Chapitre 10

Discussion

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

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

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

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

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

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

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

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

10.1 Principaux résultats obtenus

1.1 Annotation des pauses

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

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

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

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

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

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

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

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

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

1.2 Annotation de l'accent lexical

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

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

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

Les annotations du corpus CLES-FR montrent des résultats variés entre locuteurs, mais des scores globalement plus élevés pour les B2 par rapport aux B1 ($p < 0{,}001$). Toutefois, les scores de position restent faibles, avec une médiane à 30,8 % pour les B1 et 36,8 % pour les B2, et les contrastes prosodiques (f_0, intensité, durée) demeurent limités. Les locuteurs ayant un score de position élevé marquent principalement la syllabe accentuée en f_0 et en intensité, tandis que ceux ayant un score faible tendent à ne pas produire de contraste d'intensité, et produire un allongement marqué de la syllabe finale accompagné d'une montée de f_0.

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

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

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

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

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

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

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

Ces résultats confirment que les pauses de bas niveau syntaxique et les patterns accentuels incorrects ont un effet négatif sur la compréhensibilité du locuteur, en accord avec les travaux de Isaacs et Trofimovich (2012), Saito et al. (2015) et Suzuki et Kormos (2020). Contrairement aux conclusions de Nagle et al. (2019), nos observations montrent qu'une évaluation dynamique de la compréhensibilité est possible et pertinente, à condition de simplifier le protocole d'évaluation et de normaliser les patterns de clics pour tenir compte des variations individuelles des auditeurs.

10.2 Apports de notre travail

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

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

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

plusieurs métriques permettant de comparer les productions de différents groupes de locuteurs.

Notre chaîne de traitement, bien qu'adaptable, n'a pas vocation à être utilisée en l'état dans des applications finales. Son aspect modulaire lui permet de s'adapter à des données et des types de parole variés. Ainsi, PLSPP a déjà évolué en différentes versions pour s'adapter à des corpus incluant des données de parole lue, récitée, ou spontanée, produites par des locuteurs enfants et adultes de diverses langues maternelles (japonais, coréen, slovaque, portugais ou anglais) (**NishiokaAl2024** ; **NishiokaAl2025a** ; **NishiokaAl2025b** ; **Erickson2025** ; Kimura et al., 2024 ; Nakanishi & Coulange, 2024 ; Raso et al., 2024 ; Sugahara et al., 2023, 2024). Le code source de PLSPP est accessible ici[1].

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

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

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

Les trois corpus de conversations spontanées conçus dans cette thèse représentent une contribution précieuse, dans la mesure où ils mettent à disposition de la communauté une grande quantité d'enregistrements d'apprenants, accompagnés d'une évaluation précise du niveau de production orale par des évaluateurs experts

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

effectuée sur la base de ces productions. Les corpus sont accessibles aux adresses suivantes : CLES-FR[3], CLES-JP[4] et CLES-EN[5].

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

10.3 Limites & perspectives

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

3.1 Limitations de corpus

Les sessions d'interaction orale du CLES B2 offrent l'avantage d'évaluer la production orale en contexte conversationnel. Cependant, il est difficile d'estimer dans quelle mesure les candidats adaptent leur discours en fonction du niveau de leur binôme. Si ce dernier éprouve des difficultés à comprendre ou à interagir, il est naturel que le candidat simplifie ses énoncés, voire adopte une prononciation moins authentique afin d'être mieux compris. Comme l'ont souligné Nagle et al. (2022), une conver-

[3] https://hdl.handle.net/11403/cles-spontaneous-english
[4] à ajouter
[5] à ajouter
[6] https://gricad-gitlab.univ-grenoble-alpes.fr/lidilem/dynamic-rater

sation implique nécessairement un alignement entre interlocuteurs. D'après eux, ce phénomène de convergence linguistique se traduit par l'appropriation mutuelle d'expressions, de structures syntaxiques, mais aussi, dans une certaine mesure, du rythme de parole du partenaire.

Cela dit, ce biais d'adaptation est probablement atténué par le contexte d'évaluation. Bien que la conversation ait lieu entre les candidats, le véritable destinataire reste l'évaluateur, présent dans la salle. De plus, le stress associé à cette situation peut également influencer la production des candidats.

Le choix d'un contexte conversationnel pour analyser les patterns de pauses s'est révélé être une autre limite. Il est en effet difficile de distinguer les pauses destinées à gérer les tours de parole de celles ayant pour fonction de structurer l'énoncé ou résultant d'une hésitation. Dans un contexte monologal, des patterns de pauses probablement différents auraient pu être observés.

Enfin, le fait que les sessions prennent la forme de jeux de rôle, où les candidats doivent parfois défendre des points de vue différents des leurs, constitue une contrainte supplémentaire. Celle-ci peut influencer leur fluidité ou leur rythme de parole. Toutefois, cette difficulté étant partagée par tous les participants des trois corpus, elle ne devrait pas avoir un impact significatif sur les comparaisons entre groupes de locuteurs.

3.2 Limitations des analyses de parole

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

Pré-traitements

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

De plus, la compilation des segments pourrait être optimisée pour inclure systématiquement tous les segments formant un tour de parole complet. La position d'un segment dans le tour de parole influence probablement les patterns de pauses observés. Nous avons également introduit un seuil paramétrable pour tolérer, dans une certaine

mesure, les réactions de l'interlocuteur dans un segment de parole. Une adaptation dynamique de ce seuil, en fonction des caractéristiques spécifiques de la conversation (par exemple, une faible fréquence de chevauchements de parole), serait bénéfique.

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

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

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

Annotations des pauses

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

Lors de l'analyse des distributions des pauses, l'attention a d'abord été portée sur les types de constituants (propositions et syntagmes). Toutefois, cette approche s'est révélée limitée en raison de la rareté des frontières intra-syntagmes en anglais et du grand nombre de frontières inter-syntagmes laissées de côté. Le calcul du score de distribution syntaxique des pauses (DSP_n), reposant sur l'importance relative des frontières syntaxiques (estimée à partir du nombre de constituants se fermant ou s'ouvrant), a permis de pallier ces limites en intégrant l'imbrication des propositions et des syntagmes les uns dans les autres. Cependant, cette méthode nécessite de définir des seuils d'importance pour le calcul du score. Il serait intéressant de réfléchir à une approche plus continue et moins catégorique.

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

Annotations de l'accent lexical

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

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

Cette limitation concernant la mesure de durée est également due à la précision de l'alignement mot-signal effectué par Wav2Vec2.0 Baevski et al., 2020, dans la première version de PLSPP. En effet, si cet aligneur s'est révélé plus robuste à la parole

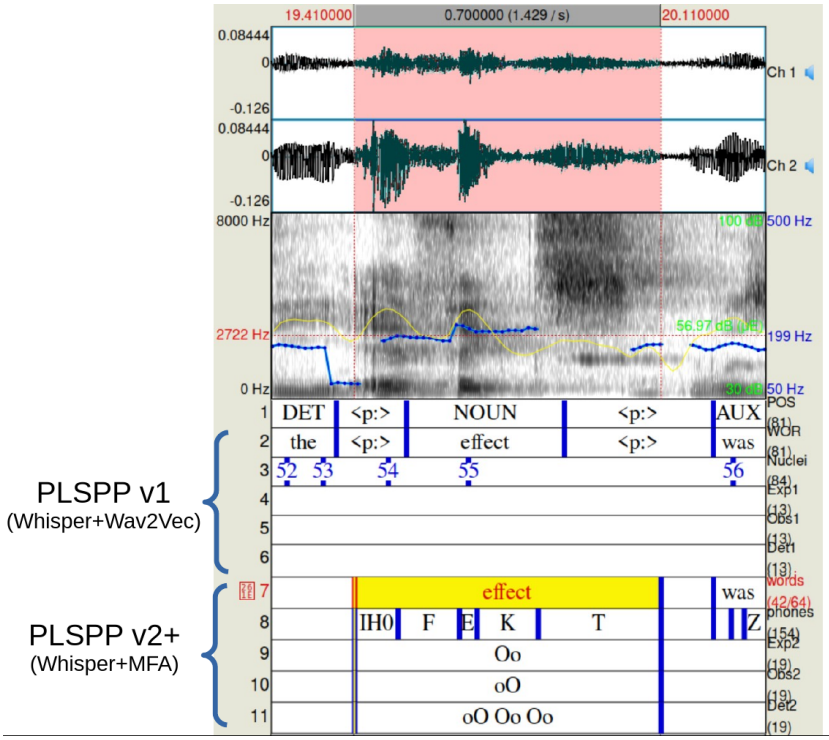

FIG. 10.1 : Illustration d'un cas d'alignement de mot trop court avec Wav2Vec2.0, bloquant l'analyse de proéminence. En comparaison, l'alignement avec MFA est plus adéquat ici. La tier 9 correspond au gabarit accentuel attendu, la 10 à celui qui est observé (accent en finale), la 11 détaille la position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

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

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

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

Mesures de f_0 Les résultats mitigés obtenus sur la parole spontanée des locuteurs anglophones natifs ont révélé que la mesure de f_0 est influencée par la tendance au dévoisement de certaines voyelles, particulièrement fréquente chez les jeunes locuteurs de l'anglais. Ce dévoisement résulte en une absence de détection de fréquence fondamentale au niveau du noyau syllabique. Cette absence est alors comblée par une interpolation linéaire de la f_0 à partir des mesures les plus proches, mais peut parfois aboutir à des résultats incohérents.

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

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

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

chez les locuteurs japonophones, qui ont tendance à insérer des voyelles dans certains groupes consonantiques (Kenworthy, 1987 ; Labrune, 2006).

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

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

3.3 Évaluation dynamique de la compréhensibilité

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

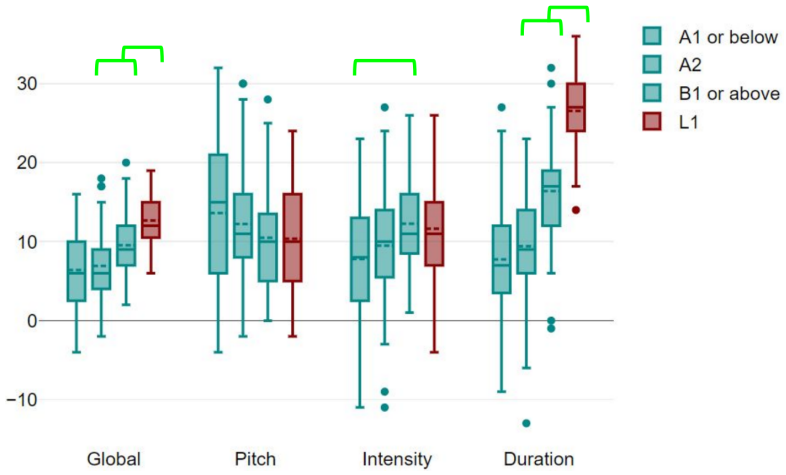

Fig. 10.2 : *Degré de contraste prosodique entre les mots grammaticaux et lexicaux à travers un échantillon aléatoire de 55 textes par groupe de niveau (A1, A2, B1 et plus, natifs, dans l'ordre d'apparition). Les crochets verts indiquent les différences significatives (ANOVA, $p < 0{,}001$)*

avec un contraste élevé ($C'' \geq 0{,}2$) concernaient seulement 17 % des mots annotés (23 sur 139).

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

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