0{,}149$ (small) $IC = [-0{,}027; 0{,}316]$). En revanche, la différence au niveau intra-syntagme est significative, indiquant que les B1 font également plus de pauses que les B2 à l'intérieur des syntagmes ($p < 0{,}05$, médianes à 21 et 18 %, $\Delta = 0{,}187$ $IC = [0{,}009; 0{,}353]$).

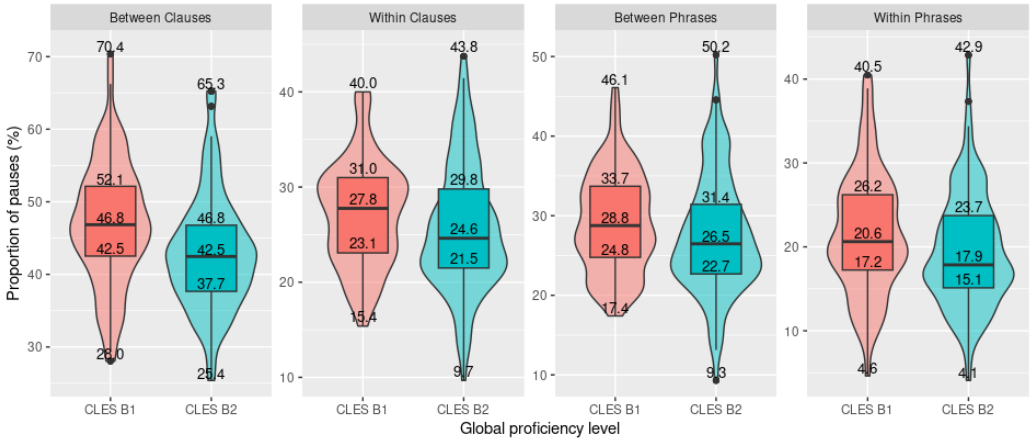


FIG. 12.4 : Proportion de pauses par type de frontière syntaxique

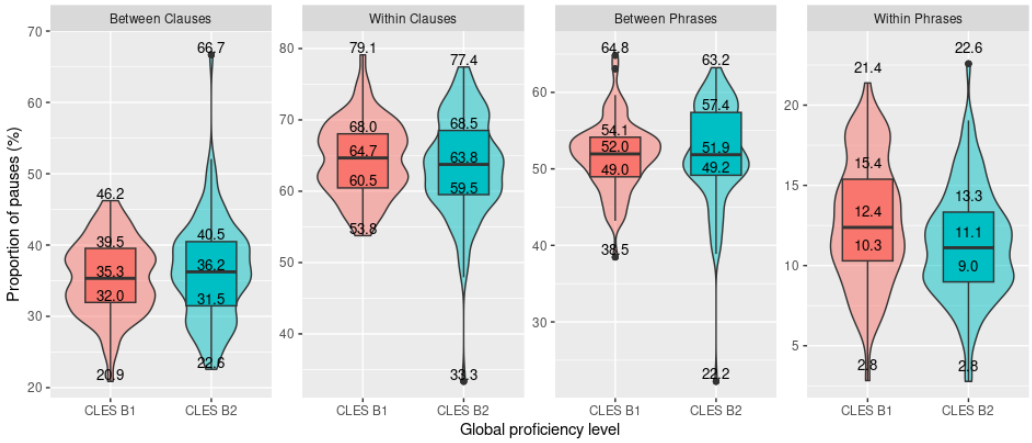


FIG. 12.5 : Proportion de pauses par type de frontière pour 100 pauses

Sachant que les locuteurs B1 ont tendance à faire plus de pauses en général, il n'est pas surprenant d'observer des fréquences plus élevées quelque soit le type de frontière. Pour neutraliser ce facteur, analysons le nombre de pauses de chaque type en fonction du nombre total de pauses par locuteur (cf. figure 12.5). On constate que la proportion de pauses inter-propositions est maintenant légèrement plus faible pour les B1 que pour les B2, mais sans différence significative (*ns.*, médianes à 35 et 36 %, $\Delta = -0{,}069$), de même pour la proportion de pauses inter-syntagmes (*ns.*, médianes à 52 %, $\Delta = -0{,}069$). La proportion de pauses intra-syntagmes est quant à elle significativement plus élevée pour les B1, mais toujours avec une taille d'effet relativement limitée ($p < 0{,}05$, médianes à 12 et 11 %, $\Delta = 0{,}216$ (small) $IC = [0{,}037; 0{,}382]$).

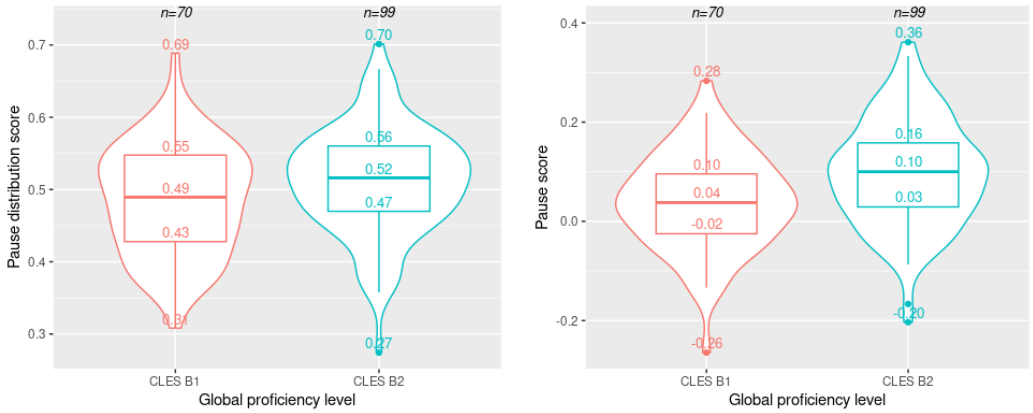
La différence entre les locuteurs B1 et B2 en terme de distribution syntaxique

des pauses semble donc se jouer principalement au niveau de la fréquence des pauses en frontière de bas niveau. Toutefois, la différence que nous observons ici reste limitée.

1.3 Score de distribution syntaxique

Calculons maintenant le score de distribution syntaxique des pauses (SDS) en fonction des catégories de pauses inter-proposition, inter-syntagme et intra-syntagme. Rappelons que calculer ce score revient à compter le nombre de pauses de chaque type et le normaliser par le nombre total de pause par locuteur, puis en faire un somme pondérée en pénalisant les pauses situées à l'intérieur des syntagmes (-1) et en favorisant les pauses inter-syntagmes (+0,5) et inter-propositions (+1). Le score obtenu permet ainsi de comparer la tendance de distribution syntaxique des pauses par locuteur : plus il est élevé, plus les pauses ont tendance à être placées en frontière de haut niveau. Les résultats indiquent que les locuteurs B1 obtienne en moyenne un score plus faible que les locuteurs B2, mais la taille d'effet reste assez limitée ($p < 0{,}05$, médianes à 0,49 et 0,52, $\Delta = -0{,}198$ (small) $IC = [-0{,}365; -0{,}019]$, cf. figure 12.6a).

On peut également calculer le SDS en fonction non plus du niveau des constituants, mais directement du nombre de constituants qui se ferment ou qui s'ouvrent à l'endroit où survient la pause. Il devient ainsi possible de prendre en compte l'imbrication des syntagmes ou des propositions les uns aux autres et permet d'avoir plus de souplesse dans le paramétrage du calcul. La différence entre les locuteurs B1 et B2 apparaît cette fois plus nettement ($p < 0{,}001$, médianes à 0,04 pour B1 et 0,10 pour B2, $\Delta = -0{,}301$ (small) $IC = [-0{,}455; -0{,}13]$, cf. figure 12.6b).



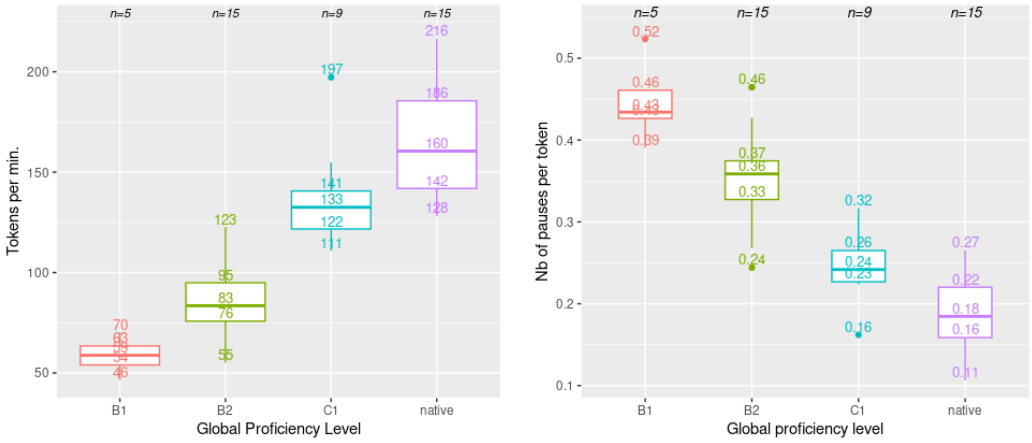
(a) Score basé sur le niveau des constituants (proposition ou syntagme)

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

FIG. 12.6 : *Scores de distribution syntaxique des pauses par locuteur (corpus CLES-FR,* 180 ms-2 s*)*

1.4 Corpus CLES-JP et CLES-EN

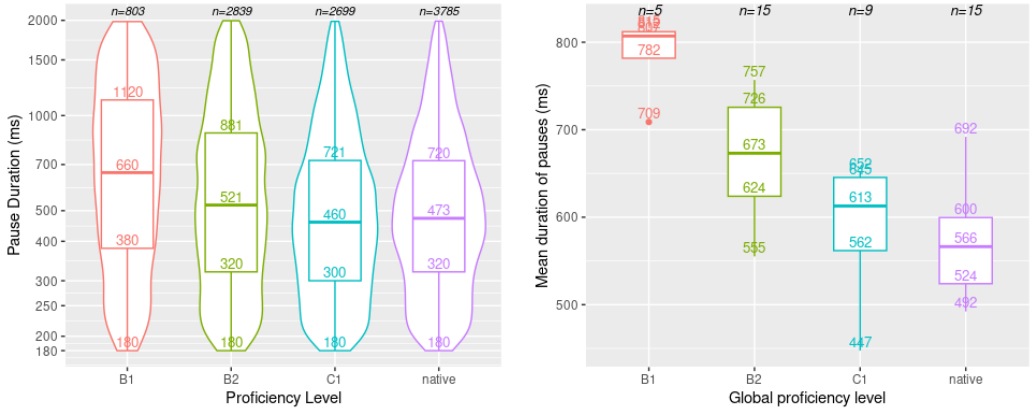
Du côté des corpus CLES-JP et CLES-EN, 21 631 intervalles pour les premiers et 20 486 pour les seconds ont été analysés. Cette fois-ci, la différence de durée des pauses entre les locuteurs B1 et B2 est significative ($p < 0{,}001$) quelque soit le seuil minimal (180 ou 250 ms) et maximal (2 ou 5 s), avec toutefois une différence faible entre les deux groupes (le Δ de Cliff variant 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 compte alors respectivement 6 341 pauses pour les locuteurs japonophones (803 pour les 5 locuteurs B1, 2839 pour les 15 B2, 2699 pour les 9 C1) et 3 785 pauses pour les 15 locuteurs natifs. Comme pour le corpus CLES-FR, plus le niveau du locuteur est élevé plus le débit de parole est élevé également (cf. figure 12.7a), amenant à une quantité de parole, et par extension de pauses observées, plus importante. Ramené au nombre de pauses par mot, la tendance est en revanche inverse, et fortement contrastée entre les niveaux : les locuteurs B1 se situent autour de 43 pauses pour 100 mots, les B2 à 36 pauses, les C1 à 24 et les locuteurs natifs à 18 (cf. figure 12.7b).



(a) Débit de parole *(b) Fréquence des pause*

Fig. 12.7 : *Débit de parole (gauche) et fréquence des pauses (droite) par locuteur dans les corpus CLES-JP et CLES-EN (*180 ms-2 s*)*

La figure 12.8a montre la distribution des durées de pauses pour chaque groupe. La différence de durée moyenne des pauses par locuteurs apparaît fortement contrastée entre les niveaux. La durée moyenne se situe entre 709 ms et 815 ms, médiane à 807 ms, tandis que la médiane des locuteurs B2 est à 673 ms, celle des C1 à 613 ms et celle des natifs à 566 ms (cf. figure 12.8b). La différence entre les B1 et B2 est significative ($p < 0{,}01$, $\Delta = 0{,}867$ (large) $IC = [0{,}338; 0{,}98]$), ainsi qu'entre les locuteurs B1,B2 et les locuteurs natifs ($p < 0{,}001$, $\Delta = 0{,}82$ (large) $IC = [0{,}54; 0{,}937]$).

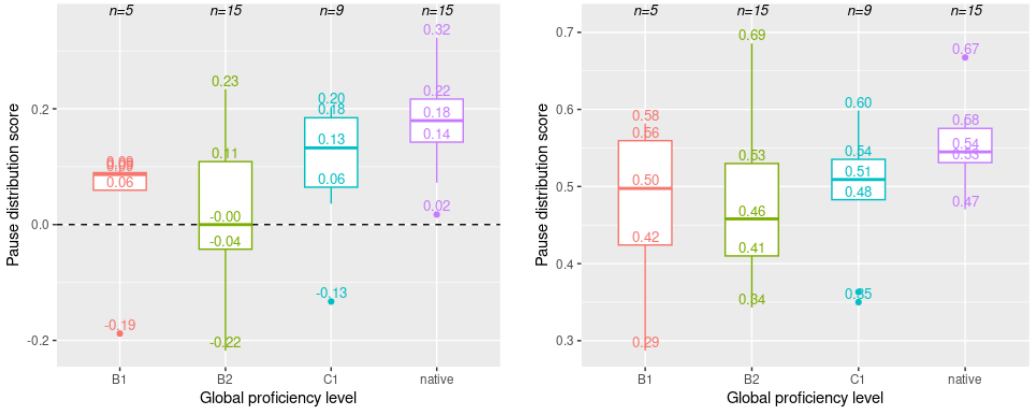


(a) Distribution des durées de pauses *(b) Durée moyenne des pauses par locuteur*

FIG. 12.8 : *Durées des pauses dans les corpus CLES-JP et CLES-EN (*180 ms-2 s*)*

En termes de distribution syntaxique des pauses, là aussi la différence apparaît plus contrastée que pour les locuteurs francophones, mais suit les mêmes tendances. Les locuteurs B1 font en moyenne significativement plus de pauses que les B2 en frontière de propositions ($p < 0{,}01$, médianes à 58 et 46 %, $\Delta = 0{,}840$ (large) $IC = [0{,}459; 0{,}96]$), de syntagmes ($p < 0{,}05$, médianes à 43 et 33 %, $\Delta = 0{,}657$ (large) $IC = [0{,}115; 0{,}876]$) mais la différence n'est pas significative au niveau intra-syntagme, probablement à cause du manque de données pour les B1 (*ns*, médianes à 30 et 23 %, $\Delta = 0{,}52$ (large) $IC = [-0{,}037; 0{,}83]$). Probablement pour les mêmes raisons, la différence de proportion de pauses par type n'est significative à aucun niveau. La comparaison entre les B1+B2 et les locuteurs natifs donne quant à elle des résultats plus fiables : les locuteurs natifs font significativement plus de pauses en frontières de proposition ($p < 0{,}05$, médianes à 35 pour les L2 et 37 % pour les L1, $\Delta = -0{,}4$ (medium) $IC = [-0{,}687; -0{,}004]$), et moins de pauses à l'intérieur des syntagmes ($p < 0{,}05$, médianes à 13 pour les L2 et 10 pour les L1, $\Delta = 0{,}483$ (large) $IC = [0{,}087; 0{,}747]$).

Le score de distribution syntaxique des pauses montre également une différence non-significative entre les B1 et B2, qu'il soit basé sur les niveaux de constituants ou sur l'importance des frontières. En revanche, la différence L1/L2 est quant à elle significative : $p < 0{,}01$, médianes à 0,48 pour les L2 et 0,54 pour les L1, $\Delta = -0{,}527$ (large) $IC = [-0{,}777; -0{,}131]$ pour le score basé sur les niveaux de constituants, et $p < 0{,}001$ médianes à 0,06 et 0,18, $\Delta = -0{,}707$ (large) $IC = [-0{,}891; -0{,}32]$ pour le score basé sur la profondeur des frontières syntaxiques (cf. figure 12.9).



(a) Score basé sur le niveau des constituants *(b) Score basé sur la profondeur des frontières*

Fig. 12.9 : *Score de distribution syntaxique des pauses par locuteur, basé sur le niveau des constituants (propositions et syntagmes, à gauche) ou le niveau de profondeur des frontières syntaxiques (à droite) (corpus CLES-JP et CLES-EN,* 180 ms-2 s*)*

12.2 Accentuation lexicale

Nous avons vu dans le chapitre 3 que la précision de position de l'accent lexical est souvent corrélée avec le jugement de compréhensibilité du locuteur. Toutefois, la plupart des études que nous avons citées s'appuient souvent sur de la parole lue ou des annotations manuelles de position de l'accent. Nous avons souhaité savoir si des mesures automatiques permettent de caractériser les patterns accentuels des locuteurs en parole spontanée, et si une différence significative est observable entre les niveaux B1 et B2. En outre, nous nous intéresserons à la qualité de l'accentuation en termes de degré de contraste prosodique entre les syllabes au niveau de la F_0, de l'intensité et de la durée.

2.1 Données analysées

L'ensemble des segments de paroles du corpus CLES-FR analysés par PLSPP compte 68 515 tokens, avec un nombre de tokens par locuteur légèrement inférieur pour les B1 par rapport à celui des B2 (médianes à 376 contre 422, non significatif). Les mesures d'accentuation syllabique que nous avons effectuées ici portent exclusivement sur les mots polysyllabiques lexicaux (noms communs, verbes, adjectifs et adverbes), cependant tous n'ont pas nécessairement été annotés par PLSPP. On compte un total de 14 873 mots polysyllabiques lexicaux, significativement plus nombreux chez les

locuteurs B2[1]. Parmi eux, le nombre de mots annotés par PLSPP n'est que de 6 468, soit seulement 43 % des mots initialement ciblés. Nous reviendrons sur ce constat dans le chapitre 14. L'ensemble des analyses de cette section portent sur ces 6 468 mots, nous les appellerons « mots analysés ».

Nombre de mots analysés par locuteur Si le nombre absolu de mots analysés par locuteur est significativement plus élevé pour les B2 ($p < 0{,}01$, médianes à 32 pour B1 contre 41 pour B2), il ne l'est pas ramené au nombre de tokens (médianes à 9 et 10 %) ni au nombre de mots polysyllabiques lexicaux (médianes à 42 et 43 %). On ne peut donc pas dire que les locuteurs B2 utilisent proportionnellement plus de mots polysyllabiques que les B1, ni que les mots qu'ils produisent sont mieux reconnus par PLSPP que ceux des locuteurs B1.

Caractéristiques des mots analysés Ce sont en grande partie des noms communs (57% des mots, doublons compris), suivis par des verbes (19%), des adjectifs (12%) et des adverbes (12%). Ils sont majoritairement composés de deux syllabes (73%), mais on trouve également des mots de 3 syllabes (21%), 4 syllabes (5%), 5 et 6 syllabes (moins de 1%). Pour limiter l'influence potentielle de l'accent secondaire, et étant donné que les mots de plus de 3 syllabes représentent moins de 6 % des mots analysés, nous focaliserons les analyses de cette section sur les mots de 2 à 3 syllabes ($n = 6\ 002$). Parmi eux, la syllabe qui porte l'accent primaire théorique (attendu, prescrit) est la syllabe initiale dans 74 % des cas (4 432 mots), médiale dans 13 % (791), et finale dans 13 % (779) des cas (cf. tableau 12.2). Pour les mots à 2 syllabes, 84 % d'entre eux sont accentués à l'initiale contre seulement 16 % en finale. Pour les mots à 3 syllabes, la majorité est accentuée en médiale (58%), puis en initiale (38%) et en finale (4%).

[1] $p < 0{,}05$, médianes à 75 pour B1 et 94 pour B2, $\Delta = -0{,}232$ (small) $IC = [-0.396; -0.052]$; ramené au nombre de tokens par locuteur : $p < 0{,}01$, médianes à 21 % et 23 %, $\Delta = -0{,}238$ (small) $IC = [-0.397; -0.065]$.

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

2.2 Patterns accentuels observés

Intéressons-nous maintenant aux patterns accentuels produits par les locuteurs. Nous parlerons maintenant de « syllabe proéminente » pour qualifier la syllabe identifiée par PLSPP comme acoustiquement proéminente, et qui sera ainsi perçue, en théorie, comme syllabe accentuée par l'auditeur. Le terme « accent théorique » fait quant à lui référence à l'accent prescrit par le dictionnaire de référence, ici le *CMU Pronouncing Dictionary* (version 0.7b). On observe que 67 % des syllabes proéminentes sont situées en finale, contre 26 % en initiale et 7 % en médiale (cf. tableau 12.2). De manière générale, le pourcentage de mots dont la syllabe proéminente correspond à la position de l'accent primaire théorique est relativement bas (36% à travers le corpus, 32% pour les mots produits par des locuteurs B1, et 38% pour ceux produits par les B2). Comme le nombre de mots analysés varie en fonction des locuteurs, il est plus pertinent de considérer la proportion de mots correctement accentués par locuteur : nous parlerons de « score de position de l'accent ». Les scores individuels sont très variés, ils s'étendent de 0 à 64,8 % (médiane à 34,6 %).

La figure 12.10 présente la distribution des scores par locuteur en fonction du niveau. On constate que les scores se chevauchent largement entre B1 et B2, mais présentent toutefois une différence significative ($p < 0{,}01$, médianes respectives à 30,8 et 36,8 %), avec toutefois une taille d'effet limitée ($\Delta = -0{,}275$ (small) $CI = [-0{,}432; -0{,}102]$). Par ailleurs, lorsque l'on considère le score obtenu en fonction de la position de l'accent théorique, on constate que le score des locuteurs B2 est effectivement plus élevé pour les mots à accent initial ($p < 0{,}01$, médianes à 25 pour B1 et 32 pour B2, $\Delta = -0{,}269$ (small) $CI = [-0{,}428; -0{,}094]$) et médial (*ns.*, médianes à 30 et 36, $\Delta = -0{,}158$ (small) $CI = [-0{,}329; -0{,}024]$),

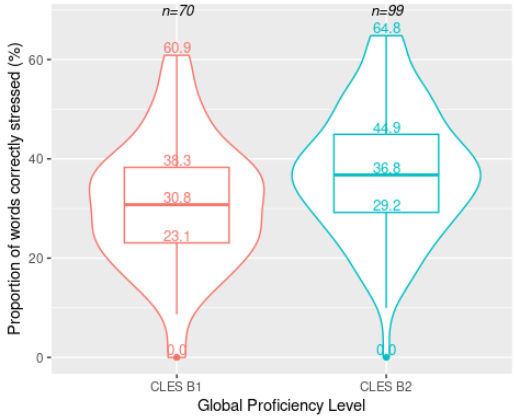


Fig. 12.10 : *Score de position de l'accent par locuteur, sur 6 468 mots lexicaux de 2 à 3 syllabes (corpus CLES-FR)*

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. 12.3 : *Liste des mots de plus de 50 occurrences dans le corpus CLES-FR*

mais pas pour les mots à accent final (*ns.*, médianes 77, $\Delta = -0{,}032$ (negligible) $CI = [-0{,}148; 0{,}210]$). Cela laisse entendre que la différence entre les deux groupes de locuteurs se joue au niveau du taux d'accentuation en initiale et en médiale, plutôt qu'en finale. Le deuxième constat est que les locuteurs B1 semblent présenter des scores plus variés que les B2 (malgré un écart type identique de 12,8 %).

Le tableau 12.3 liste les 18 mots les plus fréquents parmi les mots analysés (plus de 50 occurrences). On peut constater que les mots les plus fréquents sont loin d'être les mieux maîtrisés (“*students*” accentué en initiale seulement 24,6% des fois sur 187 occurrences ; “*people*” 20,8% sur 173 occurrences, “*teacher*” 19% (89) et “*teachers*” 10,7% (75) seulement).

2.3 Contraste prosodique

Nous nous intéressons dans cette section au degré de contraste entre les syllabes. Il ne s'agit plus simplement de savoir si la proéminence est réalisée sur la syllabe attendue ou non, mais de mesurer à quel point cette syllabe se démarque des autres sur le plan prosodique. Pour cela, nous calculons la différence entre la valeur acoustique normalisée P_s de la syllabe censée être accentuée et la moyenne $\overline{P_u}$ des autres syllabes

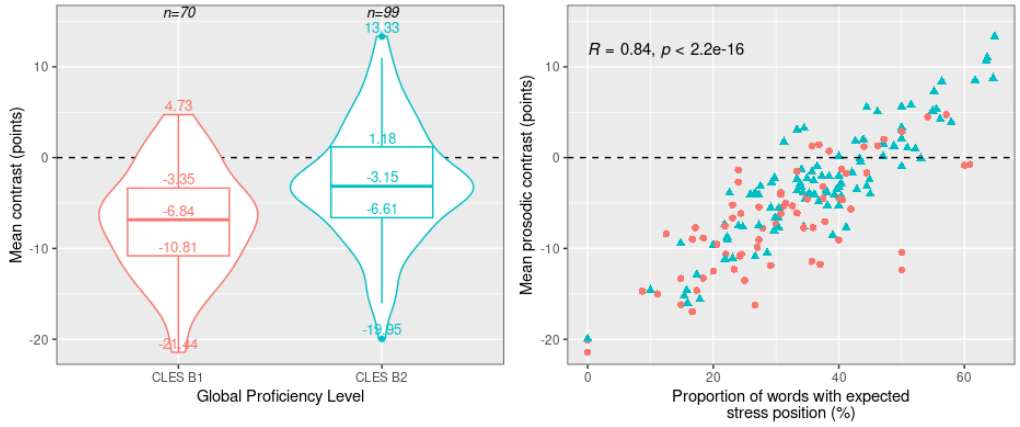


FIG. 12.11 : Contraste prosodique moyen entre P_s et $\overline{P_u}$ dans le corpus CLES-FR, à gauche en fonction du niveau du locuteur ; à droite en fonction du score de position de l'accent

du mot. La valeur obtenue varie entre -100 et 100, et indique à quel point le contraste est marqué entre la syllabe censée être accentuée et les autres syllabes. Si la valeur est positive, la syllabe proéminente correspond à la syllabe censée être accentuée, plus la valeur est élevée, plus cette syllabe est contrastée par rapport aux autres. Nous appellerons « contraste moyen » la valeur moyenne d'un contraste sur l'ensemble des mots produits par un locuteur. Il pourra s'agir du contraste de F_0, d'intensité, de durée, ou à défaut la moyenne des trois.

Le contraste moyen par locuteur dans le corpus CLES-FR s'étend de -21,44 à 13,33, avec une médiane à -4,11 indiquant que dans l'ensemble, les locuteurs ont tendance à accentuer la mauvaise syllabe. Si là encore, les locuteurs B1 et B2 se chevauchent largement, la différence entre les deux distributions est significative et plus prononcée qu'avec le simple score de position de la section précédente ($p < 0{,}001$, médianes à -6,84 pour les locuteurs B1 et -3,15 pour les B2, $\Delta = -0{,}389$ (medium) $CI = [-0{,}534; -0{,}222]$, cf. figure 12.11 gauche). On constate également, et sans surprise, que plus le score de position est élevé, plus le contraste moyen est grand, indiquant que les locuteurs qui ont tendance a bien placer l'accent produisent également un contraste plus important entre la syllabe accentuée et les autres syllabes ($R = 0{,}84$, $p < 0{,}001$, cf. figure 12.11 droite).

Quelle est la dimension prosodique pour laquelle le contraste moyen augmente le plus entre B1 et B2 ? En d'autres mots, quelle est la dimension qui représente le mieux la différence entre les deux groupes de niveau ? Considérons le contraste moyen pour chaque dimension prise séparément (figure 12.12). D'après le delta de Cliff, la plus grande différence entre B1 et B2 se situe au niveau du contraste d'intensité ($p < 0{,}001$, médianes à -5,82 en B1 et 0,77 en B2, $\Delta = -0{,}450$ (medium)

$CI = [-0{,}590; -0{,}283]$), puis vient le contraste de hauteur avec une taille d'effet un peu moins importante ($p < 0{,}001$, médianes à -7,38 et -1,44, $\Delta = -0{,}317$ (small) $CI = [-0{,}475; -0{,}138]$). Quant au contraste de durée de syllabe, il ne présente pas de différence significative entre les deux niveaux (*ns.*, médianes à -9,84 et -9,36, $\Delta = -0{,}018$ (negligible) $CI = [-0{,}196; 0{,}162]$).

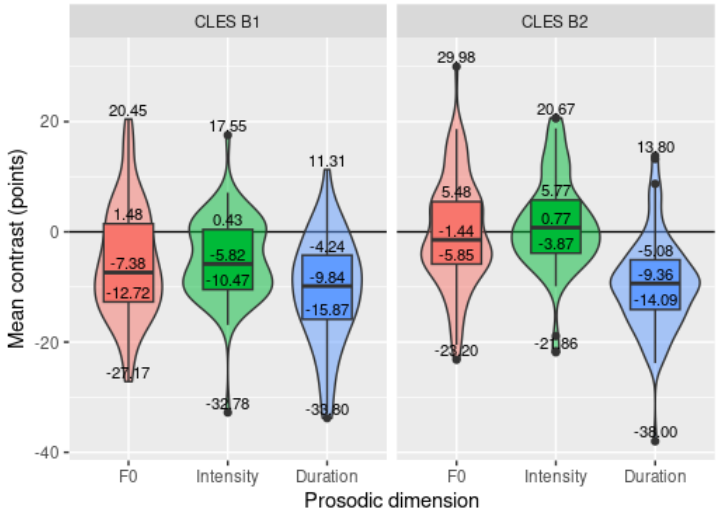


FIG. 12.12 : Contraste moyen par dimension prosodique par locuteur (corpus CLES-FR)

La figure 12.13 présente une visualisation du profil d'accentuation de deux locuteurs SpeakerA et SpeakerB. La figure indique le contraste prosodique moyen du locuteur pour chaque dimension (multidimensionnel, F_0, intensité et durée ; le premier étant la moyenne des trois suivants). Le contraste est représenté par deux cercles, le premier représente P_s (la syllabe censée être accentuée), et le second représente $\overline{P_u}$ (la moyenne des autres syllabes du mots). La valeur inscrite dans les cercles et leur taille représentent la valeur prosodique normalisée, en centile. SpeakerA est représentatif d'un locuteur de haut niveau en termes d'accentuation : son score de position est de 64,8 %, et son contraste moyen (multidimensionnel) est de 13,3. On peut voir que P_s est bien contrastée au niveau de la F_0 (contraste de 30) et de l'intensité (17), mais pas au niveau de la durée (-5). SpeakerB a quant à lui un score de position de seulement 14,8 %, et un contraste moyen de -9,4. On peut voir que c'est $\overline{P_u}$ qui a tendance à être accentué (donc la mauvaise syllabe), avec une influence en premier lieu de la durée (-21), puis de la F_0 (-11), tandis que l'intensité ne semble pas mobilisée (+2). Des tendances similaires se retrouvent chez les locuteurs au score et au contraste moyen élevé (prépondérance de la F_O et de l'intensité, neutralisation de la durée) et chez les locuteurs au score et contraste faible (impact de la durée et de la F_0, pas ou peu de contraste d'intensité).

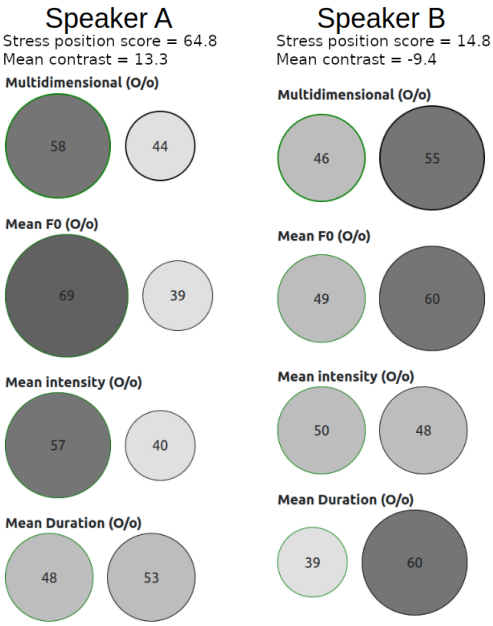


FIG. 12.13 : Contrastes prosodiques de deux locuteurs au profil différent
Écouter un extrait : Speaker A et Speaker B.

2.4 Corpus CLES-JP et CLES-EN

Quatre-cents-trente segments de parole ont été extraits des corpus CLES-JP et CLES-EN, totalisant 41 714 tokens (21 356 pour les 29 locuteurs japonophones et 20 358 pour les 15 locuteurs natifs). La proportion de mots polysyllabiques lexicaux par token est similaire dans les deux corpus (médianes à 23 % pour les locuteurs B1/B2 du CLES-JP, et 22 % pour les locuteurs natifs), cependant, et contrairement à ce à quoi on aurait pu s'attendre, la proportion de mots annotés par PLSPP est plus grande pour les locuteurs japonophones que pour les locuteurs natifs ($p < 0{,}001$, médianes à 41 % pour les locuteurs B1 et B2 ensemble et 30 % pour les natifs, $\Delta = 0{,}753$ (large) $IC = [0{,}410; 0{,}910]$). Les mots semblent donc en moyenne mieux reconnus quand les locuteurs ne sont pas natifs. Au total, 1 913 mots ont été annotés pour CLES-JP et 1 354 pour CLES-EN. La proportion des catégories grammaticales et des nombres de syllabes est sensiblement similaire entre les deux corpus, comme l'indique le tableau 12.4. Comme pour le corpus CLES-FR, nous nous focaliserons sur les mots de 2 à 3 syllabes, soit environ 96 % des mots annotés.

La figure 12.14 présente les scores de position de l'accent par locuteur. Le premier constat est qu'on n'observe pas d'amélioration significative entre les niveaux des locuteurs japonophones. Les 5 locuteurs de niveau B1 obtiennent un score variant entre 35,3 et 70,6% (médiane à 45,8 %), et les 15 locuteurs B2 entre 31,7 et 72,7 % (médiane

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

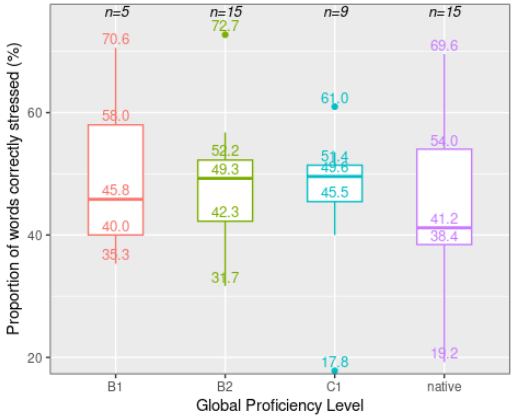


Fig. 12.14 : *Scores de position par locuteur, sur 3 014 mots lexicaux de 2 à 3 syllabes (corpus CLES-JP et CLES-EN)*

à 49,3 %, différence non significative). Les 9 locuteurs C1 n'obtiennent pas de résultats sensiblement supérieurs. Le deuxième constat est que les locuteurs natifs obtiennent des résultats très variables, s'étalant de 19,2 et 69,6 %, et en moyenne inférieurs à ceux des locuteurs japonophones (médianes à 49,3 % pour CLES-JP et 41,2 % pour CLES-EN, différence non significative, $\Delta = 0{,}170$ (small) $IC = [-0{,}223; 0{,}516]$). La détection de la syllabe proéminente semble influencée par un facteur indépendant du niveau de compétence en langue.

Le tableau 12.5 présente le nombre de mots par position d'accent théorique (attendu) ou observé (syllabe proéminente détectée par PLSPP). On peut constater que, si on s'attend à avoir environ 80 % de mots accentués en initiale, 7 % de mots en médiale et 10 % de mots en finale dans les deux corpus, PLSPP détecte une proéminence sur l'initiale dans seulement 40 % des mots, 5 % en médiale et presque 60 % en finale, pour les japonophones comme pour les anglophones natifs. Un grand nombre de proéminences sont donc détectées sur la dernière syllabe des mots, même lorsque elle n'est pas censée être accentuée, et ce quelque soit la langue maternelle du locuteur.

Lorsque l'on mesure le contraste prosodique entre la syllabe censée être accentuée P_s et la moyenne des autres syllabes $\overline{P_u}$, on constate que P_s est systématiquement plus

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

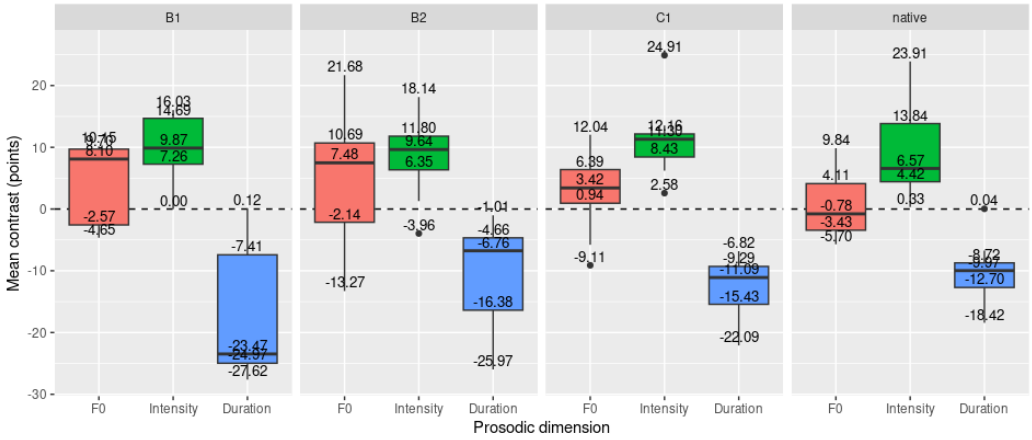


Fig. 12.15 : *Contraste moyen par locuteur entre P_s et $\overline{P_u}$ (corps CLES-JP, CLES-EN)*

haute en intensité, mais plus courte en durée que $\overline{P_u}$ (cf. figure 12.15. La F_0 a quant à elle tendance à être plus haute mais le contraste est moins fort. On observe les mêmes tendances que pour le corpus CLES-FR (cf. figure 12.11), mais les valeurs apparaissent plus extrêmes : les contrastes d'intensité et de hauteur sont assez marqués et majoritairement en adéquation avec la position de l'accent théorique ; le contraste de durée est quant à lui fortement négatif. Ce contraste négatif indique qu'une syllabe qui n'est pas censée être accentuée est plus longue que les autres, suffisamment pour rendre le contraste moyen des trois dimensions négatif, et induire en erreur la détection de la proéminence. Le problème semble être au niveau de la durée de la syllabe finale, souvent assez longue pour faire pencher la balance pour une proéminence en finale.

En analysant la corrélation entre le contraste par dimension prosodique et le score de position de l'accent des locuteurs japonophones et anglophones, on observe que les deux groupes de locuteurs ne se comportent pas de la même manière : le score des locuteurs natifs apparaît très corrélé avec le contraste d'intensité ($R = 0{,}95$, $p < 0{,}001$), tandis qu'il l'est moins pour les locuteurs japonophones ($R = 0{,}60$, $p < 0{,}001$). Le constat est inverse pour la F_O : $R = 0{,}54$ ($p < 0{,}05$) pour les locuteurs natifs mais $R = 0{,}65$ ($p < 0{,}001$) pour les locuteurs du corpus CLES-JP. Le contraste de durée, quant à lui, n'est significatif que pour les locuteurs natifs ($R = 0{,}73$, $p < 0{,}01$; contre

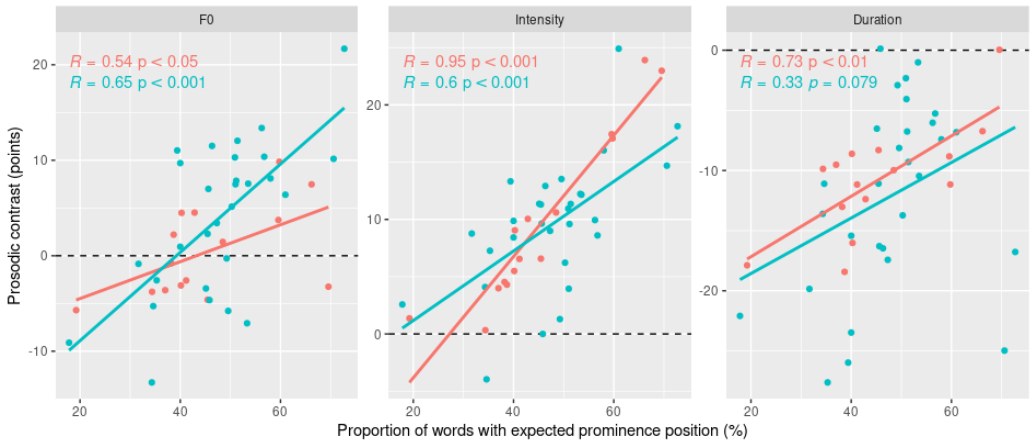


FIG. 12.16 : Corrélation entre le contraste moyen par dimension prosodique et le score de position de l'accent par locuteur (corpus CLES-JP en bleu et CLES-EN en rouge)

$R = 0{,}33$, $p = 0{,}08$ pour CLES-JP). Comment interpréter ces résultats ? Il semble possible de formuler l'hypothèse que le paramètre d'intensité est le plus fiable des trois dimensions prosodiques mesurées par PLSPP. Les locuteurs natifs qui produisent un contraste d'intensité important entre les syllabes tendent à obtenir un meilleur score global d'accentuation. De leur côté, les locuteurs japonophones semblent plutôt s'appuyer sur le contraste de hauteur : il s'agit de la corrélation la plus forte avec le score de position, mais elle est également plus forte et le contraste est plus élevé que pour les locuteurs natifs. Cela est peut-être dû au fait que les natifs accentuent parfois par une chute de F_0, ou qu'ils ont tendance à dévoiser certaines voyelles, perturbant les mesures de hauteur. Quant à la durée, elle semble aléatoire chez les locuteurs japonophones (pas de corrélation avec le score de position), ce qui pourrait être dû à une tendance à hésiter en fin de mot et ainsi allonger la dernière syllabe. Le contraste de durée apparaît pourtant fortement corrélé au score de position pour les locuteurs natifs, bien que négatif, indiquant que les locuteurs qui produisent une syllabe accentuée plus longue que les syllabes non accentuées ont bien tendance à obtenir un meilleur score de position de l'accent.

12.3 Conclusion

Nous avons vu dans ce chapitre que le nombre de pauses et leur durée moyenne dépendent du niveau de compétence en langue du locuteur, et que, bien que les locuteurs B1 fassent plus de pauses que les B2 à tous les niveaux syntaxiques (inter-proposition, inter-syntagme et intra-syntagme), ils ont tendance à faire proportion-

nellement plus de pauses intra-syntagmes que les locuteurs B2, mais pas significativement plus de pauses inter-propositionnelles – idem pour les locuteurs japonophones vis-à-vis des locuteurs natifs. Les scores de distribution syntaxique des pauses confirment que plus le niveau du locuteur augmente, plus les pauses ont tendance à être placées en frontière de haut niveau syntaxique, avec les locuteurs natifs en tête de classement.

L'analyse des annotations automatiques de proéminence syllabique montrent que les locuteurs B2 ont tendance à mieux positionner l'accent et produire un meilleur contraste acoustique entre la syllabe accentuée et les autres syllabes du mot. On constate toutefois une grande variabilité dans les mesures individuelles, le score de position de l'accent s'étalant de 0 à 65 % selon les locuteurs du corpus CLES-FR et présentant un large chevauchement entre les locuteurs B1 et B2 malgré une différence significative. Les annotations font ressortir une nette influence des patterns accentuels de la langue maternelle des locuteurs. Les francophones ont généralement tendance à augmenter la F_0 et allonger la syllabe finale, tandis que l'intensité reste stable, quelle que soit la position de l'accent théorique. 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. Du côté des locuteurs japonophones, la tendance à accentuer la syllabe finale est beaucoup moins marquée, avec de meilleurs scores de position de l'accent de manière générale (de 17,8 à 72,7 %), et une forte mobilisation de la F_0 et de l'intensité dès le niveau B1. Il n'a malheureusement pas été possible de faire ressortir de différence significative entre les niveaux étant donné le faible nombre de locuteurs du corpus. Enfin, les analyses de locuteurs natifs ont fait ressortir une forte tendance à l'allongement de la syllabe finale et un contraste moyen de F_0 limité, conduisant à des scores de position globalement bas (entre 19,2 et 69,7 %), et en moyenne inférieurs à ceux des locuteurs japonophones. Le contraste d'intensité s'est révélé quant à lui plus important que pour les locuteurs non-natifs et très corrélé avec le score de position de l'accent. Nous revenons sur les limites de l'outil de mesure dans le chapitre 14.

Chapitre 13

Mesure de l'impact du rythme sur la compréhensibilité

Introduction...

13.1 Développement de Dynamic Rater

Une application web appelée Dynamic Rater[1] a été développée spécifiquement pour les besoins de cette étude. Comme présenté dans le chapitre 8, elle s'inspire largement du fonctionnement du logiciel Idiodynamic (MacIntyre, 2012), mais propose un protocole d'évaluation plus simple et permet la 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 et illustrée en Annexe B. Nous nous concentrons ici sur la fonctionnalité d'évaluation dynamique de la compréhensibilité.

Après une phase d'entraînement, 16 stimuli audio sont présentés aléatoirement aux participants, qui doivent signaler lors de l'écoute, en cliquant sur le bouton *I'm struggling*, qu'il perçoivent une difficulté à comprendre le locuteur, quelque soit la raison (cf. figure 13.1). À chaque clic, une barre verticale s'affiche à l'endroit concerné sur la waveform, de manière à confirmer au participant que l'action a bien été prise en compte. L'application enregistre par ailleurs le timestamp du clic, qui sera envoyé sous forme d'une liste de timestamps au serveur lors de la validation.

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

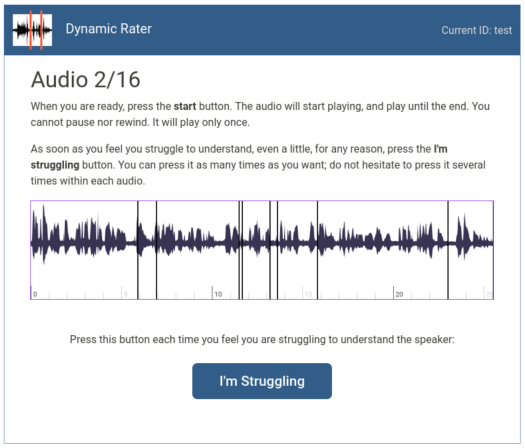


FIG. 13.1 : Aperçu de l'écran d'évaluation dynamique de Dynamic Rater

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 champs 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, une requête est envoyée au serveur avec la liste des clics et l'évaluation globale, de manière à enregistrer les résultats au fur et à mesure.

13.2 Comportements des évaluateurs

Le temps moyen de passation de l'expérience dans sa globalité est de 29 minutes. L'évaluation des 16 stimuli audio 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 (avec 4 évaluateurs excédant 45 min). Comme prévu, une grande variabilité de comportements a été observée parmi les évaluateurs, avec un nombre total de clics enregistrés allant de 12 à 272 à travers les 16 enregistrements (moyenne de 76,7 clics par évaluateur, écart type de 48,65). Cinq évaluateurs ont présenté une fréquence de clics particulièrement élevée, totalisant plus de 120 clics chacun.

Malgré cette grande variation de fréquence de clics, une tendance claire à cliquer dans les mêmes zones est observée, se traduisant par des pics de clics relativement bien contrastés comme l'illustrent les figures 13.2 et 13.3.

Le coefficient de corrélation intra-classe indique un accord absolu de 0,97 et une cohérence moyenne inter-annotateur de 0,98 pour les trois évaluations globales, soit

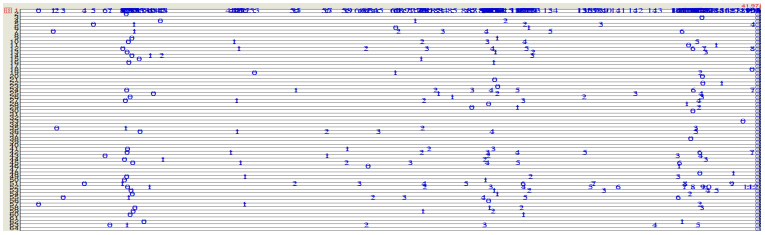


FIG. 13.2 : Aperçu des clics enregistrés par les 60 participants pour l'enregistrement n°5 (format TextGrid, un point représente un clic, un évaluateur par tier, la première tier est la somme de l'ensemble des clics)

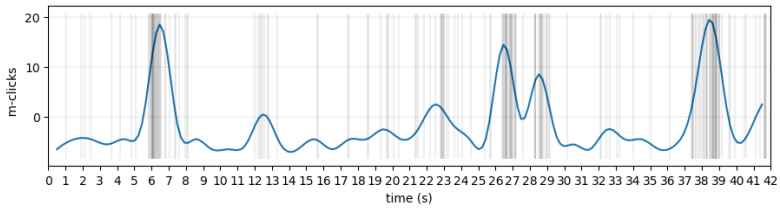


FIG. 13.3 : Somme des m-clics sur une fenêtre glissante d'une seconde, pour l'enregistrement n°5 (les clics bruts sont représentés par une barre verticale)

tout à fait satisfaisante.

13.3 Analyse des patterns de clics

Nous présentons dans cette section les résultats de l'analyse des variations du nombre de 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 13.4 (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, enfin, le nombre de m-clics semble stagner autour de 0, n'indi-

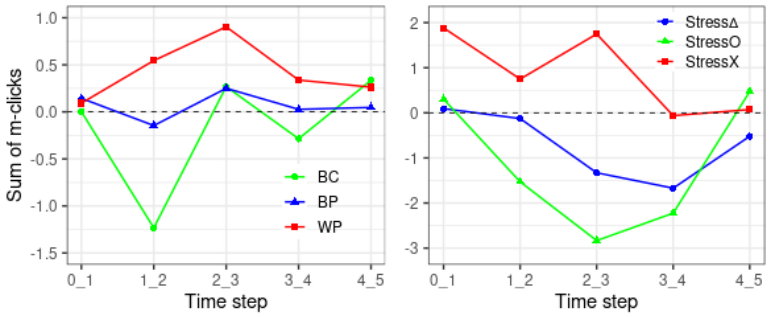


FIG. 13.4 : Nombre de m-clics moyen sur les 5 secondes suivant l'onset d'une pause (gauche) ou d'un mot pattern accentuel (droite) ; les valeurs positives indiquent une activité de clics supérieure à la moyenne

	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. 13.1 : 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, $S_w > 0{,}2$) et incorrects (StressX, $S_w < -0{,}2$), et coefficient de corrélation entre le nombre de m-clics et la valeur de S_w (— : non significatif, * : $p < .05$, ** : $p < .01$)*

quant 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 13.1.

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, $S_w < -0{,}2$, 27% des mots, n=37) reste globalement supérieure à celle des mots jugés corrects (StressO, $S_w > 0{,}2$, 17% des mots, n=23) ou ambigus (StressΔ, 57%, n=79), on observe une augmentation locale entre 2 et 3 secondes après l'onset du mot, mais une diminution évidente des clics après les mots 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 13.1.

Comme le score accentuel S_w est une valeur continue, nous avons également mesuré la corrélation entre celui-ci et le nombre de m-clics observés dans chaque fenêtre.

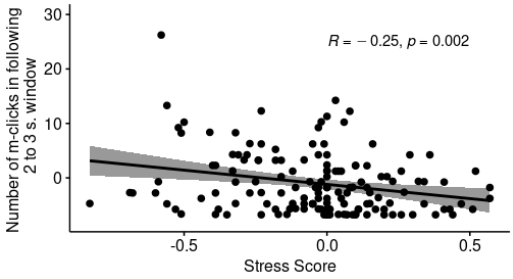


FIG. 13.5 : Projection des 139 mots cibles en fonction de leur score accentuel S_w et du nombre de m-clics enregistrés dans la fenêtre de 2 à 3 secondes après l'onset du mot

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 13.1 et figure 13.5).

13.4 Évaluations globales

Qu'en est-il des évaluations globales des enregistrements ? Les scores de qualité de prononciation, de fluidité et de compréhensibilité apparaissent de manière générale très corrélés entre eux, comme le montre la figure 13.6. On peut voir tout d'abord 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, $\Delta = -0{,}328$ (small) $IC = [-0{,}367; -0{,}288]$, figure 13.7 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 moyen 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, $\Delta = -0{,}202$ (small) $IC = [-0.243; -0.16]$, figure 13.7 milieu). Notons par ailleurs que les locutrices (n=9) ont tendance à être mieux notées que les locuteurs (n=7, $p < 0{,}001$, médianes à -0,28 pour les hommes et +0,28 pour les femmes, $\Delta = -0{,}245$ (small) $IC = [-0{,}286; -0{,}204]$, figure 13.7 droite).

Nous avons ensuite regroupé les segments en fonction de leurs tendances pour chaque dimension : ceux qui se situent en-dessus et ceux qui se situent en-dessous de la fréquence médiane des pauses en fonction de leur type et du score accentuel moyen.

Commençons par les pauses. On constate tout d'abord que les segments qui contiennent globalement moins de pauses (les 8 segments dont le nombre de pauses

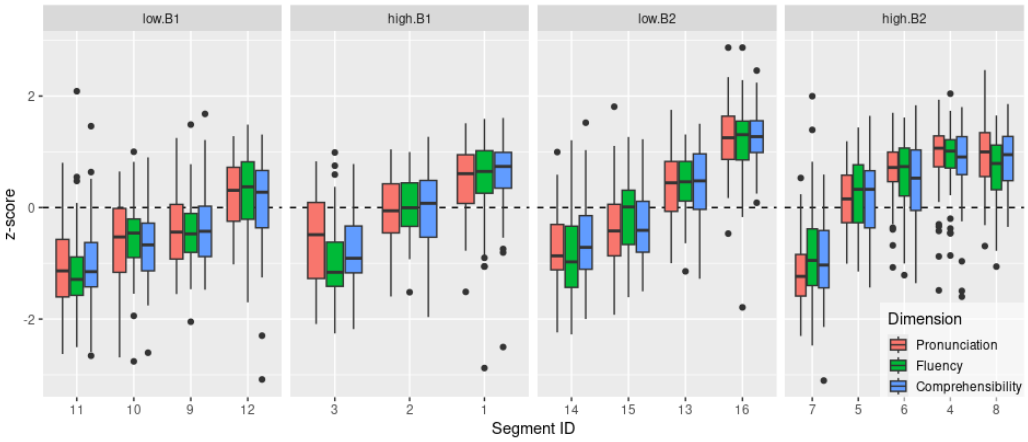


FIG. 13.6 : Évaluation globale normalisée des 16 segments en termes de qualité de prononciation (rouge) de fluidité (vert) et de compréhensibilité (bleu), en fonction du niveau du locuteur et de la catégorie du segment (low : haute proportion de pauses WP et bas score accentuel moyen ; high : basse proportion 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)

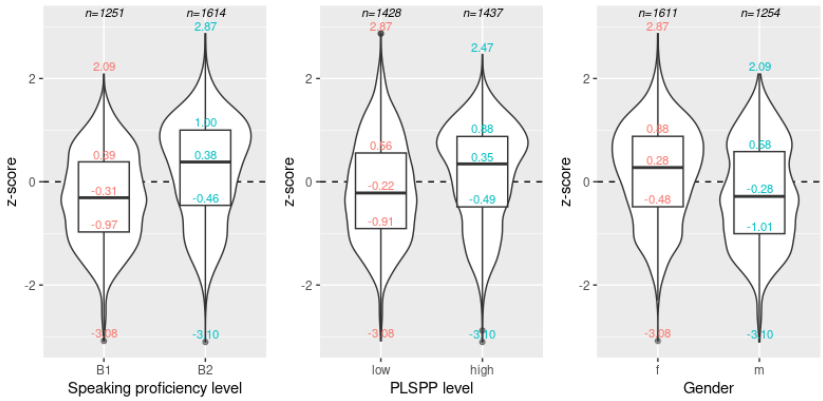


FIG. 13.7 : Évaluation globale normalisée en fonction du niveau, de la catégorie et du genre du locuteur (3 dimensions confondues, un point de donnée correspond à l'évaluation d'un segment par un évaluateur sur une dimension)

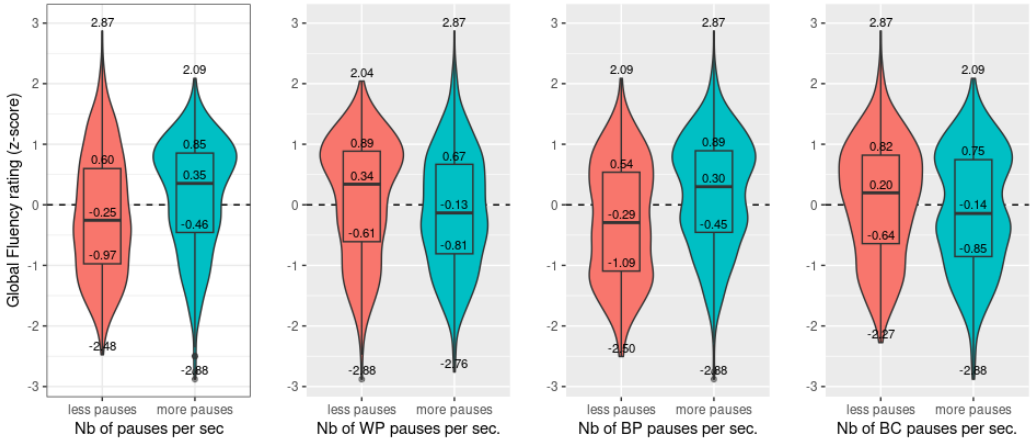


FIG. 13.8 : Enter Caption

par seconde est inférieur à la fréquence médiane) obtiennent un score de fluidité inférieur aux segments qui contiennent des pauses plus fréquentes ($p < 0{,}001$, médianes à -0,25 pour les premiers et +0,35 pour les seconds, $\Delta = -0{,}207$ (small) $IC = [-0{,}278; -0.134]$, cf. figure **??**). Qu'en est-il du score en fonction de la fréquence des pauses par catégorie syntaxique ? Comme attendu, les segments contenant moins de pauses intra-syntagme (WP) ont tendance à obtenir un score plus élevé ($p < 0{,}001$, médianes à +0,34 contre -0,13, $\Delta = 0{,}156$ (small) $IC = [0{,}082; 0{,}227]$), mais étrangement, c'est également le cas pour les pauses inter-proposition (BC) : les segments qui en ont moins on tendance à être mieux notés ($p < 0{,}05$, médianes à +0,2 contre -0,14, $\Delta = 0{,}088$ (negligible) $IC = [0{,}014; 0{,}16]$). Notons toutefois que la distribution des scores pour les segments qui contiennent plus de pauses inter-proposition présente deux gaussiennes distinctes, une positive autour de +0,8 et une négative autour de -0,3.

13.5 Analyse des commentaires libres

820 commentaires sur 955 entrées ; 11994 mots

partie IV

Discussion

Chapitre 14

Limitations et évolution

14.1 Limitations corpus

Bien que la session d'interaction orale du CLES B2 présente l'avantage d'évaluer la production orale des apprenants en contexte conversationnel, il est difficile de savoir à quel point les candidats adaptent leur discours en fonction de leur binôme. Si celui-ci montrent des difficultés pour comprendre et interagir, il est normal que le premier formule des énoncés plus simples, voire adopte une prononciation plus « française » pour se faire comprendre. Par ailleurs, en ce qui concerne l'analyse des pauses, il est difficile de déterminer quelles pauses sont utilisées dans un objectif de gestion de tours de parole – soit pour laisser la parole, soit au contraire pour la garder. Il est probable que des patterns différents soient observés en contexte monologal.

Une autre limitation évidente est bien-sûr le fait qu'il s'agisse d'un contexte d'examen, avec de surcroît la présence d'un examinateur dans la salle. Le stress engendré par cette situation peut affecter la production du candidat, sa fluidité et sa prononciation. Cette situation d'évaluation peut également impacter la conversation dans le sens où l'interlocuteur n'est plus nécessairement le second candidat, mais l'examinateur.

Enfin, gardons à l'esprit qu'il s'agit d'un jeu de rôle et que les candidats se retrouvent parfois à défendre un point de vue qui diffère du leur, et que certains points de vue peuvent s'avérer plus difficile à défendre que d'autres.

14.2 Limitations techniques

Isolation des segments de parole (pas de prise en compte du contexte de la conversation). Or les patterns de pauses sont probablement influencés par la position du segment dans le tour de parole du locuteur (au début ? à la fin ? au milieu ?). Il aurait peut-être fallu découper par tour de parole entier ? Pourrait-on adapter la sensibilité du découpage en fonction de caractéristiques de la conversation (bcp de petits tours de parole avec chevauchement vs. longs tours de parole sans chevauchement). Il faudrait pouvoir isoler la parole de chaque locuteur afin de pouvoir en conserver qu'une, même dans les cas de chevauchements.

Dans le cas où l'alignement des mots et des noyaux syllabiques ne correspond pas tout à fait, et notamment que la première ou la dernière syllabe n'est pas comprise entre les frontières de début et fin de mot, il peut être envisagé de considérer une marge de tolérance juste avant et/ou juste après le mot pour inclure la ou les éventuelles syllabes adjacentes, si elles ne font pas elles-mêmes partie du mot suivant ou précédent. Cette solution a été pensée face au constat selon lequel l'alignement de Wav2Vec2.0 a tendance à grignoter la première et la dernière syllabes des mots, résultant en une perte importante de mots cible à cause d'un problème d'alignement. Pour déterminer la durée optimale de cette marge, nous proposons de comparer la proportion de mots contenant le bon nombre de syllabes mentionnée dans le paragraphe précédent avec la proportion de mots avec marge contenant le bon nombre de syllabes. Cette marge est fixée de manière arbitraire à la moitié de l'intervalle adjacent si celui-ci est vide, et ce dans la limite de 100 ms. Elle pourra être modifiée à la baisse ou à la haute en fonction des résultats obtenus.

14.3 Concernant l'allongement final

Constat d'un important allongement final, en particulier en parole spontanée. Description étendue du phénomène par Astesano2001 p58 Lindblom1978 : allongement final = phénomène naturel mais spécifique à la langue (le degré d'allongement peut varier). En français, l'allongement final est renforcé par la présence de la syllabe accentuée. Ainsi, d'après NordAl1990 cités par Astesano2001, le contraste entre syllabe accentuée et non-accentuée est plus fort en français (accent final) qu'en anglais (accent à tendance initiale)

Chapitre 15

Implications pour le positionnement et le diagnostic

Conclusion

Bibliographie

Astesano, C. (2001). *Rythme et accentuation en français : invariance et variabilité stylistique*. L'Harmattan. https://books.google.fr/books?id=DaJBTWuv4KgC

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

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

Bredin, H. (2023). pyannote.audio 2.1 speaker diarization pipeline : principle, benchmark, and recipe. *Proc. INTERSPEECH 2023*.

Breitkreutz, J. A., Derwing, T. M., & Rossiter, M. J. (2001). Pronunciation Teaching Practices in Canada. *TESL Canada Journal*, *19*, 51-61.

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. *Proc. Speech Prosody 2002*, 199-202.

Candea, M. (2000, décembre). *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. [Theses]. Université de la 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.

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.-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. *Psychol. Bull.*, *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., The University 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*.

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/https://doi.org/10.1016/j.cognition.2006.10.010

Coulange, S., Fries, M.-H., Masperi, M., & Rossato, S. (2024, mai). A Corpus of Spontaneous L2 English Speech for Real-situation Speaking Assessment. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti & N. Xue (Éd.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (p. 293-297). ELRA ; ICCL. https://aclanthology.org/2024.lrec-main.27

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]. *Proceedings 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). 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

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

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297-334. https://doi.org/10.1007/BF02310555

Crystal, D. (2008, juin). *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. *J. Acoust. Soc. Am.*, *111*(6), 2862-2873.

Cutler, A. (2015, avril). 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/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/https://doi.org/10.1002/9781119184096.ch9

de Jong, N. (2016). *International Review of Applied Linguistics in Language Teaching*, *54*(2), 113-132. https://doi.org/doi: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.), *Proceedings of the 6th Workshop on Disfluency in Spontaneous Speech, DiSS* (p. 17-20).

de Jong, N., Pacilly, J., & Heeren, W. (2021). PRAAT scripts to measure speed fluency and breakdown fluency in speech automatically. *Assessment in Education : Principles, Policy & Practice*, *28*(4), 456-476. https://doi.org/10.1080/0969594X.2021.1951162

Delattre, P. (1963). Comparing the prosodic features of English, German, Spanish and french. *IRAL Int. Rev. Appl. Linguist. Lang. Teach.*, *1*(1).

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

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

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. *Proc. of the 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/https://doi.org/10.1006/jmla.1996.2500

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/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/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/doi: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.

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 University Press.

Fox Tree, J. (2001). Listeners' uses ofum anduh in speech comprehension. *Memory & Cognition*, *29*(2), 320-326.

Frost, D. (2023, août). 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., & Picavet, F. (2014). Putting prosody first — some practical solutions to a perennial problem : The Innovalangues project. *Res. Lang.*, *12*(3), 233-243.

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*.

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.

Grosjean, F., & Deschamps, A. (1972). *Phonetica*, *26*(3), 129-156. https://doi.org/doi: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/doi: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/https://doi.org/10.3917/lang.211.0013

Heldner, M., & Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, *38*(4), 555-568. https://doi.org/https://doi.org/10.1016/j.wocn.2010.08.002

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy : Industrial-strength Natural Language Processing in Python. https://doi.org/https://doi.org/10.5281/zenodo.1212303

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

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/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. *Proc. 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. *Lang. Assess. Q.*, *15*(2), 150-168.

Kernou, H. (2022). Les disfluences dans le discours radiophonique : signification(s) et fonction (s) communicative (s). *Multilinguales*, (17).

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. *Comput. Speech Lang.*, *45*, 326-347.

Kitaev, N., Cao, S., & Klein, D. (2019). Multilingual Constituency Parsing with Self-Attention and Pre-Training. *Proceedings 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. https://arxiv.org/abs/2012.14788

Krivokapić, J. (2007). Prosodic planning : Effects of phrasal length and complexity on pause duration. *J. Phon.*, *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/https://doi.org/10.1016/j.lingua.2005.06.010

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. *Can. J. Behav. Sci.*, *1*(1), 25-37.

Lennon, P. (1990). Investigating Fluency in EFL : A Quantitative Approach. *Language Learning*, *40*(3), 387-417. https://doi.org/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. *J. Verbal Learning Verbal Behav.*, *6*(4), 560-564.

Lezcano, M. F., Dias, F., Arias, A., & Fuentes, R. (2020). Accuracy and Reliability of AG501 Articulograph for Mandibular Movement Analysis : A Quantitative Descriptive Study. *Sensors*, *20*(21). https://doi.org/10.3390/s20216324

Li, C., Liu, J., & Xia, S. (2007). English sentence stress detection system based on HMM framework. *Appl. Math. Comput.*, *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/https://doi.org/10.1016/j.specom.2017.11.003

Lickley, R. (1995). Missing disfluencies. *Proc. of the ICPhS*, *4*, 192-195.

Lickley, R. (2015, juin). 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

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., University 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., The University 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

Martin, J., & Strange, W. (1968). The perception of hesitation in spontaneous speech. *Percept. Psychophys.*, *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.

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, janvier). *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.

Proc. Interspeech 2017, 498-502. https://doi.org/10.21437/Interspeech.2017-1386

Mertens, P. (2008). Syntaxe, prosodie et structure informationnelle : une approche prédictive pour l'analyse de l'intonation dans le discours. *Travaux de linguistique*, *n° 56*(1), 97-124.

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

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

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/https://doi.org/10.1016/0167-6393(90)90023-3

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

Pennington, M. C. (1999). Computer-Aided Pronunciation Pedagogy : Promise, Limitations, Directions. *Computer Assisted Language Learning*, *12*(5), 427-440. https://doi.org/10.1076/call.12.5.427.5693

Piccardo, E. (2016). Common European Framework of Reference for Languages : Learning, Teaching, Assessment. Phonological Scale Revision Process Report [Accessed : 2024-9-12].

Plaquet, A., & Bredin, H. (2023). Powerset multi-class cross entropy loss for neural speaker diarization. *Proc. INTERSPEECH 2023*.

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/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*.

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, janvier). *Lectures on Conversation* (G. Jefferson, Éd.). Blackwell.

Saito, K., Macmillan, K., Kachlicka, M., Kunihara, T., & Minematsu, N. (2022). Automated assessment of second language comprehensibility : Review, training, validation, and generalization studies. *Stud. Second Lang. Acquis.*, 1-30.

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, janvier). *Cognitive bases of second language fluency*. Routledge.

Shea, C., & Leonard, K. (2019). Evaluating measures of pausing for second language fluency research. *Can. Mod. Lang. Rev.*, *75*(3), 216-235.

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 University*, *30*(2), 11-18. https://cir.nii.ac.jp/crid/1520290882531293440

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations : Uses in assessing rater reliability. *Psychol. Bull.*, *86*(2), 420-428.

Siegman, A., & Feldstein, S. (1979, novembre). *Of speech and time*. John Wiley & Sons.

Siegman, A., & Pope, B. (1966). Ambiguity and verbal fluency in the TAT. *J. Consult. Psychol.*, *30*(3), 239-245.

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

Sugahara, M. (2011). Identification of English primary stress and bias toward strong word-initial syllables : native vs. Japanese listeners. *Proceedings of 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 university 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., & 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/https://doi.org/10.1111/modl.12706

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. *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, *1*, 937-940.

Thomson, R. I. (2015). Fluency. In *The Handbook of English Pronunciation* (p. 209-226). John Wiley ; Sons, Ltd. https://doi.org/https://doi.org/10.1002/9781118346952.ch12

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.

Trouvain, J., Bonneau, A., Colotte, V., Fauth, C., Fohr, D., Jouvet, D., Jügler, J., Laprie, Y., Mella, O., Möbius, B., & Zimmerer, F. (2016, mai). 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.), *Proceedings 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 University] [PhD thesis, Saarland University].

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 : Proceedings 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

White, S. (1989). Backchannels across Cultures : A Study of Americans and Japanese. *Language in Society*, *18*(1), 59-76. Récupérée septembre 27, 2024, à partir de 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.

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.

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/

Annexe A Penn Treebank II Constituent Tags

Source : https://surdeanu.cs.arizona.edu//mihai/teaching/ista555-fall13/readings/PennTreebankConstituents.html (consultée le 3 novembre 2024)

1.1 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.

1.2 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.

1.3 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

Annexe B Captures d'écran de Dynamic Rater

L'application web Dynamic Rater[1] a été développée pour les besoins de cette étude. Elle se compose de 4 vues principales : une page d'accueil avec la présentation

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

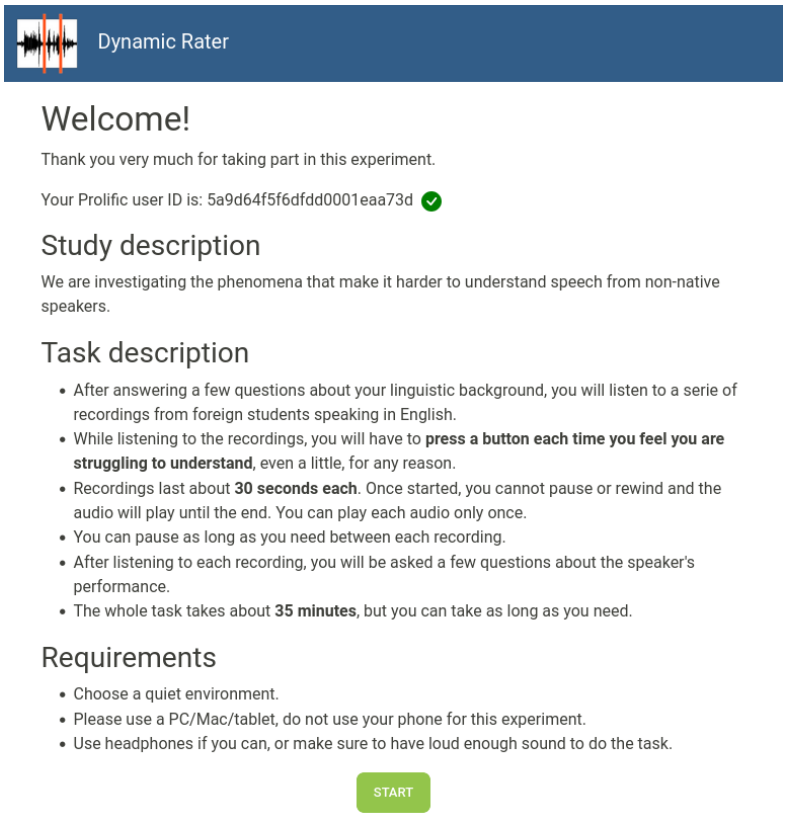


FIG. 15.1 : Page d'accueil de Dynamic Rater

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 15.1 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 partici-

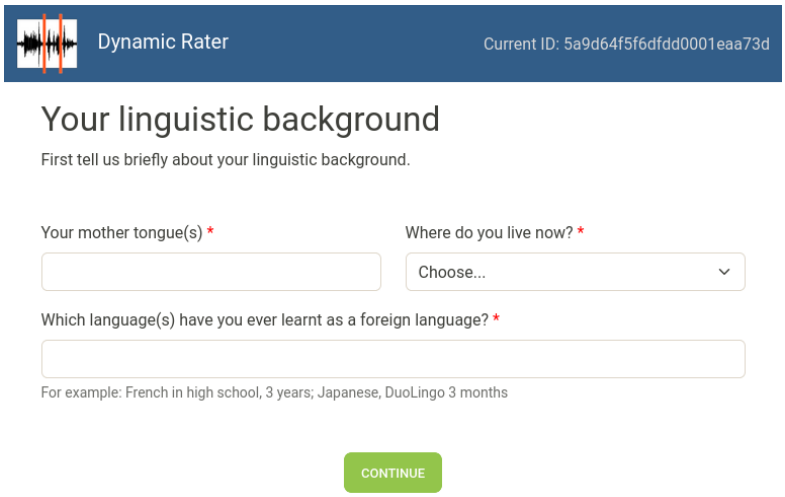


Fig. 15.2 : Questionnaire linguistique

pants 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 15.2.

La phase d'entraînement, figure 15.3, 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 15.4).

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 15.5). En cliquant sur le bouton *validate the survey*, il est redirigé vers Prolific, qui est alors informé de la fin de passation.

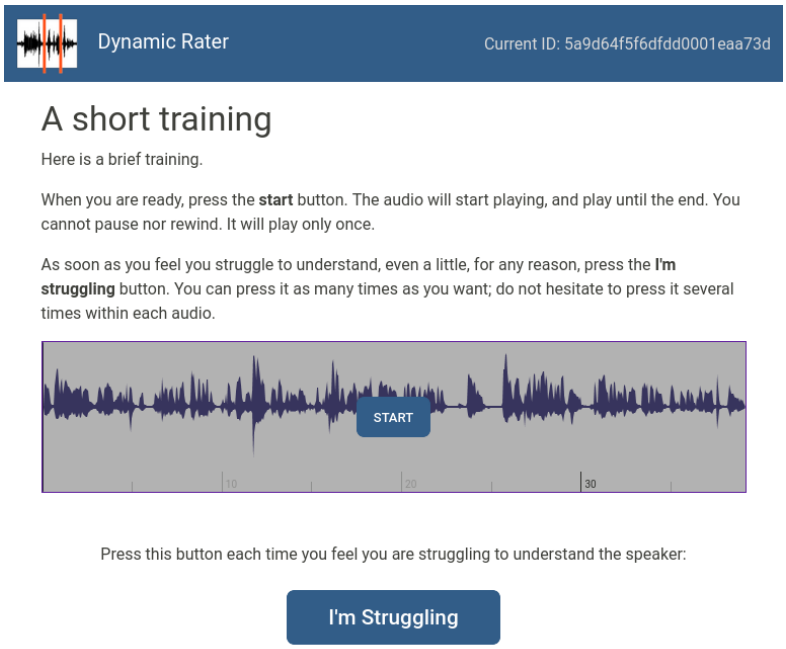


FIG. 15.3 : Phase d'entraînement

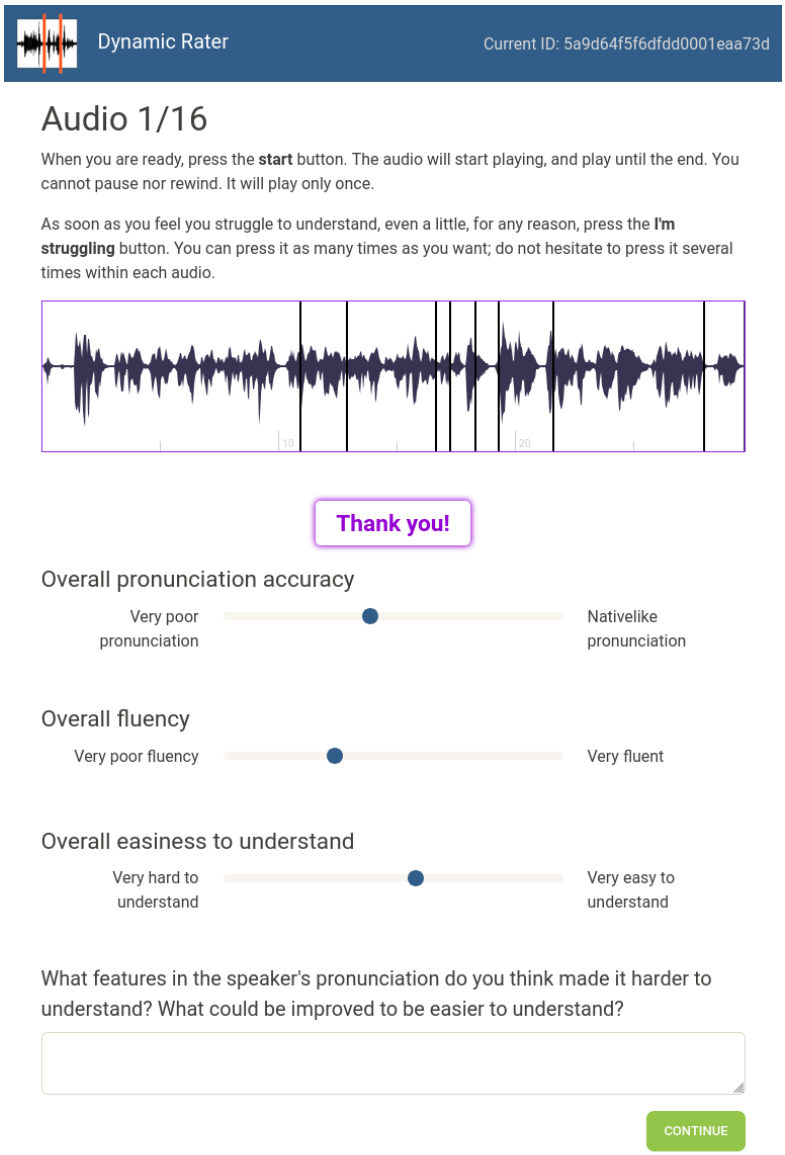


FIG. 15.4 : Évaluation globale à la suite de l'évaluation dynamique d'un enregistrement

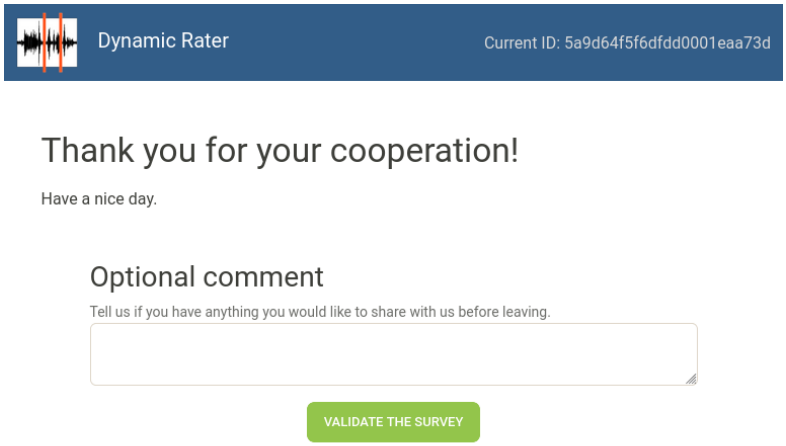


FIG. 15.5 : Écran de fin d'expérience