Caractérisation et mesure de la compréhensibilité de la parole de locuteurs non natifs dans le cadre de l'apprentissage des langues

Verdiana De Fino

RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

Université de Toulouse

Doctorat de l'Université de Toulouse

préparé à l'Université Toulouse III - Paul Sabatier

Caractérisation et mesure de la compréhensibilité de la parole de locuteurs non natifs dans le cadre de l'apprentissage des langues

---

Thèse présentée et soutenue, le 11 mars 2024 par

Verdiana DE FINO

École doctorale
EDMITT - Ecole Doctorale Mathématiques, Informatique et Télécommunications de Toulouse

Spécialité
Informatique et Télécommunications

Unité de recherche
IRIT : Institut de Recherche en Informatique de Toulouse

Thèse dirigée par
Julien PINQUIER et Isabelle FERRANE

Composition du jury
M. Sylvain DETEY, Président, Waseda University
Mme Martine ADDA-DECKER, Rapporteure, CNRS - Paris
Mme Marie TAHON, Rapporteure, Le Mans Université
M. Julien PINQUIER, Directeur de thèse, Université Toulouse III - Paul Sabatier

Membres invités
Mme Isabelle FERRANÉ, Université Toulouse III - Paul Sabatier
M. Lionel FONTAN, Archean Technologies

Remerciements

Je tiens tout d'abord à remercier chaleureusement mon directeur et mes encadrants de thèse qui m'ont suivie durant ces trois années : Julien Pinquier, Isabelle Ferrané et Lionel Fontan. La réussite d'une thèse dépend certes du doctorant, mais je peux assurer qu'elle est fortement corrélée à 0,97 à votre encadrement. Je vous remercie de m'avoir soutenue et cru en moi lorsque moi-même je n'y arrivais pas, et de toute la confiance et la patience dont vous avez fait preuve à mon égard.

Mes remerciements vont également à l'ensemble des membres du jury : Martine Adda-Decker et Marie Tahon pour avoir lu et rapporté mon manuscrit, et Sylvain Detey pour avoir présidé ce jury. Je vous remercie pour vos retours constructifs et pour avoir fait le déplacement jusqu'à Toulouse pour assister à la présentation de cette thèse.

Je remercie toutes les personnes qui ont pu rendre possible la réalisation des corpus. Je pense surtout à Corentin et Joy pour avoir pris en charge les sessions d'enregistrement au Japon, mais également aux apprenants japonais et allemands, ainsi qu'aux différents évaluateurs qui se sont portés volontaires pour mener à bien cette collecte. Merci à Reinhard Gerndt et Véronique Bizien pour leur accueil en Allemagne, et à Luis, Fynn, Bjarne, Katy et Sam pour la ronde des marchés de Noël.

Je remercie tout naturellement les membres de l'équipe SAMoVA. Je n'aurais pas pu rêver d'une meilleure équipe avec laquelle partager ces années. Merci à tous les membres permanents, Jérôme, Julie, Hervé, Thomas et Christine, vous avez tous, sans exception, été disponibles au moins une fois pour m'aider. Merci aux docteurs, doctorants et ingénieurs qui ont pu croiser mon chemin et pour certains devenir de véritables amis. Je pense à Timothy, Mathieu, Sebastião, Lucile, Vincent, Léo, Alexis, Romain, Philippe, Adrien, Ludovic, Clément, Benjamin, Amélie, Gautier, Antoine, et à ceux que j'ai peut-être oublié… Tous nos échanges, les parties de tarot à midi, les après-midi jeux de société et les soirées (à Toulouse, Capbreton, Noirmoutier ou même Séoul) resteront à jamais d'excellents souvenirs. Et je n'oublie évidemment pas tous les stagiaires qui sont passés dans cette équipe.

Merci (et félicitations !) à Étienne. Nous avons traversé les mêmes épreuves et tracas en même temps, que ce soit pour la rédaction du manuscrit ou le stress présoutenance. Merci pour avoir toujours répondu à mes questions et pour nos nombreuses pauses sur la terrasse à 19h30 durant nos derniers mois de thèse.

Merci à mes deux colocataires et amis aveyronnais du bureau 227. J'ai passé d'excellentes années à vos côtés, rythmées par la bonne humeur, les potins, les rires, encore les potins, les WTT, les pauses, les brunchs et les soirées. Merci Lila, la plus sérieuse de ce trio, pour toutes les petites attentions, les mots sur le tableau, pour ta capacité à écouter, rassurer, encourager, et pour m'avoir toujours tirée vers le haut. Merci Robin, sûrement la personne qui me ressemble le plus, pour les encouragements mais surtout pour toute l'écoute et l'aide précieuse que tu m'as apportées à n'importe quel moment, et pour avoir été à la fois un enseignant, un chef pâtissier, un meilleur ami.

Je remercie bien évidemment ma famille et tout l'investissement dont elle a fait

preuve. Vous m'avez soutenue, épaulée et remonté le moral à de nombreuses reprises, mais vous avez surtout cru en moi et en cette réussite, du début à la fin.

Mes derniers remerciements vont enfin à la meilleure rencontre que j'ai faite à l'IRIT et qui partage ma vie depuis aujourd'hui quatre ans. Merci Jim d'avoir été tout simplement toi, calme, encourageant, compréhensif. Ces trois années n'auraient pas pu aussi bien se passer sans le rôle crucial que tu as tenu.

À Zio Lino.

Table des matières

Table des figures

Liste des tableaux

Glossaire

ALE : *Anglais Langue Étrangère*.

BDD : *Base De Données*, base de stockage de données structurées.

CAF-al : *Compréhensibilité d'Apprenants de Français - Allemands*, corpus constitué de productions orales d'apprenants allemands de français ainsi que des scores de compréhensibilité associés.

CAF-jp : *Compréhensibilité d'Apprenants de Français - Japonais*, corpus constitué de productions orales d'apprenants japonais de français ainsi que des scores de compréhensibilité associés.

CECRL : *Cadre Européen Commun de Référence pour les Langues*, norme internationale permettant de décrire la compétence linguistique.

DALF : *Diplôme Approfondi de Langue Française*.

DELF : *Diplôme d'Études en Langue Française*.

FLE : *Français Langue Étrangère*.

IELTS : *International English Language Testing System*, examen utilisé pour évaluer le niveau d'anglais, administré par l'université de Cambridge (Royaume-Uni).

IHM : *Interface Homme-Machine*, interface permettant à un utilisateur de communiquer avec une machine, un programme informatique ou un système.

L1 : *Langue maternelle*, langue maternelle d'un individu.

L2 : *Langue seconde*, langue en cours d'apprentissage d'un individu.

LASSO : *Least Absolute Shrinkage and Selection Operator*, technique de régularisation utilisée dans la régression linéaire.

MAE : *Mean Absolute Error*, Erreur Absolue Moyenne. Métrique de régression qui mesure la moyenne des valeurs absolues des erreurs de prédiction.

PATY : *plateforme de traitement de Parole Atypique*, plateforme mettant à disposition des systèmes de transcription automatique de la parole développés par l'équipe SAMoVA de l'IRIT.

PER : *Phone Error Rate*, taux d'erreur de phonème.

RF : *Random Forest*, algorithme de régression non-linéaire appartenant aux méthodes d'ensemble.

TOEFL IBT : *Test Of English as a Foreign Language, Internet-Based Test*, examen utilisé pour évaluer le niveau d'anglais dans un contexte universitaire.

TTR : *Type-Token Ratio*, paramètre de diversité lexicale.

WER : *Word Error Rate*, taux d'erreur de mots.

Introduction

La communication orale tient une place importante dans de nombreux aspects de la vie quotidienne. Elle permet à chaque individu d'exprimer ses idées, d'établir des relations et surtout de partager des informations. Pour garantir une compréhension mutuelle avec des interlocuteurs, il est primordial d'être capable de structurer ses pensées et de savoir ou pouvoir les énoncer clairement. La communication orale peut cependant être dégradée par plusieurs facteurs qui peuvent avoir un effet sur l'auditeur et sa compréhension du message perçu ou qui lui est adressé, et sur le locuteur dans sa difficulté à produire un message clair et compréhensible dans un contexte de communication donné. Du point de vue de l'auditeur, une difficulté de compréhension peut par exemple être liée à une perte d'information causée par un trouble du système auditif, tel que la presbyacousie, ou par un environnement très bruité. Du point de vue du locuteur, des difficultés de production de la parole, que ce soit pour des raisons médicales ou de maîtrise de la langue, peuvent diminuer de manière importante la compréhensibilité de ce qui est dit. C'est le cas des paroles dites atypiques :

— la parole pathologique (parole dégradée, par exemple à l'issue d'un acte chirurgical tel qu'une résection à la suite d'un cancer de la cavité buccale, comme l'ablation d'une partie de la langue (Barrett *et al.*, 2004), ou causée par un trouble neurologique comme par exemple la dysphasie ou la dysarthrie),

— la parole d'enfants apprenants lecteurs (Gelin, 2022) ;

— la parole de personnes apprenant une langue étrangère (Xue *et al.*, 2019).

C'est dans ce dernier contexte que se situent les travaux de thèse présentés dans ce mémoire.

Intérêt pour les apprenants L2

L'apprentissage d'une langue étrangère est primordial afin de permettre à des interlocuteurs de différentes langues maternelles de communiquer. Le processus d'apprentissage peut mener à deux objectifs, à savoir le fait de maîtriser la langue cible comme un natif, idéalement sans accent audible, et le fait de se faire comprendre en situation de communication, malgré la présence d'un accent dû à sa langue maternelle (L1) ou d'erreurs linguistiques. Bien que le premier objectif représente l'aboutissement ultime pour la plupart des enseignants et apprenants (Derwing, 2003), maîtriser à ce

point la langue seconde (L2) représente un réel défi que peu d'adultes parviennent à surmonter, et ce malgré un apprentissage initié au plus jeune âge (Flege, 1988). Afin de parvenir à une communication L2 réussie, il est de nos jours plus important pour un apprenant, dans le domaine de la didactique des langues, d'être compréhensible que de produire une parole non accentuée ou proche de celle d'un natif (Munro et Derwing, 1995a; Derwing et Munro, 1997; Jenkins, 2000; Derwing et Munro, 2009; Munro et Derwing, 2011).

Dans tout processus d'apprentissage se trouvent des phases d'évaluation, réalisées idéalement de manière objective (c'est-à-dire ne reposant pas sur le jugement subjectif d'un individu), des phases de retours et des phases de remédiation. Ces deux dernières étapes sont essentielles, car les retours pertinents faits aux apprenants permettent d'adapter l'apprentissage pour aider à combler les lacunes et améliorer les aspects de la production de parole en langue cible qui entrent en compte dans la compréhensibilité (Lyster et Saito, 2010; Saito et Akiyama, 2017). Cependant, lorsqu'un enseignant prend en charge un grand nombre d'apprenants, il est complexe de pouvoir fournir des évaluations, des retours et des remédiations personnalisés. L'apport de retours pertinents se voit être limité dans les classes de langue, et cette limitation est principalement liée au manque de temps pour écouter et évaluer les productions orales de chaque apprenant (Muñoz, 2014). Afin de pouvoir fournir un suivi adéquat et limiter l'aspect chronophage de l'évaluation s'opérant soit *a posteriori* en réécoutant des enregistrements d'apprenants, soit directement en classe, des solutions doivent être mises en place. Celles-ci peuvent par exemple s'apparenter à de nouvelles techniques pédagogiques (Pennington, 2021) ou à la mise à disposition d'outils en ligne auxquels les apprenants peuvent accéder (Loewen *et al.*, 2019). La compréhensibilité représentant un réel objectif pour tous les apprenants de langue étrangère, il est alors primordial d'être en mesure de l'évaluer avec les outils automatiques adéquats.

Contexte et problématiques

Ce travail de doctorat s'inscrit dans le cadre du projet ANR-18-LCV3-001 (FR) ALAIA [1] (Apprentissage des Langues Assisté par Intelligence Artificielle). ALAIA est un laboratoire commun initié en 2019 entre l'IRIT (Institut de Recherche en Informatique de Toulouse) et la société Archean Technologies. Son principal objectif est de proposer des outils d'aide à l'apprentissage des langues reposant sur des méthodes d'intelligence artificielle. Ces outils sont destinés à la fois aux enseignants, lors de la gestion d'un groupe d'apprenants, et aux apprenants eux-mêmes, afin de permettre un travail en autonomie. La thèse CIFRE qui a servi de support à ce travail de recherche a démarré dans ce contexte. La question de recherche sous-jacente a porté sur la compréhensibilité de la parole non-native, et plus particulièrement sur les moyens et les méthodes permettant de la mesurer automatiquement et objectivement, puis d'en

1. https://www.irit.fr/SAMOVA/site/projects/current/labcom-alaia/

évaluer la pertinence. Plusieurs questions se sont alors posées au cours de ce travail de thèse :

Question 1 : Comment caractériser la compréhensibilité de la parole de personnes apprenant une langue étrangère ?

Question 2 : Quel protocole faut-il définir et mettre en place pour collecter des données pertinentes ?

Question 3 : Quel protocole faut-il mettre en place pour annoter ces données et obtenir une vérité terrain en termes de compréhensibilité ?

Question 4 : Quelle méthode proposer pour mesurer de manière automatique la compréhensibilité des apprenants ?

Organisation du manuscrit

Au cours des trois années de thèse, j'ai tenté de répondre à ces différentes questions. Les réflexions, propositions et résultats sont présentés dans ce mémoire organisé en quatre chapitres.

Dans le premier chapitre, nous présentons un premier travail sur la compréhensibilité telle qu'elle est définie dans la littérature et dans le contexte de l'apprentissage des langues. Suite à cette étude, et à la diversité des travaux autour de la compréhensibilité, il s'est avéré indispensable de bien définir cette notion. Nous proposons donc une définition de la compréhensibilité qui sera le fil conducteur de nos travaux. Pour répondre à la première question de recherche posée précédemment, nous avons étudié et recensé les différents facteurs considérés classiquement pour l'évaluation de la compréhensibilité dans le contexte de l'apprentissage des langues. Nous étudions ainsi la compréhensibilité de la parole d'un point de vue linguistique, en examinant les différents niveaux qui la modulent et qui peuvent jouer un rôle dans sa mesure. Ensuite, nous nous intéressons aux facteurs non linguistiques qui peuvent l'influencer, selon trois points de vue : celui de l'auditeur, du locuteur et de la tâche de production orale au travers de laquelle la compréhensibilité est évaluée. Enfin, nous présentons une première étude visant à mesurer la compréhensibilité de la parole d'apprenants japonais d'anglais de manière automatique, en utilisant uniquement des paramètres phonético-phonologiques et mélodiques (accentuation et intonation).

Le deuxième chapitre est consacré aux différents paramètres linguistiques multi-niveaux que nous pouvons extraire automatiquement des productions orales des apprenants afin de les intégrer dans la prédiction automatique d'un score de compréhensibilité de la parole. À défaut d'avoir à disposition, à ce stade de notre étude, un corpus contenant des scores de compréhensibilité correspondant à une vérité terrain, nous avons abordé l'évaluation au travers d'un corpus contenant des enregistrements d'apprenants japonais de français, et associant ces apprenants à leur niveau CECRL (Cadre Européen Commun de Référence pour les Langues ; Conseil De l'Europe, 2001). Les paramètres extraits proviennent de différents niveaux linguistiques, tels que la phonologie/phonétique, le lexique, la syntaxe et le discours. La pertinence de ces paramètres

est d'abord démontrée par le biais d'une tâche de classification des apprenants selon leurs niveaux CECRL. Les évaluations CECRL rendant généralement peu compte du niveau en expression orale, nous décrivons l'étape d'enrichissement du corpus qui nous a permis de collecter des scores de compétences en production orale. Nous utilisons ensuite ces paramètres linguistiques multi-niveaux pour vérifier leur adéquation lors d'une tâche de prédiction de ces scores en expression orale. La prédiction du niveau CECRL des apprenants de ce corpus a ainsi permis de valider les différents paramètres qui peuvent ensuite contribuer à la mesure de la compréhensibilité.

Étant donné l'absence de corpus associant enregistrements de productions d'apprenants de français et scores de compréhensibilité, une contribution importante de cette thèse a été de créer un corpus dédié à la mesure de la compréhensibilité de la parole non native. Nous présentons dans un premier temps le protocole que nous avons défini afin de développer une interface et collecter des enregistrements d'apprenants. Les collaborations dans le cadre du Labcom ALAIA avec le Japon nous ont permis de travailler avec des apprenants japonais du français. Dans un second temps, nous présentons le protocole défini afin d'obtenir des annotations de la compréhensibilité de la parole et ainsi collecter des scores subjectifs, en se basant sur les conclusions que nous avons pu tirer de l'état de l'art. Nous présentons ensuite une analyse détaillée de ce premier corpus, nommé CAF-jp (Compréhensibilité d'Apprenants de Français - Japonais). Comme nous avons plusieurs annotations par enregistrement, cette analyse permettra de vérifier, entre autres, l'accord inter-annotateurs. Cette étape nous permet de vérifier la cohérence des annotations et la fiabilité des données utilisées par la suite. Nous montrons ensuite la généricité de notre protocole de création de corpus, en adaptant la tâche de production orale à des apprenants allemands de français, créant ainsi notre deuxième corpus, nommé CAF-al (Compréhensibilité d'Apprenants de Français - Allemands).

Enfin, le quatrième et dernier chapitre est consacré à la prédiction de la compréhensibilité de la parole, au cœur de la problématique de la thèse. À partir des mesures linguistiques implémentées dans le deuxième chapitre, nous tentons de prédire la compréhensibilité des apprenants du corpus CAF-jp. En premier lieu, nous ajoutons de nouveaux paramètres lexicaux et sémantiques pouvant se mesurer uniquement à partir de certains types de productions orales. Nous effectuons et comparons les prédictions de la compréhensibilité avec deux algorithmes d'apprentissage automatique et trois stratégies de fusion. Afin d'apporter une certaine interprétabilité à nos résultats de prédiction, nous décrivons la méthode de sélection de paramètres que nous avons appliquée pour réduire notre ensemble de paramètres. Nous fournissons ensuite une analyse qualitative de notre système. Enfin, nous présentons une ouverture de notre méthode de prédiction aux apprenants de notre second corpus (CAF-al) afin d'étudier la généricité de notre système de prédiction.

1

État de l'art

Sommaire

Les années 1990 ont marqué une phase décisive dans les recherches sur la compréhensibilité de la parole d'apprenants L2. Bien que l'intérêt pour ce domaine ait connu une augmentation notable entre les années 1980 et 1990, les travaux publiés par le chercheur Munray J. Munro dans les années 1990 ont joué un rôle crucial dans le développement de ce champ de recherche. Plus particulièrement, ses études concernant l'intelligibilité, la compréhensibilité et le degré d'accent de la parole d'apprenants L2 (Munro et Derwing, 1995a) ont ouvert la voie sur la caractérisation de la compréhensibilité de la parole. En effet, de nombreuses recherches ont été menées dans le but de comprendre les différents facteurs permettant de caractériser et de distinguer la compréhensibilité de l'intelligibilité et du degré d'accent. Bien que souvent évoquée ou liée à l'intelligibilité, nous allons essayer dans ce chapitre, et en nous appuyant sur la littérature, de définir la compréhensibilité de manière claire. Nous allons également identifier les différents éléments pertinents, qu'ils soient linguistiques ou non, susceptibles d'exercer une influence sur celle-ci. Nous discuterons finalement de l'intégration de ces éléments dans le processus permettant d'établir une mesure automatique de la compréhensibilité de la parole.

1.1 Différents points de vue sur la compréhensibilité

L'établissement d'une définition de la compréhensibilité de la parole s'avère être le point de départ de toute analyse. Sans une définition claire et précise, nous ne sommes pas en mesure d'affirmer que le phénomène évalué relève bel et bien de la compréhensibilité de la parole. Nous présentons dans cette section les définitions que nous pouvons trouver dans la littérature en rapport avec le domaine de l'apprentissage des langues et nous analysons leur potentiel impact sur l'évaluation de la compréhensibilité.

1.1.1 Compréhensibilité ou intelligibilité ?

Lorsque nous nous intéressons à la compréhensibilité de la parole, il est difficile, voire impossible, de ne pas aborder l'intelligibilité de la parole. Il est en effet intéressant de noter que la plupart des études qui traitent de la compréhensibilité de la parole d'apprenants d'une langue étrangère s'appuient sur la définition de l'intelligibilité pour introduire ce concept. D'après Levis (2006), dont le positionnement sert de point de départ pour certaines études (Isaacs et Trofimovich, 2012; Trofimovich et Isaacs, 2012; Saito *et al.*, 2017), l'intelligibilité peut être définie de deux manières, à savoir au « sens strict » et au « sens large ». Ces deux niveaux peuvent se distinguer grâce à leurs méthodes d'évaluation. Dans son sens strict, l'intelligibilité est généralement mesurée en termes d'exactitude des transcriptions orthographiques, réalisées par l'évaluateur lui-même (Munro et Derwing, 1995b) (est-ce que la transcription *entière* est correcte ?) ou par le biais de questions de compréhension au sujet du contenu (Hahn, 2004). Ces

méthodes apportent une évaluation d'un point de vue phonético-phonologique (transcription de productions) et sémantico-discursif (compréhension du contenu). Dans son sens large, l'intelligibilité est généralement mesurée *via* une évaluation subjective de la facilité de compréhension (Munro et Derwing, 1999). Cette mesure se résume à demander à un auditeur d'évaluer la quantité d'efforts à fournir pour comprendre une production.

La compréhensibilité relèverait ainsi de la capacité d'un auditeur à comprendre un discours. Les éléments sous-jacents à la signification des mots « capacité » et « comprendre » dans cette définition paraissent cependant complexes à décoder. Il n'est pas spécifié, par exemple, s'il s'agit d'un effort à fournir pour une compréhension au niveau phonético-phonologique (l'auditeur réussit à discriminer les différents mots énoncés avec plus ou moins d'efforts) ou au niveau sémantico-discursif (l'auditeur *comprend* la production d'un point de vue du contenu, avec plus ou moins d'effort), ou s'il s'agit simplement de faire ressortir une impression (est-ce que l'auditeur a *l'impression* que c'est dur à comprendre, que cela nécessite beaucoup d'effort ?). Il est alors évident que la différenciation entre intelligibilité (dans son sens strict) et compréhensibilité, à partir de ces définitions, se voit être une tâche complexe. Aucun véritable consensus n'existait vraiment quant à la manière de définir et de différencier ces deux concepts, jusqu'à la récente étude de Pommée *et al.* (2022) réalisée dans le cadre de la parole pathologique.

Cette étude se concentre sur l'élaboration des définitions d'intelligibilité, de compréhensibilité, et des méthodes d'évaluations qui leur sont propres. L'application d'une méthode DELPHI (Jones et Hunter, 1995; McMillan *et al.*, 2016) auprès de différents experts (cliniciens, linguistes et informaticiens) a permis d'aboutir à un consensus. L'intelligibilité est ainsi définie comme étant la reconstruction d'un message au niveau acoustico-phonétique, et la compréhensibilité comme étant « la reconstruction d'un message au niveau sémantico-discursif, à l'issue de la reconstruction acoustico-phonétique » (Pommée *et al.*, 2022, p. 31). Ces définitions permettent de mieux concevoir les frontières qui séparent intelligibilité et compréhensibilité, qui étaient jusqu'à présent difficilement distinguables. Cependant, la reconstruction en termes de compréhensibilité ne peut avoir lieu qu'après la reconstruction acoustico-phonétique, dans le cas d'un message transmis oralement (processus dit *bottom-up* ; Fontan, 2012). En effet, même si la compréhensibilité agit au niveau de la reconstruction sémantique d'un message, cette reconstruction ne peut prendre place de manière effective que si la compréhension au niveau de l'intelligibilité se révèle suffisante.

La compréhensibilité engloberait donc une partie d'intelligibilité (voir figure 1.1). Néanmoins, une bonne intelligibilité n'induirait pas forcément une bonne compréhensibilité. En d'autres termes, un message oral pourrait être parfaitement énoncé (discrimination parfaite des différents phonèmes ou sons), sans pour autant véhiculer de sens correct. À l'inverse, un message pourrait être à la base complètement compréhensible d'un point de vue sémantique, mais déformé par une mauvaise prononciation au point d'être rendu incompréhensible.

Les études que nous analysons dans la suite de ce chapitre pour caractériser et

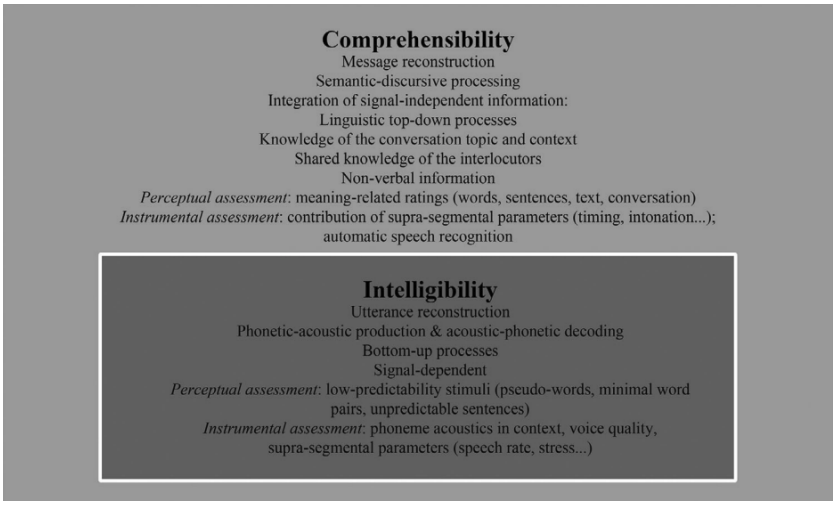


FIGURE 1.1 – L'intelligibilité et la compréhensibilité dans les productions orales (Pommée *et al.*, 2022, p. 13).

comprendre les éléments impactant la compréhensibilité de la parole la définissent également en termes de capacité à comprendre un discours, équivalent à la définition d'intelligibilité au sens large de Levis (2006). Dans la suite du manuscrit, nous nous référerons à la compréhensibilité de la parole telle qu'énoncée par Pommée *et al.* (2022) et définie par Woisard *et al.* (2013), à savoir « Capacité de l'auditeur à interpréter le sens du message oral produit par un locuteur, sans tenir compte de la précision ou de la justesse phonétique ou lexicale ».

1.1.2 L'importance d'une définition claire et explicite

Le processus d'évaluation de la compréhensibilité de la parole d'apprenants L2 se déroule généralement en quatre étapes :

1. choix d'une tâche de production orale,
2. enregistrement d'apprenants de langue étrangère,
3. définition de la compréhensibilité de la parole auprès d'auditeurs,
4. évaluation de la compréhensibilité de la parole par ces auditeurs (attribution d'un score).

Les auditeurs, après avoir reçu une brève explication de ce qu'est la compréhensibilité, ainsi qu'une définition de celle-ci, sont amenés à évaluer les productions des apprenants sur une échelle pouvant être continue à 1000 points (Saito et Akiyama, 2017) ou discrète à cinq points (Isaacs et Thomson, 2020) ou neuf points (Suzuki et Kormos, 2020). Cette étape de définition se voit être essentielle pour le bon déroulement de l'évaluation et de l'étude en cours. Parmi les définitions les plus couramment utilisées, la compréhensibilité est décrite comme étant « le degré de facilité ou de difficulté de compréhension du discours L2 » (Derwing et Munro, 2009; Saito *et al.*, 2017),

ou bien « à quel point le locuteur est facile à comprendre » (Isaacs et Trofimovich, 2012; Isaacs et Thomson, 2020), ou encore « la quantité d'efforts nécessaires pour comprendre ce que dit un locuteur » (Saito *et al.*, 2023). Comme discuté dans la sous-section précédente, ces définitions peuvent mettre les évaluateurs face à de nombreux problèmes : l'évaluation fait-elle référence à la compréhension réelle ou bien à l'effort, la facilité ou la difficulté ressentie pour comprendre ? De même, bien qu'il existe une part de subjectivité dans la compréhensibilité, car nous avons chacun notre propre conception du fait de *comprendre*, et ce malgré une explication claire et précise, ces différentes définitions laissent la place à une plus large interprétation. Lorsqu'une trop grande liberté subsiste dans la manière d'interpréter un phénomène, qu'en est-il de son évaluation ?

Prenons l'exemple de l'étude menée par Isaacs et Trofimovich (2012), centrée sur l'analyse des facteurs linguistiques pouvant influencer la compréhensibilité d'apprenants d'anglais lors d'une tâche de narration d'histoire imagée. Une des questions de recherche à laquelle cette étude souhaite répondre consiste à analyser les aspects linguistiques exerçant le plus d'influence sur l'évaluation de la compréhensibilité des apprenants par trois enseignants d'ALE (Anglais Langue Étrangère). La définition de compréhensibilité proposée pour mener à bien cette évaluation correspond à *« how easy the speaker is to understand »* (*à quel point le locuteur est facile à comprendre*). Après la phase d'évaluation de la compréhensibilité de la parole, ceux-ci sont amenés à indiquer s'ils l'ont interprétée comme étant la compréhensibilité moyenne de chaque mot pris individuellement, comme étant la compréhensibilité du récit, ou s'ils avaient adopté une tout autre interprétation. Les réponses des enseignants indiquent que deux d'entre eux ont évalué la compréhensibilité globale des récits, tandis que le troisième a évalué la compréhensibilité au niveau de chaque mot pris individuellement. Ces différences d'interprétation soulignent le caractère trop libre de la définition, et « suggèrent que la compréhensibilité devrait être définie de manière plus précise dans le cadre de la recherche et de l'évaluation en L2 que seulement *facilité de compréhension* » (Isaacs et Trofimovich, 2012, p. 489-490). D'après nous, cette conclusion pourrait également s'appliquer aux définitions faisant référence à *la quantité d'efforts nécessaires*. Les évaluateurs pourraient, en plus, bénéficier d'explications plus détaillées autour de la compréhensibilité, afin de s'assurer que les évaluations s'opèrent selon la même interprétation.

1.2 La compréhensibilité en tant que concept linguistique multi-niveaux

Il est intéressant, en vue de l'élaboration d'une mesure automatique, d'analyser et de comprendre les différents facteurs pouvant exercer une influence sur la compréhensibilité. Cette analyse a particulièrement intéressé les chercheurs du domaine de l'apprentissage des langues et a mené à des protocoles permettant une évaluation plus fine de la compréhensibilité de la parole. En plus de l'évaluer en attribuant un

score à un apprenant, les compétences linguistiques de celui-ci font également l'objet d'une évaluation subjective, soit par l'auditeur lui-même (Crowther *et al.*, 2018), soit par un ou plusieurs experts linguistes (Saito, 2015). Les évaluations concernent principalement cinq niveaux, à savoir la phonologie/phonétique (regroupant les niveaux segmentaux et suprasegmentaux), le lexique, la syntaxe et le discours. À la suite de ces évaluations, de potentielles corrélations avec la compréhensibilité peuvent être mesurées, permettant alors d'identifier les dimensions linguistiques exerçant une influence sur celle-ci.

Les différentes études menées sur le sujet ont pu démontrer que la compréhensibilité pouvait être vue comme un concept linguistique multi-niveaux. La table 1.1 présente les dimensions linguistiques généralement évaluées et les mesures qui leur sont associées (Isaacs et Trofimovich, 2012; Saito *et al.*, 2017; Crowther *et al.*, 2018; Suzuki et Kormos, 2020).

TABLE 1.1 – Dimensions linguistiques et mesures associées

Dimensions linguistiques	**Mesures**
Phonologie/phonétique	Débit de parole Pauses (pleines et silencieuses) et autres disfluences (auto-correction, répétitions, etc.) Erreurs segmentales (substitutions phonémiques) Erreurs de structure syllabique (insertions et délétions de consonnes et voyelles) Erreurs rythmiques (réduction de voyelles) Erreurs d'intonation (hauteur de ton) Erreurs d'accentuation au niveau du mot (accent mal placé)
Lexique	Richesse lexicale (diversité, densité et sophistication du lexique) Appropriation lexicale (erreurs lexicales, par exemple mauvais choix de vocabulaire)
Syntaxe	Complexité grammaticale Précision grammaticale (accord sujet-verbe, ordre des mots, temps verbaux)
Discours	Cohésion du discours (présence de connecteurs du discours) Richesse du discours (éléments de narration, nombre de catégories de propositions différentes)

Les paramètres issus des dimensions linguistiques présentées dans cette table n'exercent pas tous la même influence sur la compréhensibilité de la parole. En effet, selon le type de tâche de production orale employée, que nous analyserons plus en détail dans la section 1.3.3, la compréhensibilité se voit être plus ou moins corrélée à certains aspects linguistiques de la parole. Nous pouvons tout de même observer des similitudes dans les résultats de plusieurs travaux. Par exemple, lorsque nous analysons les études de Saito *et al.* (2016a), Crowther *et al.* (2018) et Saito et Shintani (2016), qui partagent la même tâche de production orale, les évaluations de la compréhensibilité de la parole par des auditeurs natifs démontrent que plusieurs dimensions linguistiques, aussi bien les dimensions segmentales/suprasegmentales que lexicales, syntaxiques et discursives, sont liées à la compréhensibilité. Ces résultats se trouvent être en adéquation avec ceux obtenus par Isaacs et Trofimovich (2012), Saito *et al.* (2016c) et Crowther *et al.* (2015a), qui associent la compréhensibilité à des considérations suprasegmentales (phonologie), à la phonétique, au lexique et à la syntaxe.

La table 1.2 recense les coefficients de corrélation obtenus entre la compréhensibilité de la parole et 18 paramètres linguistiques dans différentes études de 2015 à 2020. Les deux études de Crowther (Crowther *et al.*, 2015a, 2018) portant sur l'influence de respectivement deux et trois tâches de production orale sur la compréhensibilité de la parole, nous faisons apparaître les résultats obtenus à l'issue de ces tâches, à

savoir les tâches de production orale de type IELTS (*International English Language Testing System*)[2] et TOEFL IBT (*Test Of English as a Foreign Language, Internet Based Test*)[3] pour l'étude de Crowther *et al.* (2015a) et les tâches de description d'image, de production orale IELTS et TOEFL IBT pour l'étude de Crowther *et al.* (2018) (nous détaillons ces méthodes dans la section 1.3.3).

2. https://www.ielts.org/
3. https://www.toeflibt.fr/

TABLE 1.2 – Corrélation (en valeur absolue) entre les paramètres linguistiques et la compréhensibilité de la parole de plusieurs études de la littérature. Un coefficient de corrélation noté « + » se situe entre 0,3 et 0,5, « ++ » entre 0,5 et 0,7 et « +++ » indique un coefficient supérieur à 0,7. Les cases grisées indiquent que les paramètres n'ont pas été étudiés.

		Saito et Shintani (2016)	**Crowther *et al.* (2015a)**		**Saito (2015)**	**Saito *et al.* (2016b)**	**Saito *et al.* (2016c)**	**Saito et Akiyama (2017)**	**Suzuki et Kormos (2020)**	**Crowther *et al.* (2018)**			**Crowther *et al.* (2016)**	**Saito *et al.* (2016a)**
			IELTS	TOEFL						Image	IELTS	TOEFL		
Paramètres segmentaux et suprasegmentaux	Erreurs segmentales	+++	+++	+++	+++				++	+++	+++	+++	++	+++
	Structure syllabique								+++				+	
	Accentuation (mot)	++	++	++	+++				+	++	++	+++	+++	++
	Rythme		+	++					++	+++	+++	+++	+++	
	Intonation	+	++	++	++					++	++	+++	++	++
	Débit de parole	++	++	++	++	+++		+		+++	+++	+++	++	++
	Pauses pleines						+++						+	
	Pauses silencieuses								+++					
	Répétitions/auto-correction												++	
Paramètres lexicaux et syntaxiques	Diversité lexicale					+	+++	+	++	++	++	+++	+++	
	Densité lexicale									++	++	+++	+++	
	Sophistication lexicale		+	+			+			++	++	+++		
	Appropriation lexicale	+	++	++	+	++	+++	+		++	++	+++	++	+
	Complexité grammaticale		++	+						++	++	+++	++	
	Appropriation morphologique	+	+	+	++	+	+		+++	++	++	+++	++	++
	Ordre des mots	+	+	+	++				++	++	++	+++	++	++
Paramètres discursifs	Cohésion du discours												+	
	Richesse du discours		+							++	++	+++	++	

Nous pouvons observer que les résultats sont plutôt hétérogènes. Tous les paramètres sont corrélés, ne serait-ce même que faiblement, au moins une fois à la compréhensibilité de la parole. Les erreurs segmentales sont systématiquement corrélées à plus de 0,5 (ou moins de -0,5), ce qui indique que la prononciation joue un rôle important. Aucun paramètre n'est perçu comme étant systématiquement corrélé à la compréhensibilité, mais nous observons tout de même que l'appropriation lexicale et l'appropriation morphologique ont un coefficient de corrélation supérieur à 0,3 (ou inférieur à -0,3) dans neuf des dix études présentées.

D'après Saito *et al.* (2016a), les différentes dimensions linguistiques représentent ainsi des facteurs nécessaires pour pouvoir extraire le sens d'une production orale. La compréhensibilité est ainsi négativement affectée par une production orale contenant, à titre d'exemple, des erreurs phonético-phonologiques (Suzukida et Saito, 2021, les erreurs segmentales par exemple), des disfluences (Suzuki et Kormos, 2020, au niveau de la fluence phonétique), et des erreurs syntaxiques (Varonis et Gass, 1982). En effet, concernant les erreurs syntaxiques, la structuration de phrases contenant des erreurs grammaticales peut mener à des ambiguïtés et à des difficultés à saisir le sens de ce qui est exprimé. Enfin, malgré l'impact négatif que peuvent avoir certains paramètres phonético-phonologiques, la présence de ce type d'erreurs n'entrave pas systématiquement la compréhensibilité de la parole d'apprenants de langue étrangère (Munro et Derwing, 1995a). De la même façon, un apprenant peut rester compréhensible malgré la présence incontestable d'un accent dû à sa L1 (Isaacs et Trofimovich, 2012).

Nous pouvons donc envisager la compréhensibilité de la parole comme représentant un concept multi-niveaux et influencée par la phonologie/phonétique, le lexique, la syntaxe et le discours d'un apprenant. Néanmoins, l'influence des paramètres issus de ces dimensions linguistiques reste variable. Comme nous avons pu l'observer, les erreurs phonético-phonologiques peuvent affecter la compréhensibilité dans certains cas, tandis qu'elles n'ont aucun impact dans d'autres. Chaque dimension linguistique joue donc un rôle dans la compréhensibilité. Il devient alors intéressant d'essayer de parvenir à une évaluation automatique de la compréhensibilité de la parole par le biais d'une évaluation des différents paramètres linguistiques qui peuvent la composer.

1.3 Facteurs non linguistiques affectant la compréhensibilité

Comme nous venons de le voir dans la section précédente, la compréhensibilité peut être perçue comme une combinaison de plusieurs dimensions linguistiques. Néanmoins, cette seule combinaison ne suffit pas à expliquer la compréhensibilité d'un apprenant d'une langue étrangère. Comme suggéré dans l'étude de Varonis et Gass (1982), la compréhensibilité d'un apprenant est égale à une somme de plusieurs facteurs linguistiques et sociaux (la familiarité entre l'auditeur et l'apprenant, par exemple), de sa langue maternelle et du sujet abordé (que nous pouvons aussi qualifier de contexte d'énonciation). Il devient alors inconcevable de proposer une mesure de

la compréhensibilité de la parole, qu'elle soit automatique ou non, sans prendre en compte ces différents facteurs non linguistiques. Nous abordons ainsi dans cette section trois facteurs que nous retrouvons de manière récurrente dans la littérature, à savoir le rôle du profil de l'auditeur, le rôle du locuteur, particulièrement de sa langue maternelle, et le rôle de la tâche de production orale.

1.3.1 La compréhensibilité du point de vue de l'auditeur

Le premier facteur non linguistique qui peut potentiellement affecter la compréhensibilité de la parole est le profil de l'auditeur. Il ne fait presque aucun doute qu'un auditeur familier avec l'accent L1 d'un apprenant a moins de mal à le comprendre qu'un auditeur non familier (Gass et Varonis, 1984; Tauroza et Luk, 1997). La familiarité avec un accent permet de reconnaître avec plus de facilité des mots en dépit de leur prononciation atypique. De même, et selon Bradlow et Bent (2008), « les caractéristiques phonétiques de la parole présentant un accent étranger sont très systématiques et assez constantes chez les locuteurs issus de la même langue maternelle » (Bradlow et Bent, 2008, p. 708, « *the phonetic characteristics of foreign-accented speech are highly systematic and quite consistent across talkers from the same native language background* »), ce qui implique que lorsqu'un auditeur est familier avec un locuteur présentant un certain accent L1, il est moins difficile pour lui de comprendre d'autres locuteurs partageant cette même L1. Néanmoins, l'étude menée par Bergeron et Trofimovich (2017) et centrée sur les effets de la tâche de production orale et du profil des auditeurs a révélé des résultats contraires. Dans cette étude, la compréhensibilité d'apprenants de langue maternelle espagnole est évaluée par des auditeurs français, avec seulement une moitié présentant des connaissances en espagnol. La familiarité des auditeurs avec la L1 n'a pas montré d'impact sur la compréhensibilité des apprenants, mais une différence significative a tout de même pu être observée dans les résultats des évaluations de paramètres phonético-phonologiques et lexicogrammaticaux. La corrélation entre la compréhensibilité et les paramètres de phonologie/phonétique s'est avérée être plus élevée pour les évaluateurs familiers avec l'espagnol (r = 0,83) par rapport aux évaluateurs non familiers (r = 0,57). Les paramètres lexicogrammaticaux ont également montré une corrélation plus élevée avec la compréhensibilité chez les auditeurs familiers (r = 0,67). La familiarité avec l'accent L1 représente donc un facteur nécessaire à prendre en compte lors de l'évaluation de la compréhensibilité.

L'expérience linguistique ou l'expérience dans l'enseignement peut également influencer la compréhensibilité perçue, car les connaissances linguistiques de la langue cible permettent de mieux appréhender la production orale d'un apprenant. Nous qualifions d'auditeur expérimenté un auditeur familier avec la L2 de par sa profession, c'est-à-dire ayant de l'expérience, que ce soit d'un point de vue linguistique (études et/ou diplôme en linguistique) ou d'un point de vue enseignement (enseignement de la L2). Les conclusions quant à cet aspect semblent hétérogènes. Dans l'étude menée par Saito *et al.* (2017), la compréhensibilité de la parole est évaluée par 20 auditeurs natifs, la moitié représentant un groupe expérimenté, l'autre moitié un groupe non

expérimenté (évaluateurs dits novices). Les apprenants évalués par les auditeurs expérimentés se sont vus attribuer des scores de compréhensibilité plus élevés que lors de l'évaluation par des auditeurs novices, et ce de manière significative. Toutefois, certaines études semblent ne pas montrer de différence significative dans l'évaluation de la compréhensibilité de la parole selon l'expérience des auditeurs (Kennedy et Trofimovich, 2008; Isaacs et Thomson, 2013). Quant à l'étude de Trofimovich et Isaacs (2012), celle-ci ne compare pas à proprement parler le résultat de l'évaluation de la compréhensibilité menée par différents auditeurs (experts et novices), mais elle fournit une analyse des dimensions linguistiques pouvant l'affecter. Les auditeurs experts sont représentés par trois enseignants d'ALE, ayant chacun une expérience d'au moins dix ans dans l'enseignement de l'anglais, tandis que les auditeurs novices sont représentés par un groupe de 60 Anglais natifs, encore étudiants au moment de l'étude, venant de disciplines non linguistiques et indiquant une faible familiarité avec la L1 des apprenants (le français). D'après les résultats, les auditeurs novices ont tendance à faire reposer leur jugement de la compréhensibilité sur des aspects de fluence phonétique, de richesse lexicale et de précision grammaticale, tandis que les auditeurs experts ne se reposent pas sur la fluence phonétique, mais sur l'utilisation du vocabulaire et la grammaire. Tout comme la familiarité avec l'accent L1, il ne fait aucun doute que nous devons prendre en compte l'expérience d'un auditeur lors de l'évaluation de la compréhensibilité de la parole. Cette dernière peut en effet différer, selon le profil de l'auditeur, de même que les dimensions linguistiques qui pourraient également être évaluées.

Enfin, il est juste de penser qu'il existe une différence de perception, et donc de compréhension, entre des auditeurs de différentes langues maternelles à l'écoute de la même production. Selon Saito *et al.* (2019), une des caractéristiques principales de cette différence de compréhension se trouve dans la similarité de la L1 de l'auditeur avec la L2 du locuteur. En effet, en analysant cette étude, lorsque la compréhensibilité d'apprenants japonais d'anglais est évaluée par huit groupes d'auditeurs différents, les résultats indiquent qu'elle est plus élevée lorsqu' elle est évaluée par des auditeurs ayant un profil linguistique plus similaire à la langue anglaise, notamment à l'anglais comportant un accent japonais. La parole a également été évaluée comme étant plus compréhensible par les auditeurs conscients du rôle que la prononciation et le lexique jouent dans une communication réussie. Nous devons tout de même moduler ces conclusions par le fait que la compréhensibilité est généralement définie comme étant l'effort de compréhension. Un auditeur ayant une L1 proche de la L2 aurait, de ce fait, moins d'efforts à fournir.

Nous pouvons observer par le biais des différentes études abordées dans cette sous-section que le profil de l'auditeur est un aspect intéressant à prendre en compte lors de l'évaluation de la compréhensibilité de la parole. Cette dernière serait donc modulable, rendant un apprenant plus ou moins compréhensible selon son auditeur. Ainsi, dans le cas où nous souhaiterions évaluer la compréhensibilité d'apprenants dans un contexte d'immersion dans un pays étranger, il serait intéressant de l'évaluer par

le biais d'auditeurs natifs, novices, et sans familiarité élevée avec l'accent L1. Si, par contre, nous souhaitons obtenir une plus grande précision quant aux influences des différentes dimensions linguistiques dans un contexte d'enseignement, il pourrait être judicieux de faire évaluer la compréhensibilité par des auditeurs natifs experts. Quelle que soit la situation, **le profil de l'auditeur est donc à prendre en considération**.

1.3.2 La compréhensibilité du point de vue du locuteur

En plus de varier selon le profil de l'auditeur, la compréhensibilité représente également un concept propre à chaque apprenant. Des différences notables peuvent ainsi être observées selon le profil linguistique des apprenants, par exemple selon leurs langues maternelles. En effet, celle-ci a un impact significatif sur la production L2, notamment d'un point de vue phonético-phonologique (Crowther *et al.*, 2015b; Isaacs et Thomson, 2020). Les résultats de l'étude menée par Isaacs et Thomson (2020) montrent bien que la langue maternelle des apprenants exerce une influence sur leur compréhensibilité, quel que soit le niveau de familiarité des auditeurs avec la L1. Dans cette étude, la compréhensibilité d'apprenants d'ALE de L1 slave ou mandarin est évaluée par des natifs ayant une familiarité plus forte avec l'accent mandarin. Malgré cette familiarité, les productions des apprenants de L1 slave ont été évaluées comme étant plus compréhensibles que les productions des apprenants de L1 mandarin, et ce de manière significative. Selon les auteurs, ces résultats peuvent provenir du fait que le mandarin contient plus de divergences phonologiques avec l'anglais comparé aux langues slaves. La distance phonologique entre la L1 d'un apprenant et la langue cible pourrait ainsi expliquer ces observations (Bongaerts *et al.*, 2000).

De leur côté, les auteurs de Crowther *et al.* (2015b) ont analysé la compréhensibilité d'apprenants d'ALE ayant des langues maternelles chinoises, hindi/ourdou et farsi. Leurs résultats révèlent que, de manière générale, les apprenants de L1 farsi ont été évalués comme étant les plus compréhensibles, suivis des apprenants de L1 hindi/ourdou et enfin des apprenants de L1 chinois. La compréhensibilité s'est révélée ne pas être liée aux mêmes composantes linguistiques pour chaque groupe. Pour les apprenants L1 chinois, elle a été fortement associée à la phonologie/phonétique (taux d'erreurs segmentales), mais également à des aspects lexicogrammaticaux (appropriation lexicale, richesse lexicale, complexité syntaxique) pour les apprenants L1 hindi/ourdou. Elle ne s'est cependant révélée associée à aucun de ces aspects pour les apprenants L1 farsi.

La langue maternelle d'un apprenant exerce donc une influence sur la compréhensibilité de la parole, d'un point de vue phonético-phonologique, et ce indépendamment de la familiarité que pourrait avoir un auditeur avec l'accent L1. De plus, et toujours selon la langue maternelle, les dimensions linguistiques qui entrent en jeu ne sont pas les mêmes. Tandis que la compréhensibilité peut être fortement associée à des aspects phonético-phonologiques pour des apprenants d'une certaine L1, elle peut également être fortement associée à d'autres aspects linguistiques, tel que le lexique

ou la syntaxe. Les dimensions linguistiques peuvent également toutes contribuer à la compréhensibilité de manière égale, sans qu'aucune ne se différencie de manière forte.

1.3.3 La compréhensibilité du point de vue de la tâche de production orale

Nous venons de discuter des potentiels facteurs linguistiques et non linguistiques pouvant avoir une influence sur la compréhensibilité de la parole. Un dernier facteur non linguistique qu'il est important de prendre en compte concerne la tâche de production orale.

De manière générale, l'évaluation de la compréhensibilité s'effectue par le biais de tâches de production orale relativement libres. Les méthodes les plus communément utilisées pour recueillir la parole d'apprenants L2 sont listées ci-dessous :

— la narration d'histoires imagées : un exposé basé sur une histoire prenant généralement place sur sept ou huit images différentes [4] (Derwing *et al.*, 2009; Isaacs et Trofimovich, 2012; Saito *et al.*, 2017; Crowther *et al.*, 2018; Isaacs et Thomson, 2020),

— l'argumentation : une expression d'opinion sur une affirmation précise. Par exemple, exprimer son opinion sur l'affirmation « *Les Jeux Olympiques de Tokyo 2020 vont favoriser la croissance économique du Japon.* » (Suzuki et Kormos, 2020),

— l'épreuve orale du IELTS dite *long turn* : argumentation autour d'un sujet précis, à savoir « *Décrivez un évènement sportif que vous aimez regarder* » ou « *Décrivez un métier que vous voudriez exercer dans le futur* » (Crowther *et al.*, 2015a, 2018)

— l'épreuve de résumé oral du TOEFL IBT : une lecture d'un passage d'environ 100 mots, écoute de 80-90 secondes d'un audio traitant du même sujet, puis réponse à une question liée au contenu des deux sources. Les sujets abordés portent sur les effets d'audience en psychologie ou sur l'explication comportementale en sociologie (Crowther *et al.*, 2015a, 2018)

— la description d'images : un exposé de plusieurs images différentes (généralement sept), avec obligation d'utiliser trois mots-clés prédéfinis pour chaque image (Saito *et al.*, 2019).

Selon Robinson (2005) et Skehan (2009), les apprenants utilisent différentes ressources linguistiques pour réaliser différentes tâches de production orale, notamment des ressources phonético-phonologiques, lexicales et grammaticales. Le type de tâche peut affecter les productions orales de différentes manières, tant aux niveaux segmental et suprasegmental (Tarone, 1983) qu'aux niveaux lexical et syntaxique (Robinson, 2005; Skehan, 2009). Par exemple, les différences de temps de planification, de temps de réalisation, d'objectifs de réalisation et de familiarité avec le thème ou le sujet permettent de discerner des variations aux niveaux lexical et grammatical (Foster et

4. À titre d'exemple, le support le plus utilisé dans ce domaine peut être téléchargé sur le site IRIS `https://www.iris-database.org/details/G1cDs-5ebVE`

Skehan, 1996; Yuan et Ellis, 2003; Robinson, 2005). Le temps de planification d'un apprenant lui permettrait en effet de réfléchir plus longuement au vocabulaire et à la construction des phrases qu'il va produire. Il sera également plus à l'aise dans le choix de son lexique et dans la construction grammaticale de ses phrases s'il présente une familiarité plus élevée avec le sujet traité.

En plus d'observer une différence dans les ressources utilisées, le type de tâche a également un impact direct sur la compréhensibilité. Nous l'observons par exemple dans l'étude menée par Crowther *et al.* (2018). Dans cette étude, trois tâches de production orale sont comparées, à savoir la narration d'histoire imagée, l'épreuve IELT *long turn* et l'épreuve orale TOEFL IBT, en termes de leurs influences sur les évaluations de la compréhensibilité des apprenants. Les résultats montrent que la compréhensibilité est généralement évaluée comme étant plus élevée lors de la réalisation de l'épreuve IELTS, comparée aux épreuves de narration d'histoire imagée et du TOEFL IBT. Cette différence pourrait provenir de la familiarité avec le sujet traité : tandis que l'épreuve IELTS permet aux apprenants de répondre à des questions faisant appel à leur expérience personnelle, les deux autres tâches les contraignent plutôt à réaliser une production sur des sujets bien spécifiques, sans relation avec une quelconque expérience personnelle. Dans leur précédente étude, également sur le lien entre la compréhensibilité et les tâches de production orale de type IELTS et TOEFL IBT (Crowther *et al.*, 2015a), une analyse plus fine a été réalisée. Pour l'épreuve IELTS, certaines variables phonético-phonologiques telles que les erreurs segmentales et le rythme de la parole se sont avérées avoir un plus grand impact sur la compréhensibilité. Pour l'épreuve TOEFL IBT, la compréhensibilité a également été associée à des variables phonético-phonologiques, mais aussi à des variables lexicogrammaticales (appropriation lexicale, complexité et précision syntaxique). Nous pouvons ainsi émettre l'hypothèse, après analyse de ces études, que lorsque les apprenants sont familiers avec le sujet traité, ils seront plus à l'aise concernant leurs choix lexicogrammaticaux et leurs niveaux de compréhensibilité ne seraient différenciables que grâce à leurs divergences phonético-phonologiques. En revanche, lorsque la familiarité ne fait pas partie de l'équation, et que le sujet à traiter est relativement contraint, des variables lexicogrammaticales entreraient en jeu dans la différenciation de leurs niveaux de compréhensibilité.

D'un autre point de vue, les autres études essayant de définir les dimensions linguistiques qui entrent en jeu dans l'évaluation de la compréhensibilité travaillent généralement sur une seule tâche, celle de la narration d'histoire imagée. Leurs résultats ne peuvent pas permettre de comparer différentes tâches de production orale, mais donnent tout de même un aperçu détaillé sur les dimensions linguistiques les plus corrélées à la compréhensibilité pour une tâche particulière. Ainsi, pour la tâche de narration d'histoire imagée, les aspects temporels de la parole (variables phonologiques, phonétiques) et les variables lexico-grammaticales ont tous les deux tendance à contribuer au niveau de compréhensibilité des apprenants (Derwing *et al.*, 2004, 2008; Trofimovich et Isaacs, 2012; Saito *et al.*, 2016c; Saito et Shintani, 2016).

Ces différents résultats doivent toutefois être modérés par plusieurs facteurs. Premièrement, la population d'apprenants évalués n'est pas la même parmi toutes les études. Nous retrouvons différents profils linguistiques, étant donné qu'ils ne partagent pas tous la même L1. Deuxièmement, la population d'évaluateurs est également différente d'une étude à l'autre. Ceux-ci n'ont pas tous la même familiarité avec l'accent des apprenants ni la même connaissance linguistique ou didactique. Enfin, les évaluations de la compréhensibilité de la parole et des différentes dimensions linguistiques n'ont pas été menées à partir des mêmes définitions. Malgré ces différents points, nous pouvons tout de même observer que **le type de tâche de production orale exerce une influence sur la compréhensibilité d'un apprenant**. Cette observation est donc à prendre en compte lors de la mise en place d'un protocole visant à évaluer la compréhensibilité.

1.4 L'évaluation automatique de la compréhensibilité

Dans les sections précédentes, nous avons étudié différentes approches permettant d'évaluer la compréhensibilité de manière subjective, par le biais d'évaluations humaines et d'analyses non automatiques. Cependant, l'évaluation de la compréhensibilité de la parole est souvent un processus long et fastidieux à mettre en place. Ces derniers temps, une attention spéciale a été portée sur le besoin d'évaluer la compréhensibilité de la parole de manière automatique, notamment pour pouvoir réduire le temps d'évaluation (O'Brien *et al.*, 2018), pour permettre aux apprenants de s'améliorer *via* des retours personnalisés (Trofimovich *et al.*, 2016), et pour permettre une évaluation continue plus aisée, à différents stades de l'apprentissage (Isaacs *et al.*, 2018).

À notre connaissance, et parmi les études récentes sur l'évaluation de la compréhensibilité de la parole, une seule a été menée et a ouvert la voie sur l'automatisation de l'évaluation. Il s'agit de l'étude de Saito *et al.* (2023), visant à prédire la compréhensibilité sur la base de paramètres extraits des productions orales de manière automatique. Les paramètres concernés proviennent des niveaux phonético-phonologique et mélodique (accentuation et intonation étant caractérisées comme des mesures mélodiques et non phonologiques d'après les auteurs). Ces paramètres sont ensuite utilisés pour prédire un score de compréhensibilité par le biais d'un modèle de régression linéaire multiple. Le plan de recherche de cette étude se décompose en trois parties, à savoir :

1. une vérification de la bonne prédiction de la compréhensibilité,
2. une vérification de la bonne prédiction de la compréhensibilité lorsque le jeu de données change,
3. une vérification de la bonne prédiction de la compréhensibilité lorsque la tâche de production orale change.

Les locuteurs participant à la première étape correspondent à 90 apprenants japonais d'ALE et 10 locuteurs anglophones (locuteurs de référence). La tâche de production orale utilisée ici est la description de sept images différentes. Lors de cette description, les apprenants ont pour consigne d'inclure trois mots-clés prédéfinis et propres à chaque image. Pour la seconde étape, la tâche de production orale est la même, mais les locuteurs changent : 40 apprenants japonais d'ALE et cinq locuteurs anglophones non inclus dans la première étape. Enfin, concernant la troisième et dernière étape, les locuteurs sont les mêmes que ceux de la deuxième étape, en revanche la tâche de production orale n'est pas la description d'images, mais une tâche d'argumentation semblable à l'épreuve orale du IELTS. Les locuteurs ayant participé à cette étude ont été, en amont, catégorisés selon leurs différents niveaux de compréhensibilité. Afin de former les groupes, le critère pris en compte a été la durée de séjour des apprenants dans le pays de langue L2. Quatre groupes ont ainsi été formés, à savoir le groupe d'apprenants :

— *inexpérimentés* : aucune expérience d'immersion dans un pays anglophone, soit un niveau de compréhensibilité bas,

— *modérément expérimentés* : au maximum cinq ans d'immersion dans un pays anglophone, soit un niveau de compréhensibilité moyen,

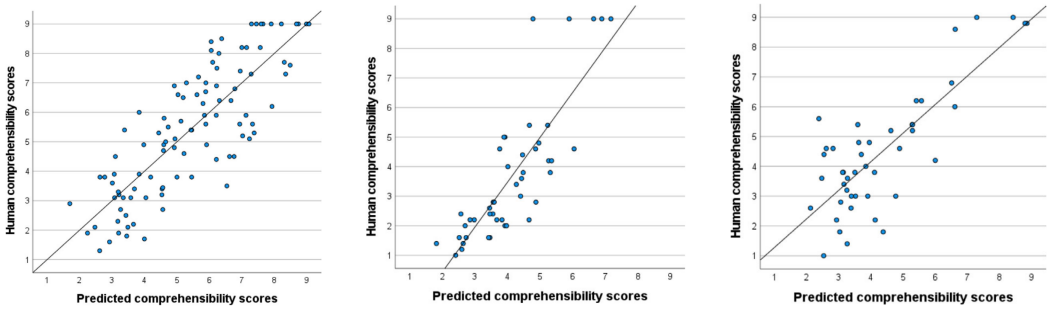
— *hautement expérimentés* : résidents d'un pays anglophone avec au minimum six ans d'immersion, soit un niveau de compréhensibilité élevé,

— *référence* : anglophones natifs.

Les évaluateurs lors de la première étape sont des auditeurs anglophones natifs ayant tous une grande familiarité avec les accents étrangers, mais une familiarité variable avec l'accent japonais. Pour les deuxièmes et troisièmes étapes, cinq auditeurs natifs ont participé, mais aucune information quant à leur familiarité avec l'accent étranger n'est indiquée. Tous les auditeurs ont évalué la compréhensibilité de la parole des apprenants en attribuant un score sur une échelle à neuf points (1 = difficile à comprendre, 9 = facile à comprendre) que nous qualifions de *scores terrains*. La compréhensibilité de la parole a été définie ainsi :

> *Ce terme fait référence à l'effort nécessaire pour comprendre ce que quelqu'un dit. Si vous pouvez comprendre (ce dont parle l'histoire illustrée) sans difficulté, alors le locuteur est hautement compréhensible. En revanche, si vous avez du mal et que vous devez écouter très attentivement, voire que vous ne comprenez pas du tout ce qui est dit, le locuteur est peu compréhensible.*
> (Saito *et al.*, 2023, p.13)

Les trois objectifs décrits plus haut ont été explorés en utilisant un modèle de régression linéaire multiple. Ce modèle a été construit lors de la première étape (prédiction de la compréhensibilité), donnant lieu à un coefficient de corrélation de Pearson de $r = 0{,}823$ ($p < 0{,}001$) entre les scores terrains et les scores prédits. La deuxième étape a résulté en un coefficient de corrélation de $r = 0{,}827$ ($p < 0{,}001$) et la troisième étape en un coefficient de corrélation de $r = 0{,}809$ ($p < 0{,}001$) (voir figure 1.2).



(a) Première étape.

(b) Même équation de régression que la première étape, jeu de données différent.

(c) Même équation de régression, jeu de données de l'étape 2, tâche de production orale différente.

FIGURE 1.2 – Nuages de points et droites des régressions linéaires multiples (Saito *et al.*, 2023, p. 251, p. 254 et p. 256).

Ces différents résultats permettent de conclure sur l'efficacité de la prédiction automatique de la compréhensibilité de la parole dans le cadre d'une tâche de description d'images effectuée par des apprenants japonais d'ALE, et à partir de paramètres extraits automatiquement aux niveaux phonético-phonologique et mélodique. De plus, l'utilisation de la même équation de régression pour prédire la compréhensibilité de nouveaux apprenants se montre concluante, de même lorsque la tâche de production orale change (épreuve orale de IELTS au lieu de description d'images). Ces résultats rendent cette méthodologie généralisable à d'autres contextes d'énonciation.

Cette étude ouvre la voie aux prédictions automatiques de la compréhensibilité de la parole, et permet de démontrer qu'un tel objectif est réalisable. Cependant, nous devons tout de même faire part de quelques remarques concernant les limites de cette étude :

— les paramètres utilisés pour la prédiction de la compréhensibilité sont uniquement issus de dimensions liées à l'intelligibilité (phonético-phonologie, mélodie). Les dimensions linguistiques qui jouent un rôle dans la compréhensibilité (voir section 1.2) devraient être incluses. Ceci pourrait améliorer les performances de prédiction. D'un point de vue pratique, et dans le but de fournir un outil d'évaluation automatique, les apprenants et enseignants pourraient bénéficier de plus de détails concernant la *qualité* des productions aux différents niveaux linguistiques.

— il n'est pas précisé si le jeu de données a été séparé en un jeu d'entraînement et un jeu de test, ou si la régression a été appliquée à l'ensemble des données. Dans le second cas, effectuer une régression sur l'ensemble des données à disposition pourrait avoir comme effet de biaiser l'évaluation des performances du modèle. Néanmoins, le fait que le modèle généralise bien aux données qu'il n'a pas encore vues (étapes 2 et 3) nous laisse supposer qu'il n'est pas dans une situation de surapprentissage (*overfitting*).

— comme cela n'est pas précisé, nous supposons qu'un seul modèle de régression

linéaire a été testé et utilisé. L'utilisation de différents modèles, qu'ils soient de type régression linéaire ou appartenant à la catégorie des méthodes d'ensemble (arbres de décision) pourrait apporter un bénéfice sur les performances.

1.5 Conclusion

Dans ce chapitre, nous avons défini le terme de compréhensibilité et discuté de sa différence avec l'intelligibilité. Tandis que cette dernière agit essentiellement au niveau acoustico-phonétique, la première agit également au niveau sémantico-discursif. Nous avons exploré les différentes manières dont la compréhensibilité pouvait être définie dans le domaine de l'apprentissage des langues, puis identifié les dimensions linguistiques qui la caractérisent. Après avoir montré qu'elle constituait une construction de dimensions linguistiques telles que la phonétique/phonologie, le lexique, la syntaxe et le discours, nous avons abordé les différents facteurs non linguistiques pouvant également entrer en considération. Enfin, nous avons analysé la possibilité de prédire de manière automatique la compréhensibilité de la parole.

Nous avons constaté au travers de cet état de l'art sur la compréhensibilité que plusieurs points de vue pouvaient être pris en compte et influaient sur ce qui pouvait être mesuré. En effet, les différentes dimensions linguistiques qui la caractérisent n'exercent pas systématiquement la même influence sur celle-ci. Le contexte d'énonciation est à prendre en compte, de même que le profil linguistique de l'apprenant. Également subjective, la compréhensibilité d'un apprenant varie selon l'auditeur, et peut être plus ou moins élevée au regard de la familiarité de celui-ci avec l'accent L1 par exemple. De même, nous avons observé que les définitions de la compréhensibilité pouvaient donner lieu à différentes interprétations. Il est alors nécessaire de mieux cadrer les expériences et d'interpréter les résultats d'évaluation en prenant en compte tous les paramètres. Lorsque nous allons chercher à évaluer la compréhensibilité, nous devons garder à l'esprit que ce que nous évaluons dépend fortement de la définition adoptée, du profil des apprenants et des évaluateurs, ainsi que du contexte d'énonciation employé.

L'évaluation automatique de la compréhensibilité de la parole est devenue un sujet primordial dans le domaine de l'apprentissage des langues. Étant donné que la compréhensibilité est caractérisée par plusieurs dimensions linguistiques, il est nécessaire de développer un outil permettant de la mesurer automatiquement sur la base de toutes ces dimensions linguistiques, à savoir la phonétique/phonologie, le lexique, la syntaxe et le discours.

Le chapitre suivant sera consacré à l'extraction et à la validation des paramètres multi-niveaux permettant de mesurer la compréhensibilité de la parole.

2

Sélection et validation de paramètres multi-niveaux

Sommaire

Comme énoncé dans le chapitre précédent (chapitre 1), la compréhensibilité englobe plusieurs niveaux linguistiques, tels que les niveaux phonético-phonologique (segmental et suprasegmental), lexical, syntaxique et discursif. À défaut d'avoir à disposition un corpus nous permettant de prédire la compréhensibilité de la parole d'apprenants de français à ce stade de notre étude, nous avons exploré la possibilité de prédire la compétence en matière de production orale. Notre approche consiste à représenter chaque apprenant par un ensemble de paramètres linguistiques multi-niveaux extraits grâce à des mesures automatiques.

Il existe à ce jour différents outils permettant de mesurer objectivement les productions *écrites* des apprenants de langue étrangère (Lu, 2010; Wetzel *et al.*, 2020). Ces mesures peuvent servir d'indicateurs pour déterminer les capacités et les niveaux d'habileté des apprenants. Cependant, les difficultés d'évaluation de l'*oral* diffèrent de celles de l'*écrit*, car la première nécessite notamment la prise en compte du plan phonético-phonologique.

Les méthodes d'évaluation automatique développées pour le niveau phonético-phonologique ont pour but principal d'évaluer la qualité de la prononciation des apprenants, tandis que celles développées pour les niveaux lexical, syntaxique et discursif ont pour but d'évaluer la compétence linguistique des apprenants. La plupart des études s'intéressant à l'évaluation automatique des compétences orales en L2 se focalisent sur un seul niveau linguistique : la prononciation (Laborde *et al.*, 2016), la fluence (Cucchiarini *et al.*, 2000) ou encore le lexique (Lindqvist *et al.*, 2011). Il est intéressant de noter que, de par sa complexité, l'évaluation automatique de la parole non-native au regard des niveaux supra-phonétiques, notamment en français, est très peu étudiée.

Dans ce chapitre, nous décrivons l'implémentation des paramètres qui permettent d'évaluer le niveau d'un apprenant de français en prononciation, fluence phonétique, lexique, syntaxe et discours. Nous étudions ensuite leur adéquation avec le niveau CECRL des apprenants non natifs de français et la possibilité de les utiliser pour prédire ce niveau. Nous appliquons notre étude sur les données du corpus CLIJAF, collectées dans le cadre méthodologique général du projet IPFC (Detey et Kawaguchi, 2008; Racine *et al.*, 2012). Plus précisément, nous utilisons deux sous-ensembles de ce corpus, le sous-ensemble CLIJAF_18 (18 apprenants) pour l'étude sur l'adéquation entre les mesures et le niveau CECRL, et le sous-ensemble CLIJAF_38 (38 apprenants) pour la prédiction de ce niveau (voir annexe A.1).

2.1 Paramètres linguistiques multi-niveaux

Dans cette section nous présentons les différentes mesures que nous avons implémentées afin d'extraire des paramètres et d'évaluer de manière automatique les compétences linguistiques d'apprenants de français langue étrangère.

2.1.1 Prononciation au niveau segmental

L'évaluation automatique de la prononciation s'effectue généralement de deux manières (Detey *et al.*, 2016). La première est basée sur le calcul de paramètres acoustiques (phonémiques et prosodiques) (Witt, 2012), la deuxième sur le calcul des scores de reconnaissance des phonèmes obtenus à l'issue de l'utilisation d'un système de reconnaissance automatique de la parole (Eskenazi, 2009; Chen et Jang, 2012; Hu *et al.*, 2013), basé sur des techniques d'alignement libre ou forcé. Nous avons choisi de l'évaluer en utilisant la deuxième méthode, par le biais d'un système de reconnaissance automatique de la parole avec alignement libre (Heba, 2021).

Nous utilisons un système de transcription dont les modèles acoustiques ont été entraînés sur 340 heures de données issues de différents corpus audio français tels que ESTER1, ESTER2 (Galliano *et al.*, 2005), EPAC (Estève *et al.*, 2010), BREF (Lamel *et al.*, 1991) et Librivox français (Kearns, 2014). À ces données s'en ajoutent d'autres issues de techniques d'augmentation courantes, comme la perturbation du rythme, l'addition de bruit de type « *cocktail-party* » et le masquage de blocs de canaux de fréquence à l'aide de l'outil SpecAugment (Park *et al.*, 2019). Ce système peut être utilisé pour reconnaître des phonèmes ou mots français et donne en sortie un texte correspondant aux unités reconnues ainsi qu'un **indice de confiance** par unité sous la forme d'une probabilité. Nous avons utilisé ce système *via* la plateforme PATY [5] (*plateforme de traitement de Parole Atypique*), développée en partenariat par l'IRIT et le LPL (Laboratoire Parole et Langage).

2.1.2 Fluence phonétique

Concernant la fluence, deux approches sont également à considérer. Dans le contexte de l'apprentissage des langues, la fluence est définie comme étant « *le degré de fluidité du discours sans pauses ni autres marques de disfluence* » (Derwing et Munro, 2015, p. 5, notre traduction). La première approche consiste à utiliser la reconnaissance de la parole comme outil pour mesurer des paramètres temporels tels que le débit de parole ou la longueur moyenne des pauses (Cucchiarini *et al.*, 2000). Étant donné que l'utilisation de tels systèmes connaît des limites, car dépendants de la langue pour laquelle leurs modèles ont été entraînés, nous nous basons sur une méthode développée pour mesurer la fluence de manière plus automatique et indépendamment de la langue cible. Nous pouvons citer par exemple l'algorithme présenté par Fontan *et al.* (2020), issu de travaux pilotes portant sur l'évaluation automatique de la fluence phonétique d'apprenants japonais de français en tâche de lecture (Fontan *et al.*, 2018; Detey *et al.*, 2020). Ces travaux s'appuient sur la méthode de segmentation *Forward-Backward Divergence* (André-Obrecht, 1988) basée sur la détection de ruptures dans la trajectoire de l'énergie du signal de parole au cours du temps et permettent, en outre, de calculer des variables à partir de pseudo-syllabes (Farinas et Pellegrino, 2001) et de pauses silencieuses, comme le débit de parole ou encore le pourcentage de parole.

5. `https://paty.irit.fr/demo`

Nous avons utilisé cet outil pour évaluer la fluence phonétique. À partir du résultat de la segmentation du signal audio correspondant à l'énoncé à traiter, nous calculons l'énergie de chaque segment. L'algorithme identifie des pseudo-syllabes et des pauses silencieuses (figure 2.1). Sur la base de ces éléments, nous pouvons mesurer quatre paramètres :

— le **débit de parole** :

$$\textit{Débit de parole} = \frac{N_{ps}}{D} \tag{2.1}$$

avec N_{ps} le nombre de pseudo-syllabes et D la durée totale du fichier audio.

— le **pourcentage de parole** :

$$\textit{Pourcentage de parole} = \frac{D_{sp}}{D} \tag{2.2}$$

avec D_{sp} la durée des segments de parole.

— l'**écart-type de la durée des pseudo-syllabes** ;

— le **nombre normalisé de pauses silencieuses**

$$\textit{Nombre normalisé de pauses silencieuses} = \frac{P_s}{D} \tag{2.3}$$

avec P_s le nombre de pauses silencieuses.

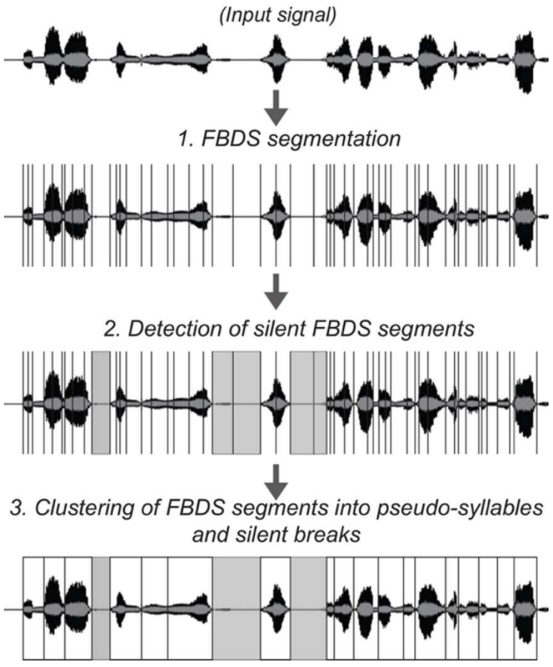


FIGURE 2.1 – Étapes de segmentation automatique du signal de parole en pseudo-syllabes (rectangles transparents) et pauses silencieuses (rectangles gris ; Fontan *et al.*, 2022).

Ces paramètres sont complétés par un score dit de « fluence normalisée » (Fontan *et al.*, 2020). Ce score est le résultat d'une régression linéaire multiple basée sur les

quatre paramètres précédents et ayant pour but de prédire la fluence phonétique d'étudiants anglophones apprenant le français, pour une tâche de lecture à voix haute. En utilisant la même équation de régression, nous pouvons mesurer des scores de fluence normalisés entre 0 (fluence la plus basse) et 1 (fluence la plus élevée).

2.1.3 Compétences lexicales

L'évaluation des compétences lexicales d'apprenants d'une langue étrangère renvoie souvent à l'évaluation de la *richesse lexicale* (Laufer et Nation, 1995; Bonvin et Lambelet, 2019). Celle-ci est composée de trois sous-dimensions :

— la **diversité lexicale**, qui caractérise la taille du vocabulaire dans un texte ou un énoncé, et définit le nombre de mots différents produits par un locuteur. Elle se mesure de différentes manières, les plus connues étant l'index de Guiraud (Guiraud, 1959) et le *Type-Token Ratio*, ou TTR (Daller *et al.*, 2003). L'index de Guiraud étant reconnu comme plus stable que le TTR pour des énoncés de longueurs variables (van Hout et Vermeer, 2007), nous l'avons implémenté pour mesurer la diversité lexicale en utilisant la formule suivante :

$$\textit{Diversité lexicale} = \frac{V}{\sqrt{N}} \tag{2.4}$$

avec V le nombre de mots distincts lemmatisés et N le nombre total de mots lemmatisés.

— la **densité lexicale**, qui caractérise la proportion de mots lexicaux dans un texte ou un énoncé. Un mot lexical est ici défini comme un mot dont la catégorie est un verbe, un nom, un adjectif ou un adverbe. La proportion de mots lexicaux se calcule en utilisant la formule suivante :

$$\textit{Densité lexicale} = \frac{V_{ml}}{N} \times 100 \tag{2.5}$$

avec V_{ml} le nombre de mots lexicaux lemmatisés distincts.

— la **sophistication lexicale**, qui caractérise le nombre de mots lexicaux relevant d'une connaissance ou d'une pratique plus avancée de la langue. Elle est définie par l'utilisation de mots rares ou peu fréquents dans la langue, qui diffère selon le niveau du locuteur non-natif (Ovtcharov *et al.*, 2006). La notion de fréquence des mots dans la langue cible est donc une information indispensable pour cette mesure. Les bases de données lexicales conçues pour des langues spécifiques intègrent généralement des informations en termes de lexique et de fréquence d'utilisation. Concernant la langue française, nous avons à disposition la BDD (Base De Données) Lexique3 (New *et al.*, 2005) qui contient plus de 135 000 entrées lexicales. Pour chaque entrée, nous retrouvons des informations telles que les différentes formes lexicales, les lemmes, et surtout les fréquences d'occurrence. Cette BDD est aussi bien adaptée pour l'évaluation de l'oral que de

l'écrit, car elle a été établie à partir de corpus de livres et de sous-titres de films. Nous avons donc implémenté la mesure de sophistication lexicale en nous basant sur les fréquences des lemmes compris dans Lexique3, et obtenus à partir des corpus de sous-titres de films (colonne *freqlemfilms2* de la BDD). La proportion de mots lexicaux dont les fréquences des lemmes sont inférieures à 10 pour un million, donc considérés comme rares [6] dans cette BDD, est mesurée classiquement comme suit :

$$\textit{Sophistication lexicale} = \frac{V_{10}}{N} \times 100 \tag{2.6}$$

avec V_{10} le nombre de mots lexicaux ayant une fréquence d'occurrence de lemme inférieure à 10.

2.1.4 Compétences syntaxiques

L'évaluation des compétences syntaxiques renvoie à l'analyse de la *complexité syntaxique* d'un texte ou d'un énoncé. L'unité la plus couramment employée pour mesurer objectivement la complexité syntaxique est la phrase. Il existe plusieurs manières de la mesurer, telles qu'en calculant le nombre moyen de mots par phrase, ou en calculant la profondeur moyenne des arbres syntaxiques par exemple (Blache, 2010; Lahuerta Martínez, 2018). La **longueur moyenne des énoncés**, en termes de mots, se mesure avec la formule suivante :

$$\textit{Longueur moyenne} = \frac{1}{N_e} \sum_{i=0}^{N_e} \text{len}(E_i) \tag{2.7}$$

avec N_e le nombre total d'énoncés produits, et $\text{len}(E_i)$ la longueur de l'énoncé i, en termes de mots.

Nous mesurons également la **profondeur moyenne des arbres syntaxiques** avec la formule suivante :

$$\textit{Profondeur moyenne} = \frac{1}{N_e} \sum_{i=0}^{N_e} \text{prof}(A_i) \tag{2.8}$$

avec $\text{prof}(A_i)$ la profondeur moyenne de l'arbre syntaxique de l'énoncé i.

Ce paramètre est mesuré en sommant les profondeurs des feuilles, puis en divisant cette somme par le nombre total de feuilles de l'arbre en cours de traitement (voir figure 2.2).

2.1.5 Compétences discursives

L'évaluation des compétences discursives, dans le cadre de l'apprentissage des langues, renvoie souvent à l'évaluation de la *cohésion du discours*. La structuration et

6. https://groups.google.com/g/lexiqueorg/c/C2fJ6JLQPK8/m/ydKYm2E9BAAJ

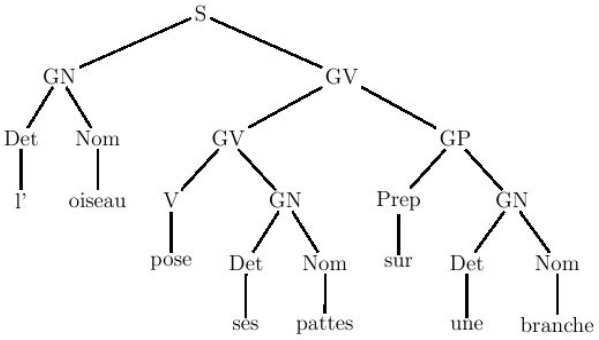


FIGURE 2.2 – Exemple d'arbre syntaxique pour l'énoncé « l'oiseau pose ses pattes sur une branche ». GN indique un groupe nominal, GV un groupe verbal et GP un groupe prépositionnel. La profondeur moyenne de cet arbre, telle que calculée avec notre formule, est 3,25. (source : `https://www.lattice.cnrs.fr/sites/itellier/poly_info_ling/linguistique006.html`)

l'articulation du discours sont souvent opérées par l'utilisation d'éléments de liaison, ou connecteurs (« ensuite », « de plus », par exemple), qui permettent de relier les propositions qui le composent (Beacco *et al.*, 2004). Il est alors évident de devoir disposer d'une liste de connecteurs du discours pour pouvoir en évaluer la cohésion. Pour la langue française, nous disposons de la liste LEXCONN (Roze *et al.*, 2012) contenant 431 connecteurs. Cette liste permet ainsi de mesurer la **proportion de connecteurs du discours** produits par un apprenant. Pour se rapprocher au plus près des critères utilisés dans le cadre du CECRL concernant les compétences discursives, la **diversité des connecteurs du discours** est également prise en compte en adaptant la formule de l'équation 2.4 :

$$\textit{Diversité connecteurs} = \frac{V_c}{\sqrt{N_c}} \tag{2.9}$$

avec V_c le nombre de connecteurs distincts et N_c le nombre total de connecteurs.

Cette formule nous permet de mesurer à quel point un énoncé contient des connecteurs du discours différents.

Nous prenons aussi en compte les disfluences pouvant survenir à ce niveau. Nous mesurons la **proportion de mots d'hésitation** (« *euh* » ou « *mh* » pour le français par exemple) et la **proportion de faux départs** (ou mots incomplets). Ces disfluences permettent de rendre compte des compétences en planification du discours.

2.2 Étude de l'adéquation des paramètres linguistiques multi-niveaux pour estimer le niveau CECRL d'apprenants

Afin d'étudier l'adéquation de nos paramètres linguistiques multi-niveaux avec les compétences en production orale d'apprenants de langue étrangère, nous menons une étude exploratoire sur le corpus CLIJAF (Detey et Kawaguchi, 2008; Racine *et al.*, 2012) contenant des enregistrements d'apprenants japonais de français. En effet, dans le cadre du LabCom ALAIA, la paire de langues L1/L2 étudiée est la paire japonais/français. Bien qu'un premier corpus soit utilisé au sein de ce LabCom, celui-ci ne contient que des productions orales de mots isolés obtenus lors d'une tâche de répétition. Afin de mener à bien notre étude, et mettre en pratique nos paramètres linguistiques, nous avons besoin de productions orales de contenu plus long. Le corpus CLIJAF se prêtait bien à notre étude, car son contenu ne se limite pas qu'à des mots isolés. De plus, il contient des métadonnées correspondant au niveau CECRL certifié des apprenants qui permettent d'attester leurs compétences.

Dans cette section nous présentons ainsi le corpus CLIJAF, et plus particulièrement le sous-ensemble de données que nous analysons. Nous étudions ensuite l'adéquation de nos paramètres linguistiques multi-niveaux avec le niveau CECRL des apprenants qui ont participé à l'élaboration de ce corpus.

2.2.1 Corpus CLIJAF

Pour cette analyse, nous nous intéressons à une partie du corpus CLIJAF. Ce sous-ensemble correspond à une tâche de production orale semi-spontanée d'apprenants japonais de français. Cette tâche consiste en un entretien individuel semi-directif en français, mené par un locuteur natif (enseignant ou doctorant). Les entretiens ont été conduits avec 43 apprenants, dont 27 suivent un cursus universitaire dans la préfecture de Tokyo (universités Waseda et TUFS (*Tokyo University of Foreign Studies*)) et 16 suivent un cursus universitaire dans la préfecture de Fukuoka (universités Seinan Gakuin et Fukuoka). Durant les entretiens, 12 questions sont posées aux apprenants et ceux-ci doivent y répondre en français. Nous retrouvons un mélange de questions simples et complexes, allant de « *Quel âge avez-vous et quelle est votre nationalité ?* » à « *Quelles sont les principales différences culturelles ou sociales entre la France et le Japon ?* ».

Chaque entretien a été enregistré dans les universités respectives des apprenants, soit en studio d'enregistrement, soit dans des salles de classe calmes. Les entretiens ont ensuite été manuellement transcrits de manière orthographique et segmentés en tours de parole. Les transcriptions conservent les informations temporelles et ont été réalisées par un des locuteurs natifs ayant participé aux entretiens. Des annotations

fines ont été apportées à ces transcriptions à l'aide du logiciel Transcriber [7] en termes de :

— correction lexico-grammaticale. Par exemple, le mot « *le* » de la production *« *je suis allé le tour Eiffel* » annoté comme « incorrect » et associé à l'annotation corrective « *à la* » [8],

— divergence phonético-phonologique. Par exemple, le mot « pâtisserie » annoté comme contenant une divergence consonantique, car réalisé avec une battue [ɾ] (patisəri) au lieu d'une uvulaire [ʁ] (patisəʁi), mais sans pour autant de transcription phonétique précise à ce stade de l'annotation.

Le corpus inclut les métadonnées des apprenants (profils linguistiques et contexte d'enregistrement de la parole par exemple) et offre des possibilités d'exploitation par le biais d'un concordancier dédié (Detey *et al.*, 2018). Pour indication, un concordancier est une IHM (Interface Homme-Machine) permettant d'interroger et de visualiser une base de données (le corpus CLIJAF dans notre cas). Ces recherches concernent principalement un mot et son contexte, et permettent d'attester de son utilisation ou de l'étudier.

2.2.2 Sous-corpus CLIJAF_18 et niveau CECRL associé

Notre objectif est de vérifier l'adéquation des paramètres multi-niveaux avec les compétences en production orale d'apprenants L2. Pour cette première étude, nous avons sélectionné six questions du corpus CLIJAF :

— *Quelles sont pour vous les plus grandes difficultés quand vous apprenez le français ?*

— *Quelle est la meilleure manière d'apprendre le français ?*

— *Où est-ce qu'on parle le meilleur français ?*

— *Quel est le français que vous souhaiteriez parler ?*

— *Est-ce que vous avez déjà eu des difficultés à comprendre des francophones ?*

— *Quelles sont pour vous les principales différences culturelles ou sociales entre la France et le Japon ?*

Ces questions conduisent à des réponses plus longues que les six autres questions du corpus, et font ressortir le plus de compétences linguistiques (par exemple, « *Quelles sont les principales différences culturelles ou sociales entre la France et le Japon ?* » amène à une réponse plus longue et met plus en avant les compétences linguistiques comparées à « *Quelles langues parlez-vous ?* »). Dans le corpus CLIJAF, 18 apprenants ont répondu à ces six questions amenant à des réponses élaborées. Nous nous sommes donc intéressés à ces 18 apprenants (3M/15F) ayant des niveaux CECRL compris entre A2 et C2, mais dont presque la moitié ont un niveau B2 (voir table 2.1). Le sous-ensemble considéré ici contient environ une heure et vingt minutes de données.

7. https://transcriber.fr.softonic.com
8. En didactique des langues, la notation « * » signifie que l'énoncé qui suit est erroné.

TABLE 2.1 – Niveau CECRL des apprenants japonais.

Niveau CECRL	A2	B1	B2	C1	C2
Effectif	1	4	8	3	2

Étant donné la nature de la tâche de production orale de ce corpus, à savoir un entretien semi-dirigé, chaque prise de parole, que ce soit par l'apprenant ou par l'examinateur, est considérée comme étant un tour de parole. Ainsi, nous disposons d'autant de fichiers audio qu'il y a eu de tours de parole pendant les échanges entre les apprenants et les examinateurs. Les fichiers audio considérés ici sont les premières réponses, spontanées, produites par les apprenants en adéquation avec les questions posées. Nous ne considérons donc pas les réponses données après d'éventuelles relances des examinateurs. Pour chaque apprenant, nous disposons de six enregistrements audio ainsi que leurs transcriptions, soit un enregistrement audio et une transcription par réponse amenée lors de l'entretien. Pour l'évaluation des compétences syntaxiques, nous considérons l'enregistrement audio (le tour de parole) comme étant l'unité d'analyse, et non la phrase, car la réalité de la syntaxe de l'oral est plus complexe à traiter que celle de l'écrit (Rossi-Gensane, 2010). Pour l'évaluation de la prononciation, de la fluence phonétique et des compétences lexicales et discursives, les tours de parole de chaque apprenant ont été groupés dans un même enregistrement audio afin de constituer une seule et même unité d'analyse et d'obtenir une mesure globale par apprenant.

2.2.3 Extraction des paramètres linguistiques et adéquation avec le niveau CECRL

Nous avons appliqué la méthodologie décrite en section 2.1, à savoir l'extraction de paramètres issus des différents niveaux linguistiques, en nous basant sur une unité d'analyse représentée par l'enregistrement audio. Notre objectif est de vérifier si ces paramètres sont cohérents avec le niveau CECRL des apprenants enregistrés.

Les différents paramètres sont rappelés dans la table 2.2.

Nous avons soumis chaque résultat obtenu par les différents paramètres phonético-phonologiques, lexicaux, syntaxiques et discursifs à un test non paramétrique de Kruskal-Wallis (Kruskal et Wallis, 1952). Ce test permet de rejeter l'hypothèse nulle selon laquelle un paramètre est le même, quel que soit le niveau CECRL.

Les résultats obtenus pour les tests de Kruskal-Wallis (table 2.3) de la diversité lexicale ($p < 0{,}001$), la profondeur moyenne des arbres syntaxiques ($p < 0{,}05$), la longueur moyenne des tours de parole ($p < 0{,}05$) et la proportion de connecteurs du discours ($p < 0{,}05$) montrent que le niveau CECRL a un effet significatif sur ces paramètres. Les figures 2.3 et 2.4 montrent leur évolution selon le niveau CECRL des apprenants.

De manière générale, ces paramètres semblent augmenter en fonction du niveau CECRL. En effet, la figure 2.3 montre qu'un apprenant de niveau CECRL donné a un discours moins diversifié en termes de lexique utilisé qu'un apprenant de niveau

TABLE 2.2 – Paramètres multi-niveaux.

Niveaux linguistiques	Paramètres
Prononciation (segmental)	Indice de confiance (reconnaissance des phonèmes)
Fluence phonétique (suprasegmental)	Débit de parole Pourcentage de parole Écart-type de la durée des pseudo-syllabes Nombre normalisé de pauses silencieuses Score de fluence normalisé
Lexique	Diversité lexicale Densité lexicale Sophistication lexicale
Syntaxe	Longueur moyenne des tours de parole Profondeur moyenne des arbres syntaxiques
Discours	Proportion de connecteurs du discours Diversité des connecteurs du discours Proportion d'hésitations Proportion de faux départs

TABLE 2.3 – Résultats des tests de Kruskal-Wallis sur les différents paramètres. Les *p-value* significatives sont représentées par * ($p < 0{,}05$) et ** ($p < 0{,}01$).

Paramètres	H
Diversité lexicale	$13,4^{**}$
Profondeur moyenne des arbres syntaxiques	$11,29^{*}$
Longueur moyenne des tours de parole	$11,13^{*}$
Proportion de connecteurs du discours	$10,53^{*}$
Débit de parole	$9,21$
Proportion d'hésitations	$9,21$
Pourcentage de parole	$7,82$
Nombre normalisé de pauses silencieuses	$6,41$
Reconnaissance des phonèmes	$6,23$
Proportion de faux départs	$5,11$
Écart-type de la durée des pseudo-syllabes	$3,82$
Score de fluence normalisé	$3,68$
Sophistication lexicale	$2,77$
Diversité des connecteurs du discours	$0,97$
Densité lexicale	$0,30$

CECRL plus élevé. Il en est de même pour les trois autres paramètres, la profondeur moyenne des arbres syntaxiques des tours de parole, la proportion des connecteurs du discours et la longueur moyenne des tours de parole. Nous remarquons cependant que les apprenants de niveau C2 ont des résultats moyens plus élevés que les apprenants de niveau C1 uniquement pour ce dernier paramètre. Ce phénomène pourrait s'expliquer par le fait que les certifications des niveaux CECRL sont plus fortement basées sur l'écrit que sur l'oral, et qu'un apprenant de niveau CECRL C2 n'aurait pas forcément un niveau C2 en expression orale. Ces premiers résultats sont tout de même cohérents avec l'hypothèse selon laquelle plus un apprenant a un niveau élevé, plus son discours est diversifié, long (en termes de mots) et structuré avec des connecteurs du discours.

Ces résultats très encourageants ouvrent la possibilité de proposer une mesure

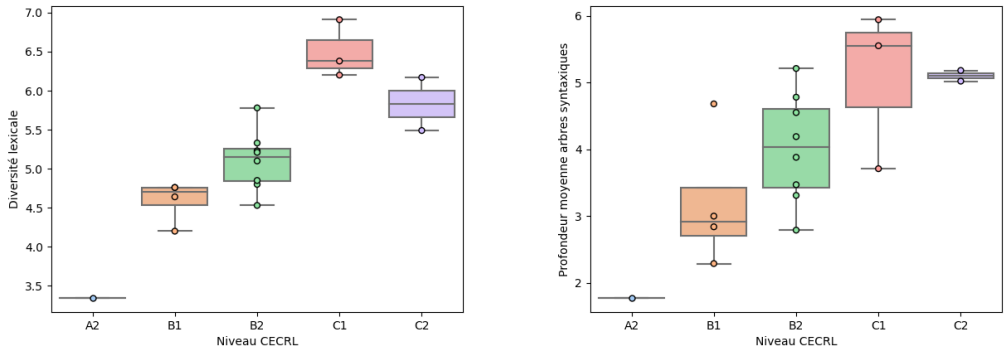


FIGURE 2.3 – Évolution des paramètres de diversité lexicale (à gauche) et de profondeur moyenne des arbres syntaxiques des tours de parole (à droite) selon le niveau CECRL des apprenants.

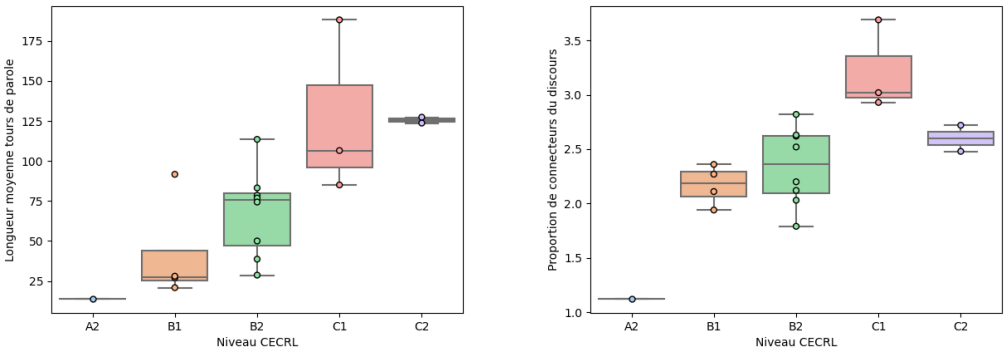


FIGURE 2.4 – Évolution des paramètres de longueur moyenne des tours de parole (à gauche) et de proportion de connecteurs du discours (à droite) selon le niveau CECRL des apprenants.

en lien avec le niveau de compétences des apprenants japonais, et basée sur des paramètres linguistiques multi-niveaux.

2.2.4 Regroupement des apprenants selon les niveaux CECRL

Après avoir montré qu'il existe un lien entre le niveau CECRL des apprenants et différents paramètres linguistiques multi-niveaux, nous voulons aller plus loin et voir si un regroupement des apprenants sur la base des paramètres qui les caractérisent est possible et cohérent avec les niveaux CECRL associés.

Nous avons utilisé l'algorithme d'apprentissage non-supervisé *k-means*, dit de « *clustering* » (Lloyd, 1982). Cet algorithme a pour objectif de diviser un ensemble de données en groupes, ou « *clusters* », de sorte que les éléments au sein d'un même groupe soient similaires, et ce sans connaissance *a priori* des étiquettes qui leur sont associées. Nous avons indiqué un nombre de *clusters* égal à cinq dans le but de regrouper les apprenants en cinq groupes, idéalement un groupe par niveau CECRL.

En prenant en compte tous nos paramètres, chaque apprenant serait représenté par un vecteur de 15 paramètres. Dans notre contexte d'apprentissage des langues, et plus particulièrement en vue de développer un outil pouvant servir à l'évaluation automatique des compétences en production orale, il est plus intéressant de pouvoir fournir un résultat basé sur un nombre minimum de paramètres pour avoir une meilleure interprétabilité et pouvoir fournir des retours cohérents aux apprenants ou enseignants. Comme nous disposons de cinq niveaux linguistiques, nous avons posé une contrainte stipulant qu'un apprenant devait être représenté par un jeu de cinq paramètres minimum. Ainsi, dans le meilleur des cas, au moins un paramètre par niveau linguistique constituerait le vecteur représentant un apprenant.

Nous avons réalisé une phase d'essai de toutes les combinaisons de paramètres possibles permettant de représenter un apprenant tout en respectant la contrainte énoncée précédemment (dimension minimum de 5). Afin de choisir la meilleure combinaison, nous évaluons le partitionnement associé avec une métrique d'évaluation, le *score de pureté*. Ce score permet d'évaluer la qualité des résultats de partitionnement, et nous informe sur l'importance de l'appartenance des éléments d'un même *cluster* à la même classe. Il se mesure avec la formule suivante :

$$\textit{Score de pureté} = \frac{1}{N}\sum_{i=1}^{k}\max|c_i \cap t_j| \tag{2.10}$$

avec N le nombre d'éléments (ici, le nombre d'apprenants), k le nombre de *clusters*, c_i le *cluster* i et t_j la classe la plus représentée dans le *cluster* c_i.

Le score de pureté est compris entre 0 (pureté très faible) et 1 (pureté parfaite).

Une fois que nous avons obtenu le meilleur score de pureté, nous calculons une deuxième métrique appelée *indice de Rand* associée au partitionnement obtenu avec la meilleure combinaison. En effet, le score de pureté connaît certaines limites, notamment le fait qu'il ne prend pas en compte la structure interne des *clusters* ou leur cohérence. Il peut donc être élevé malgré la présence de *clusters* redondants, c'est-à-dire des *clusters* contenant des éléments très similaires. L'indice de Rand permet d'effectuer une évaluation plus en détails de la qualité du partitionnement. Il permet d'évaluer à quel point les *clusters* obtenus correspondent aux regroupements réels, et peut être vu comme une mesure du pourcentage de décisions correctes prises par l'algorithme. Son calcul est donné par la formule :

$$\textit{Indice de Rand} = \frac{VP + VN}{VP + FP + FN + VN} \tag{2.11}$$

avec VP le nombre de vrais positifs, VN le nombre de vrais négatifs, FP le nombre de faux positifs et FN le nombre de faux négatifs. L'indice de Rand est compris entre 0 (correspondance aléatoire) et 1 (correspondance parfaite).

Le meilleur score de pureté a été obtenu avec une combinaison de six paramètres. Ainsi, chaque apprenant est représenté par un vecteur contenant les résultats des paramètres de diversité lexicale, longueur moyenne des tours de parole, proportion des connecteurs du discours, pourcentage de parole, débit de parole et scores de fluence

normalisés. Tous les niveaux linguistiques sont ainsi présents, à l'exception du niveau segmental (mesure de la prononciation). Celui-ci ne représenterait donc pas, pour nos données, une information utile afin de regrouper les apprenants. Le partitionnement en *clusters*, à l'issue de l'application de l'algorithme *k-means* et de la représentation en deux dimensions obtenue par le biais d'une analyse en composantes principales, est représenté sur la figure 2.5.

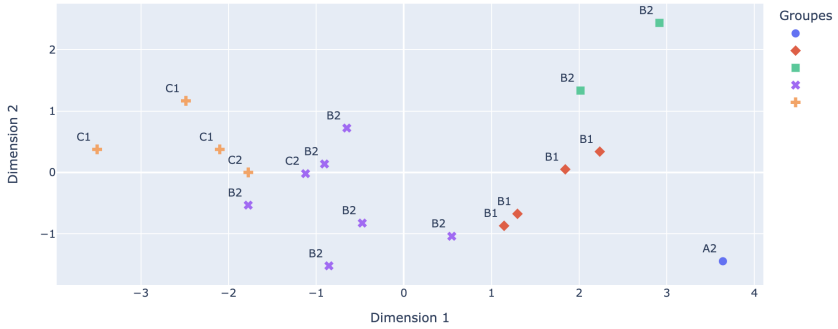


FIGURE 2.5 – Coordonnées des 18 apprenants selon les deux composantes principales résultant du partitionnement en *k-means* en utilisant six paramètres.

Nous obtenons un score de pureté de 0,88, indiquant une qualité de *clusters* élevée, et un indice de Rand de 0,85, indiquant que le regroupement obtenu est proche du niveau CECRL réel.

2.2.5 Bilan

Dans le cadre de cette étude portant sur l'adéquation des paramètres multi-niveaux avec le niveau CECRL des apprenants, nous obtenons des résultats très satisfaisants dans nos premières analyses concernant les mesures de diversité lexicale, longueur moyenne des tours de parole, proportion des connecteurs du discours et profondeur moyenne des arbres syntaxiques. Dans ce contexte, nous pouvons dire que l'évolution de ces paramètres joue un rôle significatif dans le lien qui peut être fait avec le niveau CECRL des apprenants. Il est à noter que les niveaux C1 et C2 ne semblent pas se comporter comme attendu. En moyenne, les apprenants du groupe C1 obtiennent de meilleurs résultats que les apprenants du groupe C2. Cette différence pourrait provenir du fait qu'il existe une certaine disparité entre l'évaluation du niveau global et celle d'habiletés plus spécifiques, telles que la production orale. Les niveaux CECRL sont établis d'après des certifications basées plus fortement sur l'écrit que sur l'oral. D'après les informations dont nous disposons, un étudiant de niveau certifié globalement C2 ne va pas systématiquement avoir un niveau C2 en production orale. Elle pourrait aussi provenir du fait que les niveaux CECRL indiqués par les apprenants ont été acquis *via* différents examens, passés plus ou moins longtemps avant les enregistrements. Certains niveaux ne refléteraient alors pas exactement le réel niveau d'un apprenant. De plus, la prise de parole à l'oral, et notamment dans le contexte d'apprentissage des langues, peut être freinée par des aspects sociologiques,

psycholinguistiques ou cognitifs (Gardner, 1985, 2010). Le manque de données quant à ces trois aspects et l'effectif assez faible de données ne permet cependant pas de conclure sur ce point, qui reste tout de même une hypothèse plausible.

Le regroupement en *clusters* des apprenants, sur la base des six paramètres de diversité lexicale, longueur moyenne des tours de parole, proportion des connecteurs du discours, pourcentage de parole, débit de parole et scores de fluence normalisés, et à l'aide de l'algorithme d'apprentissage non-supervisé *k-means*, s'est lui aussi montré très concluant. Nous obtenons un score de pureté de 0,88 et un indice de Rand de 0,85. Il est également important de souligner que les paramètres utilisés pour représenter les apprenants sont mesurés de manière automatique, et ne nécessitent pas forcément de recourir à une connaissance *a priori* de la langue cible. Ces travaux ont donné lieu à une publication dans la conférence francophone JEP (Journées d'Études sur la Parole) (De Fino *et al.*, 2022a).

Nos conclusions doivent cependant être nuancées, car le volume de données à disposition est réduit et ne nous permet pas de généraliser. Nous analysons donc dans la section suivante un plus grand sous-ensemble du corpus CLIJAF. Notre méthode pouvant s'avérer pertinente pour l'évaluation automatique des productions orales, nous enrichissons également le corpus grâce à l'évaluation du niveau en production orale des apprenants réalisée par des experts en FLE (Français Langue Étrangère) afin de pouvoir comparer nos mesures aux vérités terrain ainsi collectées.

2.3 Prédiction du niveau en production orale

Dans cette section nous mettons en application notre extraction automatique de paramètres afin de prédire le niveau CECRL d'apprenants non-natifs en tenant compte de leurs compétences dans leurs productions orales. Nous utilisons un sous-ensemble du corpus CLIJAF (CLIJAF_38) contenant plus d'apprenants que dans la section précédente, et nous faisons évaluer en amont leur niveau en production orale par des enseignants experts.

2.3.1 Sous-corpus CLIJAF_38

Pour cette seconde étude, nous avons sélectionné les questions du corpus CLIJAF ayant été répondues par le plus grand nombre d'apprenants. Nous disposons ainsi des réponses aux quatre questions suivantes :

— *Quel âge avez-vous et quelle est votre nationalité ?*

— *Quelles langues parlez-vous ?*

— *Quelles sont pour vous les plus grandes difficultés quand vous apprenez le français ?*

— *Quelles sont pour vous les principales différences culturelles ou sociales entre la France et le Japon ?*

Trente-huit apprenants japonais (8M/30F) ont répondu à ces questions (voir annexe A.1). Le sous-ensemble considéré ici contient environ 80 minutes de données. Comme pour l'étude précédente, les enregistrements audio considérés correspondent aux premières réponses produites par les apprenants en adéquation avec les questions posées. Nous disposons ainsi de quatre enregistrements audio par apprenant (un par question posée) ainsi que de leurs transcriptions. De même, le tour de parole représente notre unité d'analyse pour l'évaluation des compétences syntaxiques. Pour l'évaluation de la prononciation, de la fluence phonétique, des compétences lexicales et discursives, nous avons groupé ces tours de parole afin de constituer une seule et même unité d'analyse (total de parole produite) et d'obtenir une mesure globale par apprenant.

2.3.2 Évaluation humaine du niveau en production orale

Afin de pouvoir prédire de manière automatique le niveau en production orale des apprenants, nous devons au préalable disposer de vérités terrain. Les niveaux CECRL déjà présents dans le corpus étant basés sur des certifications portant principalement sur des tâches écrites, nous présentons le protocole que nous avons défini et appliqué afin de collecter des niveaux CECRL terrain, en termes de production orale seule, auprès d'experts.

Évaluateurs

Nous avons fait évaluer le niveau CECRL en production orale des apprenants par trois enseignants (2M/1F) de FLE. Ces trois enseignants sont des évaluateurs officiels des examens DELF (Diplôme d'Études en Langue Française) et DALF (Diplôme Approfondi de Langue Française), permettant d'évaluer les compétences écrites et orales d'apprenants du français, du niveau débutant au niveau expérimenté. Ils détiennent également une expérience d'enseignement de FLE au Japon.

Interface d'évaluation

Les quatre fichiers audio ont été concaténés afin d'en obtenir un seul par apprenant. Les enseignants ont écouté chaque fichier exactement une fois, dans un ordre aléatoire, en utilisant une interface graphique que j'ai développée dans le langage de programmation Python avec la librairie Prodigy[9] (ExplosionAI GmbH, Berlin, Allemagne), version 1.11.7. Les enseignants ont été invités à écouter chaque stimulus, puis à évaluer la compétence orale de l'apprenant en lui attribuant un niveau CECRL de production orale (voir figure 2.6).

9. `https://prodi.gy/`

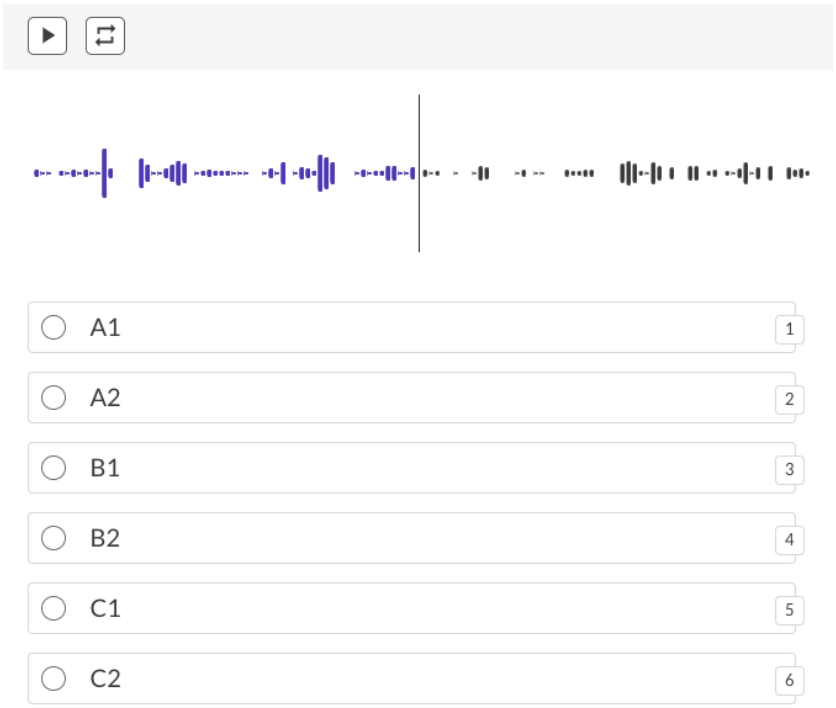


FIGURE 2.6 – Interface d'évaluation du niveau CECRL en production orale - exemple de l'évaluation d'un apprenant. En haut, la visualisation du signal de parole regroupant l'ensemble des productions (en bleu ce qui a déjà été écouté, en noir le reste du signal), et en dessous les niveaux possiblement attribuables. L'utilisateur coche le niveau qui correspond à son évaluation.

Résultats de l'évaluation

Chaque apprenant s'est vu attribuer trois niveaux CECRL en prononciation, soit un par enseignant (identifiés ci-après par les numéros 1, 2 et 3). Les niveaux ont été traduits en une échelle numérique discrète, allant de 1 (niveau A1) à 6 (niveau C2), comme indiqué en regard de chaque niveau dans la figure 2.6. Nous avons mesuré des corrélations de Spearman entre chaque paire d'enseignants pour évaluer l'accord inter-annotateurs, tout en vérifiant leur significativité (*p-value*, voir table 2.4).

TABLE 2.4 – Coefficients de corrélations de Spearman (ρ) entre chaque paire d'enseignants, accompagnés des *p-value* respectives.

Paire d'enseignants	ρ	***p-value***
1, 2	$0,78$	$< 0,001$
1, 3	$0,79$	$< 0,001$
2, 3	$0,70$	$< 0,001$

Nous remarquons un accord inter-annotateurs fort entre chaque paire d'enseignants ($\rho \geq 0,70$). Ces corrélations élevées nous permettent de faire la moyenne des trois scores obtenus par chaque apprenant. Ainsi, un apprenant n'est plus représenté par trois scores distincts, mais par un score moyen.

Globalement, les apprenants ont reçu un score moyen de 2,98 qui correspond à un niveau CECRL légèrement inférieur au niveau B1.

2.3.3 Prédiction du niveau des apprenants

Afin de prédire automatiquement les scores moyens attribués par les enseignants, chaque apprenant est représenté par un vecteur contenant les résultats du calcul de 14 paramètres, à savoir tous les paramètres décrits précédemment hormis le score de fluence normalisé (table 2.2). Nous avons écarté ce paramètre car il est fortement corrélé avec les quatre autres paramètres de fluence, étant donné qu'il représente une combinaison linéaire de ceux-ci. De plus, étant donné que nous souhaitons utiliser un modèle de régression linéaire pour la prédiction du niveau des apprenants, nous ne pouvons pas inclure ce paramètre, car une telle collinéarité affecte les performances et l'interprétabilité des modèles de régression linéaire (Wold *et al.*, 1984).

Comme pour l'étude présentée en section 2.2.4, nous voulons observer s'il est possible de prédire le niveau CECRL d'un apprenant sur la base d'un nombre réduit de paramètres extraits de manière automatique. Nous utilisons une régression linéaire de type LASSO (*Least Absolute Shrinkage and Selection Operator* ; Tibshirani, 1996) pour déterminer quels paramètres contribuent le plus à la prédiction du niveau en production orale.

De manière générale, en apprentissage supervisé, le jeu de données est découpé en jeux d'entraînement, de validation et de test. Le jeu d'entraînement permet d'entraîner un modèle, dans notre cas un modèle de prédiction, à prédire les étiquettes associées aux données. Le jeu de validation permet d'ajuster les hyper-paramètres du modèle. Le modèle est ainsi entraîné sur le jeu d'entraînement (phase d'apprentissage), et optimisé sur le jeu de validation (phase d'optimisation des hyper-paramètres). Une fois le processus d'entraînement (ou d'apprentissage) réalisé, et les hyper-paramètres optimaux définis, les performances du modèle sont évaluées sur le jeu de test. Les données constituant le jeu de test n'ont ainsi pas été prises en compte lors de l'apprentissage du modèle.

Dans notre cas, la taille réduite de notre jeu de données ne nous permet pas de le découper en jeu d'entraînement, de validation et de test. Nous utilisons ainsi une stratégie de validation croisée de type *leave-one-out* pour séparer tout d'abord nos données en jeu d'entraînement et de test. Cette stratégie consiste à entraîner notre modèle sur 37 apprenants (tous les apprenants sauf un) et de le tester sur le dernier, et ainsi de suite jusqu'à ce que tous les apprenants aient été présents exactement une fois dans le jeu de test.

Optimisation du modèle de régression linéaire LASSO

Les modèles de régression linéaire de type LASSO contiennent un hyper-paramètre α, dit de régularisation. Plus cet hyper-paramètre est grand, plus les coefficients de nos paramètres linguistiques sont égaux à 0, équivalant à une contribution nulle pour

la prédiction. Une phase d'optimisation du α est donc primordiale afin d'obtenir les meilleurs résultats possibles et d'observer quels paramètres contribuent le plus à la prédiction.

L'utilisation de la même stratégie de validation croisée et du même jeu de données pour l'estimation de l'hyper-paramètre optimal du modèle et de la meilleure prédiction conduit à une évaluation biaisée des performances du modèle. En d'autres mots, si nous utilisons la stratégie *leave-one-out* et le même jeu de données pour estimer l'hyper-paramètre α optimal et prédire les scores des apprenants en production orale, nous ajouterions un biais dans notre modèle de prédiction, le rendant optimal seulement pour notre jeu de données. Pour pallier ce problème, nous adoptons la stratégie de validation croisée dite « imbriquée », la *nested cross-validation*, que nous couplons avec la stratégie *leave-one-out* (voir figure 2.7). Il s'agit d'effectuer une validation croisée imbriquée dans la validation croisée initiale, créant ainsi une boucle interne et une boucle externe.

Le déroulement de la validation croisée imbriquée se présente comme suit :

1. validation croisée externe : division du corpus en K1 sous-ensembles,
2. pour chaque sous-ensemble d'entraînement de la boucle externe, division en K2 sous-ensembles pour la validation croisée interne,
3. validation croisée interne : pour chaque sous-ensemble d'entraînement K2, recherche du α optimal permettant d'optimiser les prédictions du sous-ensemble de validation K2,
4. sélection du meilleur hyper-paramètre α à l'issue de la validation croisée interne,
5. boucle externe : utilisation de cet α pour entraîner le modèle sur le sous-ensemble d'entraînement K1 et évaluation sur le sous-ensemble de test K1,
6. répétition de toutes les étapes autant de fois qu'il y a de sous-ensembles K1 dans la boucle externe,

Les validations croisées appliquées sur les boucles externes et internes correspondent à notre stratégie *leave-one-out* (un apprenant dans le jeu de test, tous les autres dans le jeu d'entraînement).

Nous avons testé des configurations avec un α variant de 0,001 à 1. Nous évaluons les performances du système en termes de MAE (*Mean Absolute Error*, erreur absolue moyenne) avec la formule suivante :

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{2.12}$$

avec N le nombre total d'exemples dans le jeu de données, y_i la prédiction et $\hat{y}_i$ la vérité terrain.

Ainsi, la valeur optimale de l'hyper-paramètre α est celle qui minimise cette métrique d'évaluation.

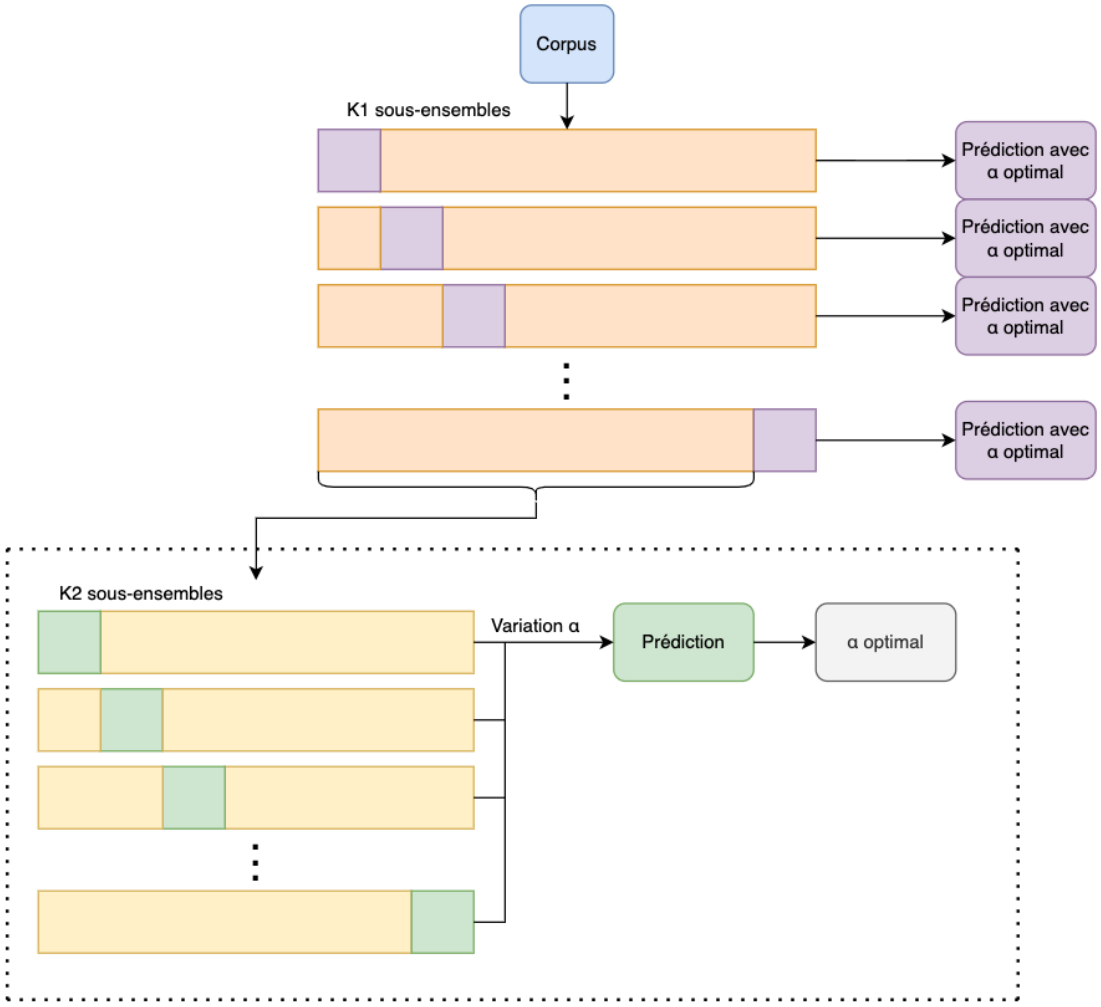


FIGURE 2.7 – Schéma explicatif de la validation croisée imbriquée. En orange, le jeu d'entraînement contenant 37 apprenants, en violet le jeu de test contenant un apprenant (niveau externe), en jaune le jeu d'entraînement contenant 36 apprenants et en vert le jeu de test contenant un apprenant (niveau interne).

Résultats de la prédiction du niveau attribué en production orale

À l'issue de la régression linéaire LASSO, cinq paramètres ont été identifiés comme apportant une contribution à la prédiction. Il s'agit du débit de parole, de la proportion de faux départs, de la proportion de mots d'hésitation, de la longueur moyenne des tours de parole en termes de mots et de la densité lexicale. Les coefficients normalisés attribués à ces cinq paramètres à l'issue de la régression linéaire LASSO sont présents sur la table 2.5 ci-dessous. Nous rappelons que plus un coefficient est proche de 0, moins il contribue à la prédiction.

Parmi ces paramètres, un permet d'évaluer la fluence phonétique (débit de parole), deux la planification du discours (proportion de faux départs et de mots d'hésitation), un la syntaxe (longueur moyenne des tours de parole) et un le lexique (densité lexicale). Nous remarquons que le débit de parole obtient le coefficient le plus élevé (en valeur

TABLE 2.5 – Coefficients normalisés attribués aux cinq paramètres qui contribuent à la prédiction du niveau en production orale des apprenants japonais.

Paramètres	**Coefficients normalisés**
Débit de parole	$0,160$
Proportion de faux départs	$-0,108$
Proportion de mots d'hésitation	$0,037$
Longueur moyenne tours de parole	$0,019$
Densité lexicale	$-0,007$

absolue), ce qui indique que ce paramètre contribue le plus à la prédiction des scores moyens en production orale des apprenants japonais.

Le résultat de la régression linéaire LASSO appliquée à nos données pour prédire les scores moyens en production orale des apprenants est visible sur la figure 2.8. Chaque point du nuage présent sur cette figure représente un apprenant autour de la droite de régression mesurée entre les scores moyens prédits et les scores moyens dits de terrain.

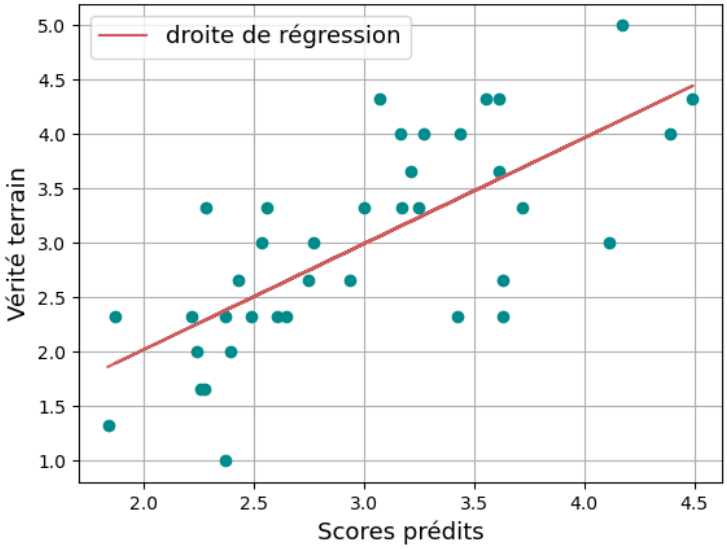


FIGURE 2.8 – Nuage de points représentant les scores moyens en production orale prédits par la régression LASSO par rapport à la vérité terrain.

Nous obtenons un coefficient de corrélation de Pearson de 0,71 entre les scores prédits et la vérité terrain, ainsi qu'une MAE de 0,53. Afin de vérifier si nos résultats se rapprochent des évaluations des enseignants, nous avons calculé des corrélations de Pearson entre nos prédictions et les scores attribués par chacun des enseignants pris séparément (voir table 2.6).

Nous remarquons dans cette table que nous obtenons une forte corrélation entre nos prédictions et les scores moyens attribués par l'enseignant 3 (r = 0,73). Nous

TABLE 2.6 – Coefficients de corrélations de Pearson (r) entre les scores prédits et les scores par enseignant, avec les *p-value* associées.

Enseignants	***r***	***p-value***
1	$0,60$	$< 0,001$
2	$0,59$	$< 0,001$
3	$0,73$	$< 0,001$

pouvons donc en déduire que notre modèle de prédiction a un comportement proche de celui-ci, et pourrait être considéré comme un quatrième évaluateur.

Comme indiqué dans la table 2.5, le débit de parole obtient le coefficient le plus élevé (en valeur absolue) à l'issue de la régression linéaire. La figure 2.9 illustre l'évolution de ce paramètre en fonction des scores moyens attribués par les trois enseignants. Il est intéressant de noter que plus le débit de parole est élevé, plus les scores moyens en production orale attribués par les enseignants lors de l'évaluation sont élevés.

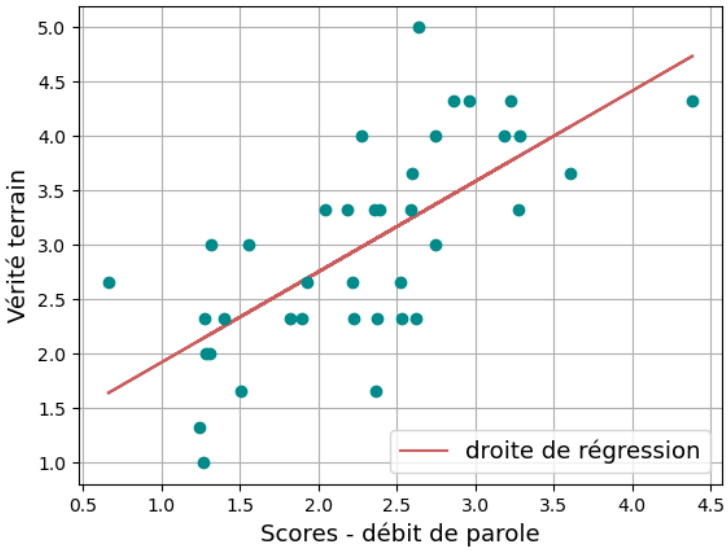


FIGURE 2.9 – Nuage de points représentant les valeurs du paramètre de débit de parole par rapport à la vérité terrain de production orale.

2.3.4 Bilan

Les résultats obtenus en termes de prédiction des scores moyens en production orale d'apprenants japonais sont très satisfaisants. En effet, cette étude démontre que des scores attribués par des enseignants de FLE peuvent être prédits en utilisant des paramètres automatiques multi-niveaux basés sur la fluence phonétique (débit de parole), le lexique (densité lexicale), la planification du discours (proportion de mots d'hésitation et de faux départs) et la syntaxe (longueur moyenne des tours de parole). La sélection de ces cinq paramètres a été réalisée grâce à l'application d'une méthode

de régression linéaire LASSO, dédiée à la prédiction du niveau CECRL des apprenants en production orale, en utilisant une stratégie *leave-one-out* couplée d'une validation croisée imbriquée. L'utilisation de seulement cinq paramètres pour prédire le niveau en production orale d'apprenants japonais de français s'est montrée très concluante à l'échelle de cette étude, avec une MAE de 0,53 et une corrélation de Pearson de 0,71. De plus, la forte corrélation entre nos résultats et les scores attribués par l'enseignant 3 ($r = 0{,}73$) montrent que notre modèle pourrait être considéré comme un évaluateur. Ces travaux ont également donné lieu à une publication dans la conférence internationale Interspeech 2022 (De Fino *et al.*, 2022b).

2.4 Conclusion

Dans ce chapitre, nous avons implémenté des mesures permettant d'extraire de manière automatique des paramètres aux niveaux segmental (prononciation), suprasegmental (fluence phonétique), lexical (richesse lexicale), syntaxique (complexité syntaxique) et discursif (cohésion et planification du discours). Nous avons ensuite démontré que l'évolution de ces paramètres est en adéquation avec les niveaux CECRL d'apprenants japonais de français, en exploitant le sous-corpus CLIJAF_18. De même, sur la base de six paramètres, les résultats du regroupement des apprenants selon leurs niveaux CECRL à l'aide de l'algorithme d'apprentissage non-supervisé *k-means* se sont montrés très concluants, avec un score de pureté de 0,88 et un indice de Rand de 0,85.

Étant donné que les niveaux CECRL présents dans ce corpus sont des niveaux dits « globaux », c'est-à-dire évalués sur la base de compétences orales et écrites, et que nos mesures sont uniquement basées sur les compétences orales, nous avons enrichi le sous-corpus CLIJAF_38 en intégrant des évaluations de niveau en production orale des apprenants japonais par trois enseignants de FLE, évaluateurs officiels des examens DELF et DALF, par le biais d'une interface graphique développée avec la librairie Prodigy. À l'issue des évaluations dites « humaines », chaque apprenant s'est vu attribuer un score par enseignant en production orale. En utilisant nos paramètres multi-niveaux et l'algorithme de régression linéaire LASSO couplé à une stratégie *leave-one-out* et d'une validation croisée imbriquée, nous avons prédit les scores moyens en production orale des apprenants (la moyenne des trois scores). Nous avons obtenu un score de MAE de 0,53 et un coefficient de corrélation de 0,71 entre nos prédictions et les scores attribués par les enseignants. Les cinq paramètres ayant contribué à ces résultats sont le débit de parole, la proportion de faux départs, la proportion de mots d'hésitation, la longueur moyenne des tours de parole et la densité lexicale.

Notre méthode est basée sur l'extraction automatique de paramètres linguistiques multi-niveaux. Ces paramètres ne dépendent pas de la L1 et peuvent facilement être implémentés pour d'autres langues cibles. Notre méthode est donc d'autant plus intéressante qu'elle peut facilement être adaptée pour l'évaluation d'autres paires de

langues L1-L2. Nous pouvons également envisager l'utilisation de ces paramètres pour la prédiction de la compréhensibilité de la parole.

Nous avons utilisé le corpus CLIJAF dans ce chapitre pour prédire le niveau CECRL d'apprenants japonais de français. Bien que ce corpus contienne des productions orales semi-spontanées d'apprenants, il n'est pas adapté pour la mesure de la compréhensibilité de la parole. Notre définition posée dans le chapitre 1 dédié à l'état de l'art fait référence à la compréhensibilité du *sens du message oral*. Pour collecter une vérité terrain et pouvoir annoter des enregistrements en termes de compréhensibilité, il est alors nécessaire d'avoir connaissance du réel message censé être véhiculé par un apprenant. Le corpus CLIJAF ne contient malheureusement pas cette référence, nous ne sommes donc pas en mesure de savoir si les réponses données aux questions correspondent exactement au message que les apprenants voulaient faire passer. De plus, les productions orales semi-spontanées amènent trop de variabilité : pour une même question posée, les réponses peuvent être plus ou moins longues selon les apprenants. Cette disparité permet certes de différencier les niveaux CECRL des apprenants, mais l'évaluation de la compréhensibilité de la parole requiert, à ce stade de notre étude, une tâche de production orale plus cadrée.

Dans le chapitre suivant, nous proposons un protocole permettant de constituer un corpus contenant des productions orales d'apprenants plus cadrées. Les évaluations humaines, basées sur notre définition de la compréhensibilité, incluses dans ce corpus, nous permettront d'évaluer automatiquement la compréhensibilité de la parole des apprenants de langue étrangère.

3

Protocoles de collecte et d'annotation

Sommaire

Nous proposons dans ce chapitre une méthodologie permettant de constituer un corpus en collaboration avec les membres du projet de recherche de type « KAKEN(B) » financé par la *Japanese Society for the Promotion of Science* intitulé « *From corpus to target data as steps for automatic assessment of L2 speech : L2 French phonological lexicon of Japanese learners* » [10] afin de répondre à deux problématiques. La première concerne les travaux de cette thèse, à savoir la prédiction automatique de la compréhensibilité de la parole d'apprenants de langue étrangère, la deuxième concerne la désambiguïsation des erreurs morphophonologiques au niveau lexical, problématique du projet KAKEN.

Nous définissons tout d'abord un protocole de collecte d'enregistrements audio d'apprenants de L2. Ces enregistrements sont récoltés *via* une tâche de production orale cadrée, permettant de connaître le réel sens du message qu'un apprenant doit véhiculer. Nous définissons ensuite un protocole d'annotation subjective de la compréhensibilité de ces enregistrements par des évaluateurs natifs de la langue cible. Enfin, nous présentons des analyses réalisées sur les annotations ainsi récoltées.

Nous appliquons tout d'abord nos protocoles de collecte et d'annotation avec des apprenants japonais de français. Puis, dans l'optique de vérifier l'aspect généralisable de notre méthodologie, nous l'appliquons également sur des apprenants allemands de français.

3.1 Protocole de collecte des données

Dans cette section nous présentons le protocole défini afin de collecter des données d'apprenants L2 par le biais d'une tâche de production orale. Nous définissons plus précisément cette tâche et les moyens logiciels mis en place afin de parvenir à enregistrer des productions orales d'apprenants.

3.1.1 Co-construction de la tâche de collecte

Nous utilisons une tâche de traduction orale d'énoncés écrits en L1 pour collecter des productions orales d'apprenants de langue étrangère. L'utilisation d'une tâche de traduction d'énoncés cibles nous permet de connaître la vérité terrain, le sens qui doit être véhiculé par l'apprenant, c'est-à-dire le sens cible. La spontanéité de la parole issue d'une tâche de traduction pourrait être remise en question, elle nous semble néanmoins être la meilleure solution pour étudier et prédire de manière automatique la compréhensibilité d'un apprenant. En effet, si nous étudions la compréhensibilité à partir de productions orales issues d'une tâche de parole conversationnelle, nous ne pourrions étudier que la compréhensibilité perçue *a priori* par un auditeur, car l'information quant à l'intention réelle derrière le message d'un apprenant serait manquante.

10. Detey, S. (dir.) (2020-2024). JSPS : Grant-in-Aid for Scientific Research (B) 20H01291.

3.1.2 Création du matériel de traduction

Les énoncés à traduire contiennent du vocabulaire de niveaux CECRL A1-A2, et sont créés de sorte à ne pas contenir d'ambiguïté de traduction. Chaque énoncé contient une difficulté courante que rencontrent les apprenants lorsqu'ils s'expriment en L2. Ces difficultés peuvent intervenir aux niveaux lexical, syntaxique ou morphosyntaxique, et sont propres à la L1 des apprenants. Par exemple, une erreur courante pour les apprenants japonais de français au niveau lexical est de remplacer le mot « *salle* » par le mot « *chambre* », au niveau syntaxique de remplacer « *mercredi* » par « *à mercredi* » et au niveau morphosyntaxique de remplacer « *le* » par « *du* ». Il est donc possible qu'un apprenant japonais commette les erreurs de traduction suivantes :

— traduire « 先生は10番教室にいます。 » (« Le professeur est dans la salle dix ») par « *Le professeur est dans la chambre dix* »,

— traduire « 私は水曜日に約束があります。 » (« J'ai rendez-vous mercredi ») par « *J'ai rendez-vous à mercredi* » ;

— traduire « 私はコーヒーが好きです。 » (« J'aime le café ») par « *J'aime du café* ».

3.1.3 Interface d'enregistrement

L'interface graphique utilisée pour enregistrer les productions orales des apprenants correspond à un site internet développé dans le langage de programmation JavaScript avec la librairie React. Ce site, développé et hébergé par Archean Technologies, permet de créer deux types de comptes, un *compte enseignant* et un *compte apprenant*. Le compte enseignant donne la possibilité de créer un exercice de traduction en renseignant les énoncés, les difficultés ciblées et les erreurs attendues pour ces difficultés. Une fois l'exercice créé, il est assigné au groupe d'apprenants ciblé et devient visible dans le compte apprenant. Deux exercices ont ainsi été créés, un exercice d'entraînement et un exercice de traduction. Un exercice complet (d'entraînement ou de traduction) est composé de trois à quatre étapes, comme représenté sur la figure 3.1. Seule l'interface correspondant à la première étape est détaillée ci-après. Les autres étapes sont décrites en annexe, car non directement liées à l'étude menée dans le cadre de ce travail de thèse.

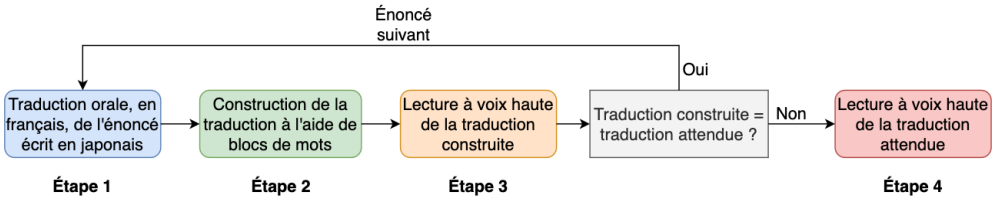


FIGURE 3.1 – Schéma explicatif des différentes étapes de l'exercice de traduction, exemple de la paire de langue japonais/français.

Durant la première étape, un énoncé écrit dans la langue maternelle de l'apprenant s'affiche à l'écran, et celui-ci doit le traduire à l'oral en français tout en s'enregistrant

(voir figure 3.2). Il est demandé à l'apprenant de s'enregistrer dès qu'une première traduction lui vient à l'esprit, afin de limiter l'aspect trop préparé de la parole produite. Dans le cas où l'apprenant ne connaît par la traduction d'un ou plusieurs mots, il lui est demandé de prononcer le pseudo-mot « *tatata* ». Un seul essai est autorisé, mais il est cependant possible pour l'examinateur de décider d'enregistrer une seconde fois la production de l'apprenant si un quelconque problème technique est survenu et impacte la qualité de l'enregistrement.

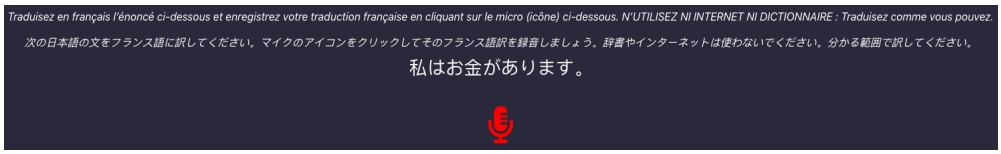


FIGURE 3.2 – Interface d'enregistrement de l'apprenant - étape 1 (exemple japonais/français) : traduction à l'oral d'un énoncé écrit en japonais. Sont présents sur cette figure de haut en bas : l'instruction en français, l'instruction en japonais, l'énoncé à traduire et le bouton d'enregistrement.

Durant la deuxième étape, l'apprenant doit construire une version écrite de la traduction attendue en s'aidant de blocs présents à l'écran (voir figures B.1 et B.2 en Annexe). Ces blocs se composent d'une difficulté ciblée (un ou plusieurs mots), d'une seule erreur typique par énoncé (un ou plusieurs mots) et des autres mots constituant l'énoncé (un bloc par mot).

La troisième étape consiste à enregistrer l'apprenant en train de lire la traduction construite à l'étape précédente. Dans le cas où celle-ci correspond mot pour mot à la traduction attendue, l'apprenant passe directement à l'énoncé suivant (voir figure B.3 en Annexe). Dans le cas contraire, une quatrième étape prend place et consiste à enregistrer l'apprenant en train de lire la traduction attendue (voir figure B.4 en Annexe).

Une session d'enregistrement d'un apprenant se déroule alors comme suit :

— **Instructions**. Une première phase de cinq à dix minutes est consacrée à la signature des documents RGPD [11] et à l'explication du déroulement de la session. Il est également demandé à l'apprenant d'indiquer son niveau CECRL en L2, soit obtenu après certification officielle, soit auto-évalué ;

— **Entraînement et test du microphone**. L'apprenant doit effectuer une session d'entraînement sur cinq énoncés non inclus dans l'exercice de traduction. Cet entraînement lui permet de se familiariser avec l'interface et permet à l'examinateur de vérifier que les instructions sont bien comprises. Durant cette phase, l'apprenant peut s'enregistrer autant de fois qu'il le veut, permettant ainsi à l'examinateur de vérifier le bon réglage du microphone et de le modifier le cas échéant ;

11. Règlement de l'Union Européenne qui constitue le texte de référence en matière de protection des données à caractère personnel.

— **Exercice de traduction**. Une fois l'entraînement terminé, les instructions claires et le microphone réglé, l'apprenant doit effectuer la vraie session d'enregistrement sur les N énoncés à traduire.

3.2 Protocole d'annotation

Maintenant que nous avons présenté notre protocole pour collecter des productions orales d'apprenants de langue étrangère, nous nous concentrons sur une méthodologie d'évaluation subjective de la compréhensibilité. Cette section décrit le protocole que nous avons défini pour annoter nos données en termes de compréhensibilité.

3.2.1 Sélection des évaluateurs

Comme expliqué au début de ce manuscrit (voir Introduction), nous avons vu que, dans le domaine de la didactique des langues, il n'est plus tant question d'amener les apprenants à atteindre un niveau de production orale équivalent à celui d'un natif, que d'atteindre un bon niveau de compréhensibilité. En tant qu'apprenant d'une langue étrangère, il est plus intéressant et primordial de pouvoir se faire comprendre par les natifs de la langue L2, dans un contexte de « vie de tous les jours », plutôt que par une population connaissant la langue maternelle ou habituée à l'accent L1 de l'apprenant. C'est dans cette optique que nous décidons de sélectionner les évaluateurs selon le critère de non-familiarité et de non-connaissance linguistique de la L1, ces derniers pouvant avoir un impact sur le degré de compréhensibilité (Trofimovich et Isaacs, 2012; Winke *et al.*, 2013). D'un point de vue perceptif, une personne familière avec l'accent d'un apprenant aurait des facilités à reconnaître des mots en dépit de leur prononciation atypique. D'un point de vue lexical et syntaxique, une personne connaissant la langue maternelle d'un apprenant serait plus sensibilisée aux potentielles erreurs d'énonciation, et aurait donc des facilités, le cas échéant, à comprendre l'intention communicative de l'apprenant. Pour finir, afin de limiter une potentielle source de variabilité liée aux performances auditives des auditeurs, nous avons défini un âge maximal de participation de 40 ans, l'étude de Cruickshanks *et al.* (1998) ayant montré qu'après 48 ans près d'une personne sur deux souffre de presbyacousie. Nous avons ainsi défini la liste de critères suivants :

— être natif L2,

— avoir entre 18 et 40 ans,

— ne pas avoir de problème de presbyacousie connu,

— ne jamais avoir étudié la L1,

— ne pas avoir de familiarité avec l'accent L1.

3.2.2 Annotation de la compréhensibilité de la parole

Les productions orales sont évaluées de manière individuelle en termes de compréhensibilité. Pour collecter des scores reflétant la compréhensibilité des apprenants vis-à-vis du réel sens du message censé être véhiculé, nous avons défini différentes étapes dans le processus d'annotation. Ces étapes sont les suivantes :

1. **Écoute de l'enregistrement**. Une seule écoute est autorisée afin de capturer la première impression et d'éviter un biais potentiel lié à des écoutes répétées du même enregistrement. Une deuxième écoute peut être possible, uniquement si l'examinateur considère qu'un problème matériel ou logiciel perturbe la première ;

2. **Transcription de l'enregistrement**. L'évaluateur transcrit avec des mots ce qu'il vient d'entendre. Tout mot jugé impossible ou trop difficile à décoder doit être remplacé par un point d'interrogation (« ? », exactement un point d'interrogation par mot). Les pseudos-mots *tatata* (voir section 3.1.3) doivent être transcrits tels quels ;

3. **Première évaluation de la compréhensibilité**. Un premier score de compréhensibilité est attribué à l'aide d'une échelle graduée à cinq points (1 = compréhensibilité nulle, 5 = compréhensibilité totale) et selon la définition de Woisard *et al.* (2013). Cette première évaluation représente la compréhensibilité *a priori*, sans connaissances sur le message initial et le sens à véhiculer ;

4. **Lecture de la traduction cible**. L'évaluateur prend connaissance de la traduction que devait produire l'apprenant ;

5. **Deuxième évaluation de la compréhensibilité**. Un deuxième score de compréhensibilité est attribué sur une deuxième échelle graduée à cinq points. Cette deuxième évaluation représente la compréhensibilité *a posteriori* et diffère de la première, car l'évaluateur a maintenant accès au vrai sens du message que l'apprenant devait véhiculer. Elle permet ainsi de réévaluer le message transmis initialement, qui pouvait sembler correct sans correspondre à ce qu'il devait transmettre.

Pour éviter l'apparition d'une certaine familiarité au cours du processus d'annotation, nous répartissons les enregistrements audio de sorte qu'un évaluateur n'ait pas à annoter deux fois la compréhensibilité d'un même apprenant et d'une même traduction (voir table 3.1). Afin de limiter l'aspect chronophage de la tâche d'annotation, nous ne faisons évaluer qu'un sous-ensemble de productions par évaluateur. Chaque production est évaluée par deux évaluateurs différents pour garantir une certaine fiabilité dans les annotations de la compréhensibilité de la parole, et une série de productions est évaluée par le même binôme d'évaluateurs.

TABLE 3.1 – Répartition des enregistrements audio par évaluateur. $a_N e_M$ correspond au fichier audio produit par l'apprenant N lors de la traduction de l'énoncé M (exemple avec 80 évaluateurs, 40 apprenants et 40 énoncés à traduire).

Évaluateurs	**Fichiers audio**
1	a_1e_1 a_2e_2 a_3e_3 ... $a_{38}e_{38}$ $a_{39}e_{39}$ $a_{40}e_{40}$
2	a_1e_{40} a_2e_1 a_3e_2 ... $a_{38}e_{37}$ $a_{39}e_{38}$ $a_{40}e_{39}$
3	a_1e_{39} a_2e_{40} a_3e_1 ... $a_{38}e_{36}$ $a_{39}e_{37}$ $a_{40}e_{38}$
...	...
41	a_1e_1 a_2e_2 a_3e_3 ... $a_{38}e_{38}$ $a_{39}e_{39}$ $a_{40}e_{40}$
42	a_1e_{40} a_2e_1 a_3e_2 ... $a_{38}e_{37}$ $a_{39}e_{38}$ $a_{40}e_{39}$
43	a_1e_{39} a_2e_{40} a_3e_1 ... $a_{38}e_{36}$ $a_{39}e_{37}$ $a_{40}e_{38}$
...	...
80	a_1e_2 a_2e_3 a_3e_4 ... $a_{38}e_{39}$ $a_{39}e_{40}$ $a_{40}e_1$

3.3 Création et analyse du corpus CAF-jp

Nos protocoles de collecte de données et d'annotation en termes de compréhensibilité de la parole étant défini, nous les appliquons afin de constituer un corpus composé de données d'apprenants japonais de français, le corpus CAF-jp.

3.3.1 Création des énoncés de traduction

Nous créons 40 énoncés à traduire pour les apprenants japonais de français (voir table C.1 en Annexe). Parmi ces énoncés, 18 contiennent des difficultés au niveau lexical, 22 contiennent des difficultés aux niveaux syntaxique ou morphosyntaxique et 5 contiennent des difficultés morphosyntaxiques de genre (réalisation de « *italien* » au lieu de « *italienne* » par exemple). Chaque énoncé a été relu et corrigé par un locuteur japonais.

3.3.2 Enregistrement des apprenants

Les différentes productions orales des apprenants japonais ont été recueillies à Tokyo (Japon), dans les universités de Waseda et TUFS, dans un laboratoire de phonétique ou dans une pièce calme. Nous avons pu enregistrer 42 participants. Les sessions d'enregistrement étaient individuelles, encadrées par un examinateur, et ont duré approximativement une heure par apprenant.

Nous avons utilisé l'interface graphique développée par Archean Technologies pour mener à bien les sessions, et un microphone casque Audio-Technica (Stanmore, Australie) modèle ATH-102USB. Étant donné le climat de pandémie mondiale qui régnait lors des sessions d'enregistrement, il a été demandé aux apprenants de porter un masque de type chirurgical pour leur propre sécurité. Le port de ce type de masque n'a pas eu d'impact négatif majeur sur la qualité des enregistrements, car il ne produit qu'une atténuation modeste dans la plupart des fréquences les plus importantes de la parole (Munro et Stone, 2021).

Étant donné l'aspect collaboratif de ce travail de collecte, nous n'avons utilisé pour notre étude sur la compréhensibilité de la parole que les productions orales issues de la première étape dite de *traduction spontanée*. En excluant les enregistrements produits lors des phases d'entraînement, et les enregistrements de deux apprenants trop bruités pour être intégrés au corpus, nous avons retenu les enregistrements audio produits par 40 apprenants japonais (13M/27F), soit respectivement 1600 productions (40 apprenants ayant traduit chacun 40 énoncés), représentant une heure et cinquante-et-une minutes d'enregistrement. Ces données constituent ainsi notre corpus CAF-jp. Enfin, le niveau CECRL de ces apprenants, indiqué en début de session d'enregistrement, se situe dans l'intervalle [A1,C1] (voir table 3.2).

TABLE 3.2 – Niveau CECRL connu selon les apprenants japonais.

Niveau CECRL	A1	A2	B1	B2	C1
Apprenants japonais	12	9	13	4	2

3.3.3 Interface et annotation du corpus CAF-jp

Nous avons recruté 80 francophones natifs (26F/54M), âgés entre 18 et 40 ans, ne présentant pas de problème de presbyacousie connu ni de familiarité avec le japonais pour annoter nos données. Les sessions d'annotation ont pris place dans des pièces calmes en utilisant un microphone casque Jabra (Copenhague, Danemark) modèle Evolve 20 HSC016 et ont duré approximativement une heure pour chaque participant. Chaque évaluateur a effectué une session individuelle, que j'ai moi-même supervisée. Durant une session, un évaluateur annote la compréhensibilité de 40 enregistrements, soit un enregistrement par apprenant et par traduction (voir table 3.1).

Pour mener à bien les évaluations, j'ai développé une interface graphique dans le langage de programmation Python avec la librairie Streamlit [12]. Nous retrouvons sur cette interface une page d'instructions, une page dédiée à l'entraînement et une page dédiée à l'évaluation de la compréhensibilité de la parole des apprenants japonais. Les paragraphes ci-dessous présentent les différentes étapes de l'annotation et les pages de l'interface, telles que présentées aux évaluateurs.

Instructions

Les dix premières minutes des sessions d'évaluation sont consacrées à l'explication des instructions (voir figure 3.3). La définition de la compréhensibilité de la parole, telle qu'adoptée au cours de mes travaux de thèse et tirée de Woisard *et al.* (2013), est proposée et complétée par un exemple afin de s'assurer que l'évaluateur ait compris que l'accent est mis sur le **sens du message** et non sur la justesse phonétique ou lexicale. L'exemple se compose des deux énoncés suivants :

— la voiture va vite,

12. `https://streamlit.io/`

— l'automobile roule rapidement.

Ici, bien que les énoncés soient construits avec un choix lexical différent, il est clair qu'ils véhiculent littéralement le même sens. Dans le premier énoncé, il est question de « voiture », qui désigne le même objet que le mot « automobile » dans le second. De même, les notions de « aller vite » et « rouler rapidement » dans ce cadre précis font référence à la même action. Le déroulement de la session est ensuite expliqué à l'évaluateur, conformément au protocole que nous avons défini. Les différentes étapes qui le composent sont alors répétées autant de fois qu'il y a d'enregistrements à évaluer, soit 40 fois concernant les apprenants japonais.

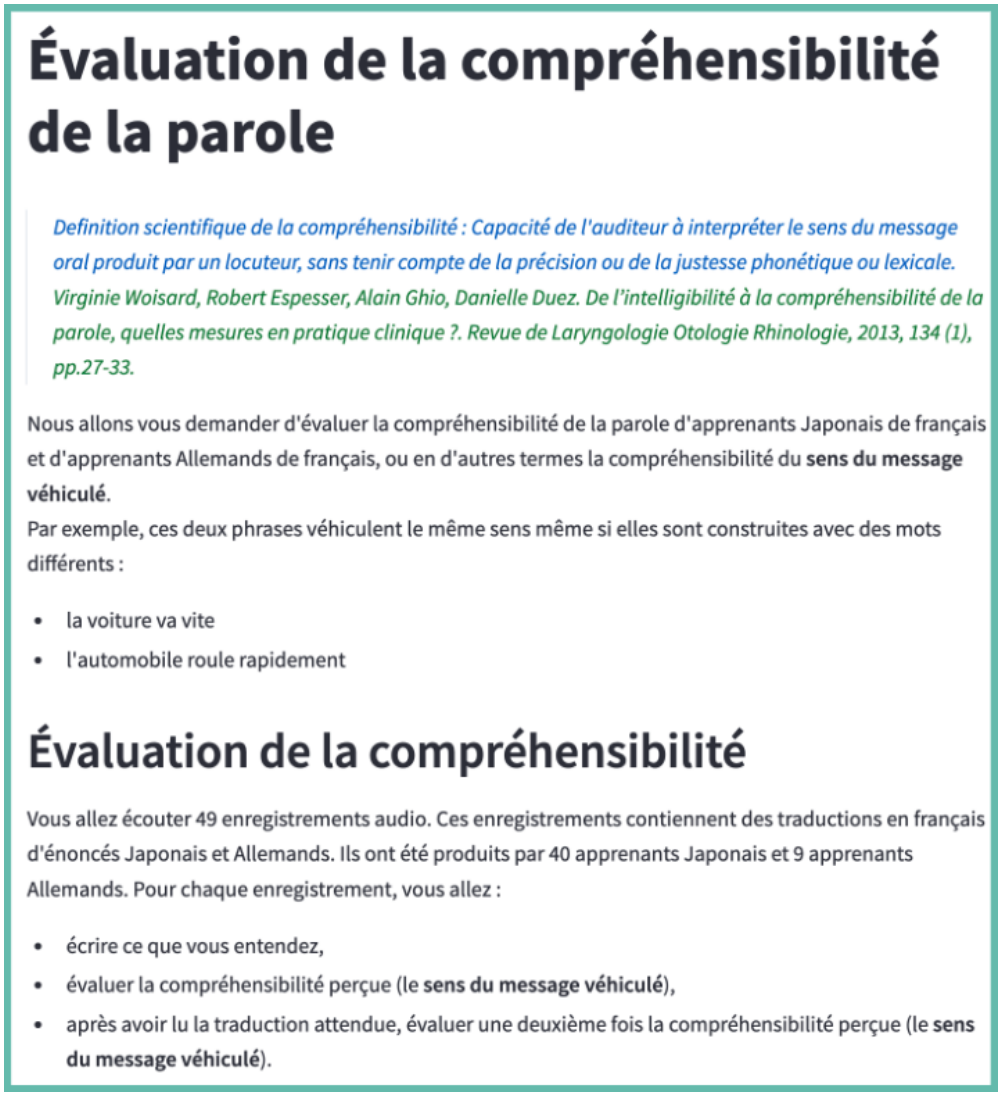


FIGURE 3.3 – Page d'instructions de l'interface d'évaluation de la compréhensibilité de la parole.

Entraînement

Une session spéciale dite d'entraînement est réalisée sur cinq enregistrements en amont de la phase proprement dite d'évaluation portant sur les données à annoter. Ces enregistrements correspondent aux productions récoltées pendant la phase d'entraînement des apprenants japonais (section 3.1.3). Bien que la perception de la compréhensibilité de la parole soit très subjective et qu'une production puisse être perçue différemment selon les auditeurs, nous avons tout de même essayé de les sélectionner de sorte qu'elles couvrent un maximum de possibilités auxquelles feraient face les évaluateurs. Se trouvent donc :

— deux productions parfaitement intelligibles et compréhensibles correspondant aux énoncés « j'ai 20 ans » et « il est étudiant » ;

— une production qui peut être perçue soit comme totalement incompréhensible et inintelligible, soit comme faiblement compréhensible. Cette production est censée correspondre à la traduction de l'énoncé « ils sont allés à Paris » ;

— une production parfaitement compréhensible lors de l'écoute, mais dont la compréhensibilité peut baisser lorsque la traduction attendue est révélée. Le message véhiculé est « je regarde la télé maintenant », entendu tel que « j'ai regardé la télé maintenant », mais la traduction attendue est « je regarde la télé chaque jour ». L'utilisation erronée d'adverbes de temps impacte le sens du message, de même que le fait d'entendre du passé avec « j'ai regardé » et de revenir au présent avec « maintenant » ;

— une production contenant le pseudo-mot « *tatata* », telle que « il mange *tatata* chaque jour », dont la traduction correspond à « ils mangent ensemble chaque jour ». Le fait de remplacer « ensemble » par « *tatata* » fait perdre, à l'oral, l'information du pluriel (en admettant que le son /s/ de « ils » ne soit pas prononcé). La première évaluation de la compréhensibilité se trouve donc être différente de la seconde évaluation, après lecture de la traduction attendue, mais ne devient pas nulle pour autant car l'action de manger et la récurrence donnée par « chaque jour » font partie du message effectivement transmis. De même, cet exemple permet de préparer et familiariser les évaluateurs avec ce pseudo-mot.

Évaluation de la compréhensibilité de la parole des apprenants

Conformément à notre protocole, l'évaluation commence par l'écoute d'un enregistrement, déclenchée par un clic sur le bouton « Écouter » (voir figure 3.4). Une fois celle-ci terminée, le bouton « Écouter » se grise pour ne pas permettre à la personne qui évalue de réécouter l'enregistrement. Une zone de saisie de texte s'affiche ensuite pour permettre de transcrire ce qui vient d'être entendu. Un score de compréhensibilité de la parole *a priori* est ensuite attribué en utilisant une échelle à cinq points graduée (voir figure 3.5) et dont les descripteurs sont détaillés dans la table 3.3.

Seuls les descripteurs associés à une compréhensibilité « Nulle » et « Totale » sont visibles en permanence et situés aux extrémités de l'échelle. Les autres deviennent

FIGURE 3.4 – Interface d'évaluation - écoute de l'enregistrement. L'évaluateur doit cliquer sur le bouton « Écouter » lorsqu'il est prêt. En bas de ce bouton se trouve le nom de l'enregistrement, utile dans le cas où une deuxième écoute serait nécessaire.

TABLE 3.3 – Descripteurs utilisés pour l'échelle de compréhensibilité.

Descripteurs	Nulle	Faible	Modérée	Élevée	Totale
Scores de compréhensibilité	1	2	3	4	5

visibles lorsque l'utilisateur déplace le curseur sur chacun des intervalles de l'échelle. L'évaluateur peut également bénéficier d'une aide située en haut à droite de l'échelle et représentée par un bouton contenant un point d'interrogation en son centre. Cette aide comporte le texte ci-dessous :

```
Échelle de compréhensibilité perçue
Compréhensibilité Nulle : même avec énormément d'effort, je ne comprends
rien du message
Compréhensibilité Faible : je comprends une faible partie du message
Compréhensibilité Modérée : je comprends la moitié du message
Compréhensibilité Élevée : je comprends la quasi-totalité du message
Compréhensibilité Totale : je comprends l'ensemble du message
```

Une fois le score de compréhensibilité *a priori* donné, la traduction attendue, que nous appelons *vérité terrain*, et l'échelle pour évaluer la compréhensibilité *a posteriori* s'affiche à l'écran. Dans l'exemple présent sur la figure 3.6, l'évaluateur a transcrit l'enregistrement audio par « j'ai le tablier », a évalué une première fois la compréhensibilité comme étant modérée puis, après avoir lu la traduction attendue qui était « j'ai apporté mon ordinateur », a évalué une seconde fois la compréhensibilité comme étant nulle. En effet, d'après lui, la traduction produite était au départ moyennement compréhensible, mais s'est avérée être totalement incompréhensible, car le sens du message véhiculé se trouvait être complètement différent du sens cible.

Le corpus CAF-jp est donc constitué de 6400 scores de compréhensibilité attribués par les 80 évaluateurs recrutés, dont 3200 correspondent à la compréhensibilité *a priori* et 3200 à la compréhensibilité *a posteriori*.

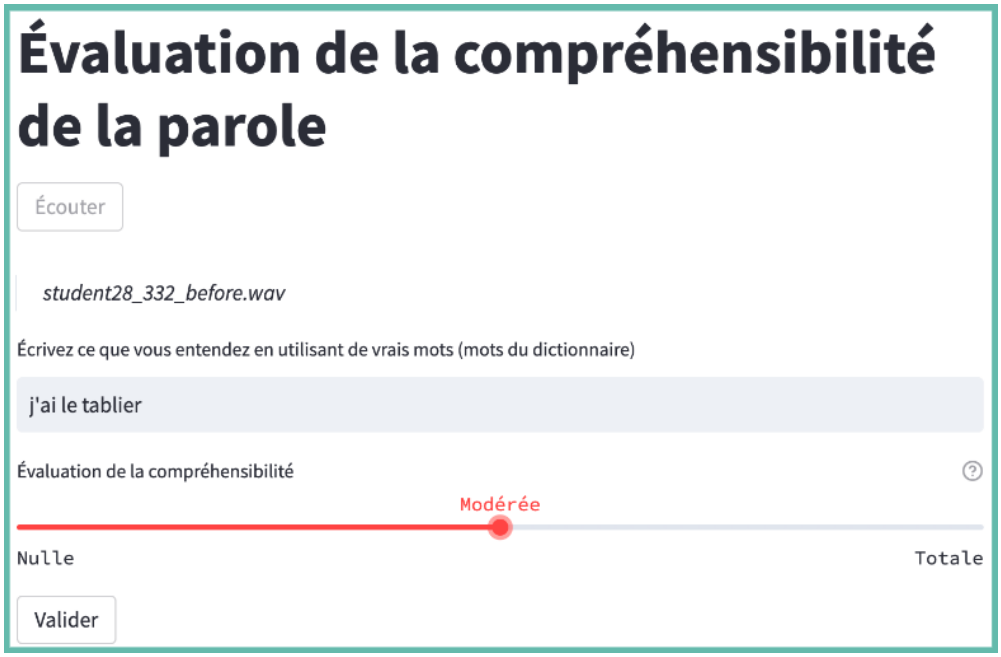


FIGURE 3.5 – Interface d'évaluation - transcription et première évaluation de la compréhensibilité de la parole.

3.3.4 Analyse des annotations

Cette section décrit les différentes analyses menées à l'issue de la collecte des évaluations de la compréhensibilité de la parole sur le corpus CAF-jp. Nous menons tout d'abord des analyses concernant les évaluateurs, afin d'étudier la cohérence de leurs scores étant donné la subjectivité de leurs appréciations. La deuxième analyse concerne les apprenants pour étudier les scores reçus et leurs liens avec les niveaux CECRL. Enfin, nous étudions les énoncés au travers des scores attribués. L'analyse des accords inter-annotateurs nous permet de vérifier la fiabilité des données, tandis que l'analyse de la distribution des scores de compréhensibilité permet de vérifier que la variabilité est suffisante pour effectuer des prédictions. De plus, dans cette section nous cherchons à vérifier si les évaluations de la compréhensibilité *a priori* et *a posteriori* sont significativement différentes et si les niveaux CECRL et les scores de compréhensibilité sont positivement corrélés.

Évaluateurs

Chaque production orale ayant été évaluée par exactement deux évaluateurs distincts, un accord inter-annotateurs a pu être mesuré. La table 3.4 présente les différents résultats observés pour les évaluations des productions des apprenants japonais concernent :

— les corrélations de Spearman (ρ) : calcul d'un coefficient de corrélation entre

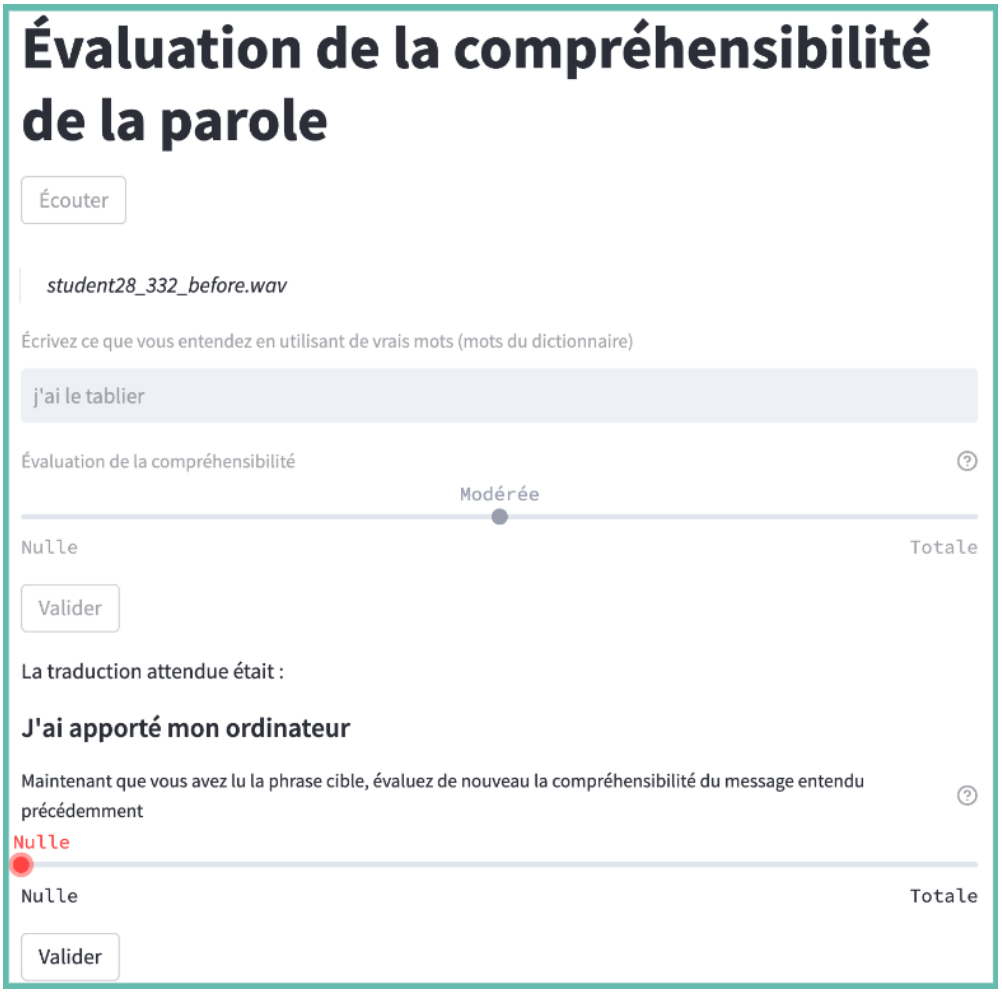


FIGURE 3.6 – Interface d'évaluation - deuxième évaluation de la compréhensibilité de la parole.

chaque paire d'évaluateurs avec la formule suivante :

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \tag{3.1}$$

avec n le nombre de points de données des variables et d_i la différence de rang de l'élément i,

— les Kappa de Cohen avec la formule suivante :

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{3.2}$$

avec p_o la proportion de l'accord entre les évaluateurs et p_e la probabilité d'un accord aléatoire ;

— les écarts inter-annotateur (écart moyen entre les scores donnés) : consiste à mesurer une erreur moyenne entre les scores de compréhensibilité de chaque paire d'évaluateurs en utilisant la formule de la MAE.

TABLE 3.4 – Résultats des accords et écarts inter-annotateurs moyens.

	Évaluation *a priori*	Évaluation *a posteriori*
ρ moyen	$0,72$	$0,75$
Écart-type	$0,1$	$0,1$
κ moyen	$0,42$	$0,46$
Écart-type	$0,1$	$0,1$
Écart inter-annotateurs moyen	$0,43$	$0,48$
Écart-type	$0,12$	$0,14$

Nous observons que les scores donnés par les différents évaluateurs sont en moyenne fortement corrélés, que ce soit pour la première ou la seconde évaluation. Le Kappa de Cohen moyen nous indique que les évaluateurs ont en moyenne eu un accord modéré concernant les scores de compréhensibilité aussi bien *a priori* qu'*a posteriori*. Tous les résultats moyens sont d'ailleurs légèrement plus élevés concernant les scores *a posteriori*. L'écart inter-annotateur moyen est lui aussi très intéressant : en moyenne, les scores attribués par les évaluateurs des mêmes productions orales varient de 0,43 pour l'évaluation *a priori* (respectivement 0,48 pour l'évaluation *a posteriori*), ce qui ne représente que 8,6% (respectivement 9,6%) de variation. De plus, les scores moyens *a priori* des productions des apprenants japonais sont de 4,25 et ceux *a posteriori* sont de 3,98 (voir table 3.5).

TABLE 3.5 – Scores donnés par les évaluateurs à l'issue de l'évaluation de la compréhensibilité de la parole des apprenants japonais (scores moyens, scores médians et écarts-types).

	Évaluation *a priori*	Évaluation *a posteriori*
Scores moyens	$4,25$	$3,98$
Scores médians	5	5
Écart-type	$1,22$	$1,39$

Les évaluateurs ont en moyenne donné des scores *a priori* plus élevés que les scores *a posteriori* lors de l'évaluation de la compréhensibilité de la parole. Nous avons testé l'hypothèse nulle stipulant que ces scores proviennent de la même distribution en utilisant le test non-paramétrique Wilcoxon (Wilcoxon, 1945). Nous obtenons comme résultat $p < 0,001$, indiquant que les scores *a priori* sont significativement supérieurs aux scores *a posteriori*. Les scores médians, quant à eux, sont identiques. D'après la figures 3.7 ci-dessous, nous pouvons observer que les évaluateurs ont majoritairement attribué le score de 5 en termes de compréhensibilité de la parole, quelle que soit l'évaluation.

La différence de scores moyens entre les conditions *a priori* et *a posteriori* pourrait s'expliquer par les différentes compréhensibilités perçues. En effet, le score de compréhensibilité lors de la première évaluation reflète une évaluation subjective, sans référence au contenu sémantique du message qui devait être transmis, le fait de comprendre ou non ce qui est produit. Le score de compréhensibilité à la suite de la seconde évaluation, quant à lui, reflète une évaluation en toute connaissance de cause, c'est-à-dire la justesse du sens du message véhiculé par rapport au message

initial à transmettre, ici la traduction cible. Un message peut ainsi être parfaitement compréhensible dans la condition *a priori*, mais moins compréhensible lorsque nous apprenons le réel sens à transmettre dans la condition *a posteriori*.

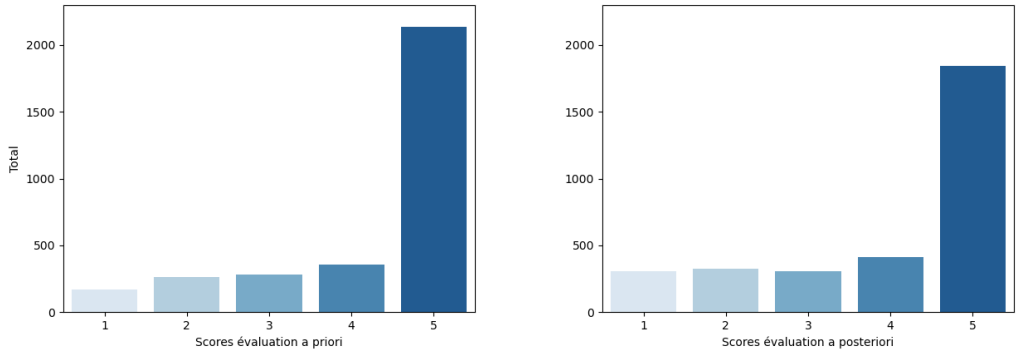


FIGURE 3.7 – Scores de compréhensibilité donnés aux productions des apprenants japonais selon les évaluations *a priori* et *a posteriori*.

Concernant les transcriptions, et pour vérifier que nos données ne sont pas aberrantes, nous avons mesuré le PER (Phone Error Rate) moyen entre chaque paire d'évaluateurs. Nous rappelons que nous leur avons donné pour consigne de transcrire exactement ce qu'ils entendaient avec des mots, et de représenter le ou les mot(s) impossible(s) à décoder par un point d'interrogation (exactement un point d'interrogation par mot). Pour faire abstraction des potentielles erreurs d'orthographe (par exemple, écriture de « Il est aller » à la place de « Il est allé »), nous avons au préalable converti les mots en suites phonémiques. Nous avons également retiré les points d'interrogation avant de mesurer le PER. Cette mesure est généralement utilisée pour évaluer les performances d'un système de reconnaissance de la parole ou d'un système de traduction automatique au niveau des unités phonétiques, nous l'avons néanmoins appliquée pour observer la similarité des transcriptions entre les paires d'évaluateurs. Celui-ci se mesure avec la formule suivante :

$$PER = \frac{S + O + I}{N} \tag{3.3}$$

où S correspond au nombre de substitutions, O correspond au nombre d'omissions, I au nombre d'insertions et N au nombre de phonèmes de référence.

Lorsque nous mesurons le PER entre deux transcriptions, nous identifions la première comme étant la référence et la seconde comme étant la transcription générée. Plus le PER est proche de 0, plus les deux transcriptions sont phonétiquement identiques. Le PER moyen mesuré entre chaque paire d'évaluateurs varie de 5,91 à 15,14 (voir figure 3.8). La paire d'évaluateurs [18,58] a produit les transcriptions les plus proches, tandis que la paire [32,72] a produit les transcriptions les plus différentes.

Nous remarquons sur ces figures, et notamment celle de droite, que nous ne possédons pas de valeurs aberrantes. Nous pouvons donc émettre l'hypothèse qu'aucune

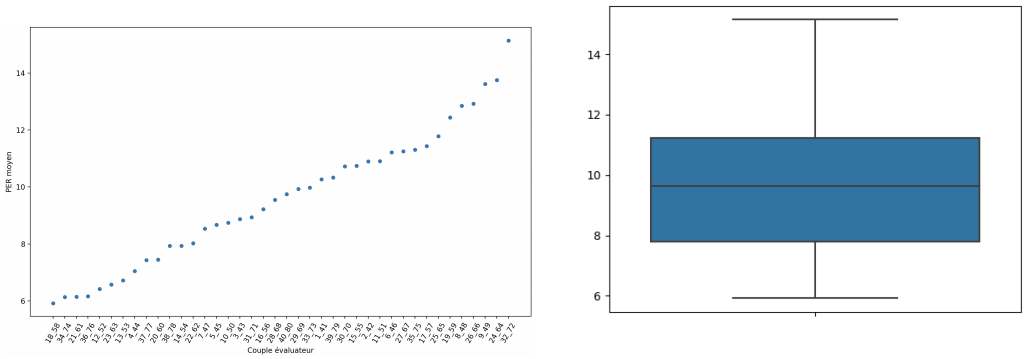


FIGURE 3.8 – PER moyen entre chaque paire d'évaluateurs pour les productions des apprenants japonais.

paire d'évaluateurs n'a eu à évaluer de production beaucoup plus compliquée ou facile en termes de compréhensibilité par rapport aux autres paires.

Apprenants

Nous avons étudié la distribution des scores obtenus globalement par chaque apprenant. Suivant la phase d'évaluation, nous pouvons observer sur la figure 3.9 que 27 apprenants (respectivement 33 et 38) ont reçu au moins une fois le score de 1 (respectivement 2 et 3) et tous au moins une fois les scores de 4 et 5 pour l'évaluation *a priori*. Concernant l'évaluation *a posteriori*, 39 apprenants (respectivement 32) ont reçu au moins une fois le score 2 (respectivement 1) et tous ont reçu au moins une fois les scores 3, 4 et 5. Nous remarquons également que plus d'apprenants ont reçu au moins une fois les scores de 1, 2 et 3 lors de la condition *a posteriori*, indiquant de nouveau une surestimation de la compréhensibilité lors de la condition *a priori*.

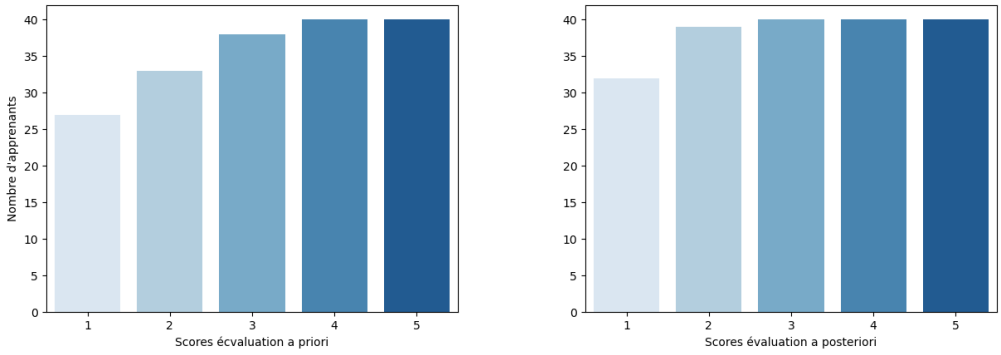


FIGURE 3.9 – Répartition des scores de compréhensibilité selon les apprenants japonais.

L'accord inter-annotateurs présenté dans la section précédente s'étant révélé élevé, nous pouvons nous permettre de calculer la moyenne des scores de compréhensibilité

obtenus. Pour rappel, chaque production d'un apprenant a été évaluée deux fois par deux évaluateurs différents. Étant donné que les apprenants ont chacun produit 40 enregistrements, nous disposons de 80 scores de compréhensibilité *a priori* et 80 scores de compréhensibilité *a posteriori* par apprenant. Pour chaque apprenant, et après avoir calculé la moyenne des scores de chaque production orale, nous passons de 80 scores à 40 scores moyens. Une analyse plus fine de ces scores nous permet d'observer leur adéquation avec le niveau CECRL renseigné lors de la phase d'enregistrement des productions orales. La figure 3.10 présente leur évolution en fonction du niveau CECRL.

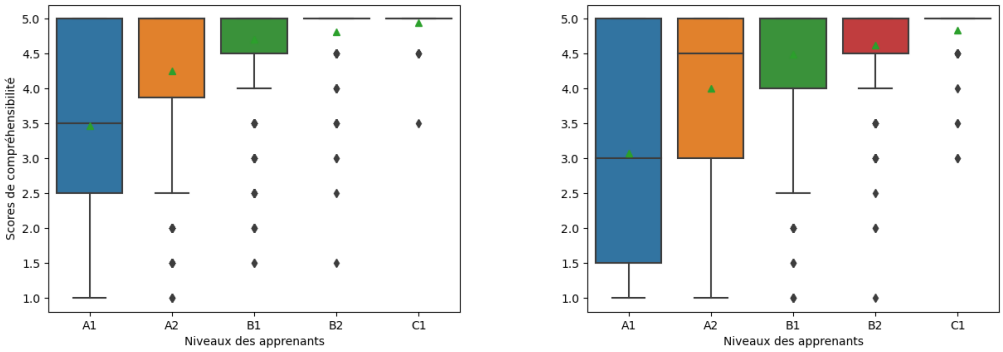


FIGURE 3.10 – Scores moyens de compréhensibilité selon le niveau des apprenants japonais pour l'évaluation *a priori* (à gauche) et l'évaluation *a posteriori* (à droite). Les triangles verts représentent les scores moyens, les traits horizontaux les scores médians.

Nous voulons vérifier s'il existe une différence significative entre les groupes de niveau CECRL selon les scores de compréhensibilité attribués aux apprenants. Les niveaux CECRL étant des données ordinales, nous appliquons le test non-paramétrique de Kruskal-Wallis. Les résultats nous informent qu'il existe une ou plusieurs différences significatives entre les groupes CECRL ($p < 0{,}001$), que ce soit pour l'évaluation *a priori* ou l'évaluation *a posteriori* de la compréhensibilité. Afin de mener notre analyse plus en détail, il est intéressant de regarder quels groupes CECRL sont significativement différents. Nous appliquons pour cela un test dit *post-hoc*, le test de Mann-Whitney (Mann et Whitney, 1947) entre chaque paire de niveau CECRL. Étant donné que nous testons 10 hypothèses (car 10 paires de niveau CECRL), nous devons ajuster le seuil de significativité pour ne pas risquer de conclusions erronées sur les significativités. Nous appliquons ainsi l'ajustement de Bonferroni (Bonferroni, 1936; Abdi *et al.*, 2007). Les groupes CECRL sont significativement différents si la *p-value* est inférieure à 0,05/10 = 0,005 (0,05 représentant la valeur usuelle de comparaison pour déduire la significativité). La table 3.6 présente les différents résultats obtenus entre chaque groupe CECRL.

Nous remarquons une différence significative entre les groupes CECRL, à l'exception des paires [B1,B2] et [B2,C1]. Ces résultats confirment tout de même que plus un

TABLE 3.6 – Indices de significativité (*p-value*) entre chaque groupe CECRL à l'issue du test non-paramétrique Mann-Whitney. Les groupes sont considérés comme significativement différents si $p < 0{,}005$ (ajustement de Bonferroni).

		A2	B1	B2	C1
Évaluation *a priori*	A1	$p < 0,001^*$	$p < 0,001^*$	$p < 0,001^*$	$p < 0,001^*$
	A2		$p < 0,001^*$	$p < 0,001^*$	$p < 0,001^*$
	B1			$p < 0,05$	$p < 0,001^*$
	B2				$p < 0,05$
Évaluation *a posteriori*	A1	$p < 0,001^*$	$p < 0,001^*$	$p < 0,001^*$	$p < 0,001^*$
	A2		$p < 0,001^*$	$p < 0,001^*$	$p < 0,001^*$
	B1			$p < 0,05$	$p < 0,001^*$
	B2				$p < 0,5$

apprenant a un niveau élevé dans la langue cible, plus son niveau de compréhensibilité est élevé.

Énoncés

Ce corpus est constitué de 40 énoncés. Chaque apprenant ayant été évalué deux fois en termes de compréhensibilité, nous disposons au total de 160 scores par énoncés, dont 80 issus de l'évaluation *a priori* et 80 issus de l'évaluation *a posteriori*.

Les scores de compréhensibilité *a priori* moyens varient entre 3,11 et 4,98 et ceux *a posteriori* varient entre 2,5 et 4,95 (voir figure 3.11). Les premiers scores de compréhensibilité sont en moyenne significativement plus élevés que les seconds (test non-paramétrique de Wilcoxon, $p < 0{,}001$). De plus, les deux énoncés ayant reçu les scores moyens les plus faibles (respectivement les plus élevés) sont les mêmes dans le cas des deux évaluations. Il s'agit des énoncés « Je dois répondre avant la semaine prochaine » (321), « Il porte un chapeau » (325), « J'aime le chocolat » (308) et « J'aime la musique » (342). Ces résultats nous permettent d'observer une certaine variabilité dans les scores moyens des énoncés traduits. Nous pouvons donc émettre l'hypothèse que les énoncés qui composent notre corpus amènent à des « degrés » de compréhensibilité différents.

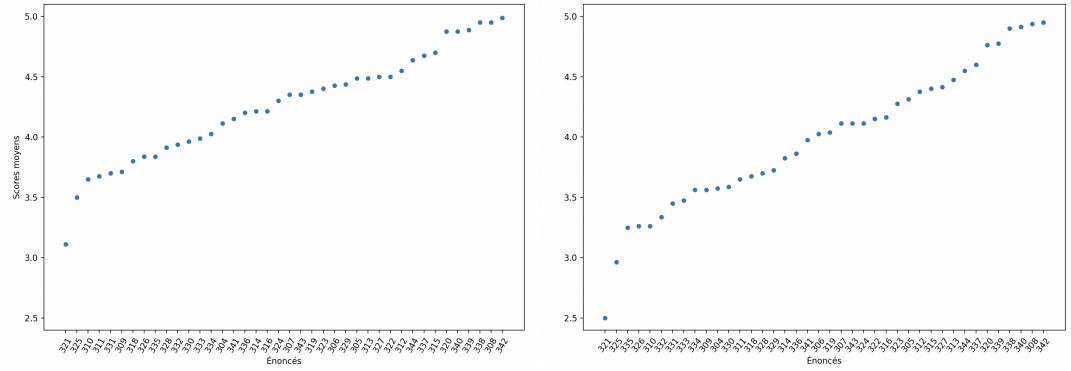


FIGURE 3.11 – Scores moyens de compréhensibilité *a priori* (à gauche) et *a posteriori* (à droite) des énoncés des apprenants japonais. En abscisse, les identifiants des énoncés tels qu'utilisés dans l'interface d'enregistrement des productions orales.

3.4 Généralisation à une autre paire de langues

Nous avons souhaité appliquer notre méthodologie de collecte de données et d'annotation de la compréhensibilité de la parole à une autre population d'apprenants. J'ai eu l'opportunité durant cette thèse de faire une mobilité de deux mois à l'université d'Ostfalia à Wolfenbüttel en Allemagne. Ceci m'a permis d'appliquer le même processus pour des apprenants allemands de français. Les étapes sont les mêmes que pour les apprenants japonais, seuls les énoncés ainsi que le focus sur les erreurs communément commises diffèrent. J'ai ainsi pu constituer un second corpus.

3.4.1 Création des énoncés de traduction

Comme pour les apprenants japonais, nous avons créé des énoncés contenant du vocabulaire censé être acquis par un apprenant de niveau CECRL A2 mais tout de même accessible à un apprenant de niveau CECRL A1. Les énoncés contiennent également des difficultés courantes que rencontrent les apprenants allemands lorsqu'ils s'expriment en français. Ces difficultés touchent également les niveaux lexical, syntaxique et morphosyntaxique. Par exemple, une erreur courante au niveau lexical est de remplacer « *obtenir* » par « *reçevoir* », au niveau syntaxique de remplacer « *répondre à* » par « *répondre* » et au niveau morphosyntaxique de remplacer « *vieil* » par « *vieux* ». Un apprenant allemand pourrait donc traduire de manière erronée « Ich habe mein Diplom erhalten. » (« J'ai obtenu mon diplôme ») par « *J'ai reçu mon diplôme* », « Wir müssen diese Frage beantworten. » (« Nous devons répondre à cette question ») par « *Nous devons répondre cette question* » et « Das ist ein alter Mann. » (« C'est un vieil homme ») par « *C'est un vieux homme* ».

En concertation avec des enseignantes de FLE à l'université d'Ostfalia, nous avons listé 45 énoncés à traduire comportant un ensemble de difficultés ciblées. Parmi ces énoncés, 20 contiennent des difficultés au niveau lexical, 22 aux niveaux syntaxique ou morphosyntaxique et 3 au niveau morphosyntaxique de genre (réalisation de « *mexicain* » au lieu de « *mexicaine* » par exemple). Chaque énoncé a été relu et corrigé par une enseignante de FLE de cette université.

3.4.2 Enregistrement des apprenants

Les enregistrements audio ont été recueillis à Wolfenbüttel (Allemagne) dans une pièce calme dans l'université Ostfalia. Étant donné le climat post-covid et la politique de cours de langues à distance en Allemagne durant cette période, nous n'avons pu enregistrer que 9 apprenants sur les 40 souhaités.

Nous avons utilisé la même interface graphique développée par Archean Technologies pour mener à bien les sessions d'enregistrement, et un microphone casque Jabra (Copenhague, Danemark) modèle Evolve 20 HSC016. Les sessions se sont déroulées exactement comme pour les apprenants japonais, avec les explications des

instructions, un entraînement (sur cinq énoncés) et un exercice de traduction (sur 40 énoncés, voir annexe C.2). En excluant les enregistrements produits lors de la phase d'entraînement, nous avons collecté 360 productions orales (représentant une durée de vingt-et-une minutes) : 40 énoncés traduits chacun par 9 apprenants allemands (2F/7M), constituant notre second corpus CAF-al.

Le niveau CECRL des apprenants se situe dans l'intervalle [A1,B1] (voir table 3.7). Contrairement aux apprenants japonais, la répartition des niveaux CECRL est ici plus homogène, mais l'étendue de ces niveaux est plus réduite.

TABLE 3.7 – Niveau CECRL selon les apprenants allemands.

Niveau CECRL	A1	A2	B1
Apprenants allemands	2	4	3

3.4.3 Annotation du corpus CAF-al

Nous avons suivi le même protocole que pour les apprenants japonais concernant l'évaluation subjective de la compréhensibilité des productions collectés auprès des apprenants allemands. Nous avons profité de l'évaluation effectuée sur les données des apprenants du corpus CAF-jp pour faire également évaluer les productions des apprenants du corpus CAF-al. Les 80 évaluateurs ont ainsi évalué neuf enregistrements audio issus des productions des apprenants allemands.

3.4.4 Analyse des annotations

Nous menons ici les mêmes analyses que pour les apprenants japonais (voir section 3.3.4), à savoir des analyses concernant les évaluateurs, les apprenants et les énoncés.

Évaluateurs

Nous avons mesuré des corrélations de Spearman (ρ), des Kappa de Cohen et des écarts inter-annotateurs moyens (écart moyen entre les scores donnés) entre les différents évaluateurs pour les évaluations des productions des apprenants allemands (voir table 3.8).

TABLE 3.8 – Résultats des accords inter-annotateurs et écart moyens.

	Évaluation *a priori*	Évaluation *a posteriori*
ρ moyen	$0,72$	$0,77$
Écart-type	$0,21$	$0,16$
κ moyen	$0,41$	$0,46$
Écart-type	$0,23$	$0,16$
Écart inter-annotateurs moyen	$0,55$	$0,54$
Écart-type	$0,29$	$0,27$

Les scores donnés par les différents évaluateurs sont fortement corrélés, que ce soient ceux concernant l'évaluation *a priori* ou ceux concernant l'évaluation *a posteriori*. Le coefficient de corrélation de Spearman moyen et le Kappa de Cohen moyen sont plus élevés concernant les scores *a posteriori*, mais l'écart inter-annotateurs moyen est, quant à lui, plus élevé concernant les scores *a priori*. Comme pour les apprenants japonais, les évaluateurs ont attribué des scores qui suivent la même tendance, avec un accord modéré et une variation inter-annotateur de 11% pour l'évaluation *a priori* et 10,8% pour l'évaluation *a posteriori*.

Les scores moyens *a priori* des productions des apprenants allemands sont de 3,92 et ceux *a posteriori* sont de 3,57 (voir table 3.9).

TABLE 3.9 – Scores donnés par les évaluateurs à l'issue de l'évaluation de la compréhensibilité de la parole des apprenants allemands (score moyen, score médian et écart-type moyen).

	Évaluation *a priori*	Évaluation *a posteriori*
Scores moyens	$3,92$	$3,57$
Scores médians	5	4
Écart-type	$1,32$	$1,47$

Nous observons les mêmes tendances que pour les résultats concernant l'évaluation des apprenants japonais, les évaluateurs ont en moyenne donné des scores plus élevés lors de l'évaluation de la compréhensibilité de la parole *a priori* qu'avec l'évaluation *a posteriori*. Les scores médians *a priori* sont eux aussi plus élevés. La différence observable entre les scores *a priori* et *a posteriori* est elle aussi significative, avec un résultat au test non-paramétrique de Wilcoxon de $p < 0,001$. D'après la figure 3.12 ci-dessous, nous pouvons observer que les évaluateurs ont majoritairement attribué le score de 5 en termes de compréhensibilité de la parole aux apprenants, quelle que soit la phase d'évaluation.

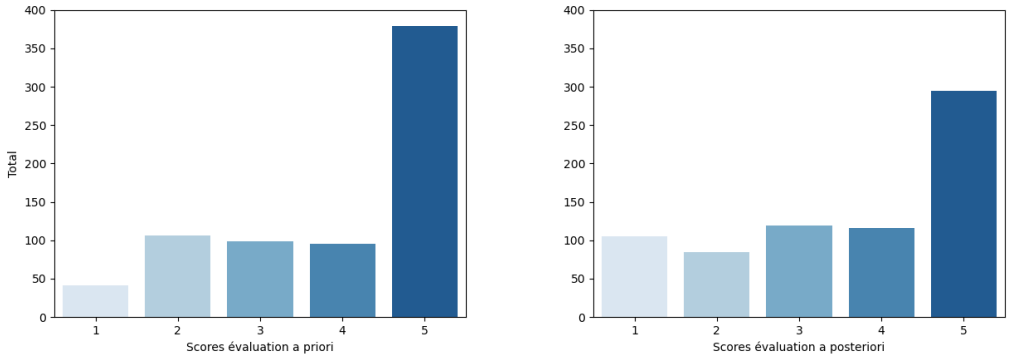


FIGURE 3.12 – Scores de compréhensibilité donnés aux productions des apprenants allemands selon les évaluations *a priori* et *a posteriori*.

Concernant les transcriptions produites par les évaluateurs, nous mesurons également le PER moyen entre chaque paire d'évaluateurs pour observer la similarité des transcriptions. Le PER moyen varie de 0,35 à 24,86 (voir figure 3.13). La paire

d'évaluateurs [23,63] a produit les transcriptions les plus ressemblantes, tandis que la paire [26,66] a produit les transcriptions les plus différentes.

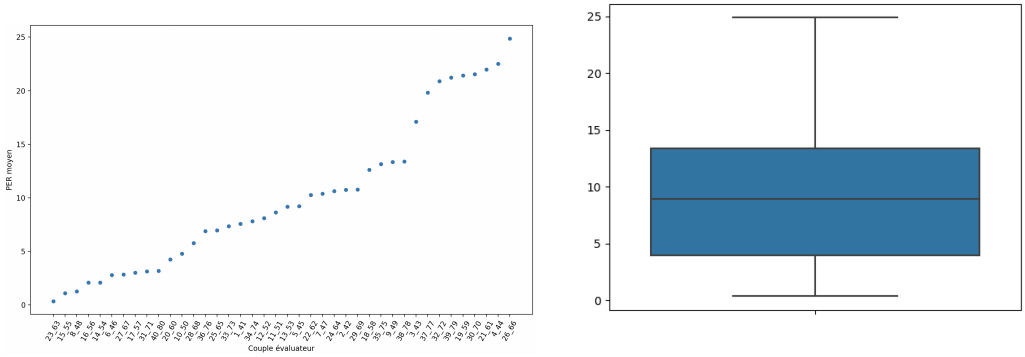


FIGURE 3.13 – PER moyen entre chaque paire d'évaluateurs pour les productions des apprenants allemands.

Nous observons également que nous ne possédons aucune valeur aberrante, ce qui pourrait signifier qu'aucune paire d'évaluateurs n'a eu à évaluer de productions beaucoup plus compliquées ou faciles par rapport aux autres paires.

Apprenants

La figure 3.14 présente la répartition des scores selon les apprenants allemands. Nous pouvons observer que tous les apprenants ont reçu au moins une fois chacun des scores aux deux évaluations, excepté un apprenant qui n'a jamais reçu le score 1 pour l'évaluation *a priori*.

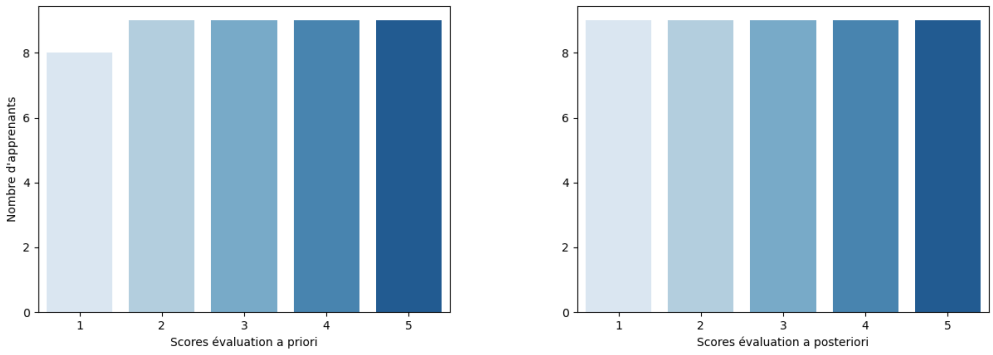


FIGURE 3.14 – Répartition des scores de compréhensibilité selon les apprenants allemands.

L'accord inter-annotateurs présenté précédemment nous permet de calculer la moyenne des scores de compréhensibilité obtenus. Ainsi, nous pouvons comparer les

scores moyens obtenus avec les niveaux CECRL des apprenants, informations obtenues en amont du processus de collecte. La figure 3.15 présente l'adéquation de ces scores en fonction du niveau CECRL.

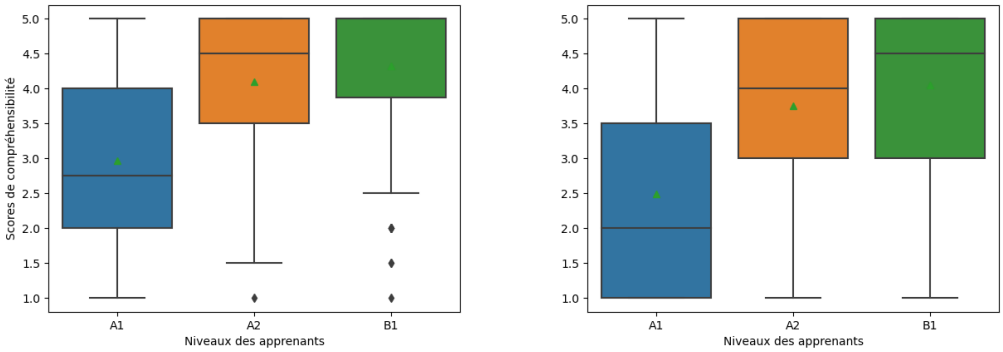


FIGURE 3.15 – Scores moyens de compréhensibilité selon le niveau des apprenants allemands pour l'évaluation *a priori* (à gauche) et l'évaluation *a posteriori* (à droite). Les triangles verts représentent les scores moyens, les traits horizontaux les scores médians.

Nous avons également appliqué le test non-paramétrique de Kruskal-Wallis à nos données, montrant qu'il existe une ou plusieurs différences significatives entre les groupes de niveau CECRL (p < 0,001) quelle que soit la phase d'évaluation. Les résultats du test *post-hoc* de Mann-Whitney entre chaque paire de niveau CECRL sont indiqués sur la table 3.10. Nous testons ici trois hypothèses (3 paires de niveaux CECRL), le seuil de significativité est donc d'environ 0,01 à l'issue de l'ajustement de Bonferroni.

TABLE 3.10 – Indices de significativité (*p-value*) entre chaque groupe CECRL.

		A2	B1
Évaluation *a priori*	A1	$p < 0,001^*$	$p < 0,001^*$
	A2		$p < 0,05$
Évaluation *a posteriori*	A1	$p < 0,001^*$	$p < 0,001^*$
	A2		$p < 0,05$

Nous remarquons qu'il existe des différences significatives entre les groupes CECRL [A1,A2] et [A1,B1] quelle que soit l'évaluation. Aucune différence significative n'est présente pour le groupe [A2,B1]. Ces résultats sont tout de même cohérents avec ceux obtenus pour les apprenants japonais (voir figure 3.10) et montrent une fois encore que plus un apprenant a un niveau élevé dans la langue cible, plus son niveau de compréhensibilité est élevé.

Énoncés

Le corpus CAF-al est lui aussi constitué de 40 énoncés, évalués deux fois pour chacun des neuf apprenants. Nous disposons donc au total de 36 scores par énoncés, dont 18 issus de l'évaluation *a priori* et 18 issus de l'évaluation *a posteriori*.

Les scores de compréhensibilité *a priori* moyens varient entre 2,66 et 4,94 et ceux *a posteriori* varient entre 2,11 et 4,83 (voir figure 3.16). Nous remarquons également que les scores de compréhensibilité *a priori* sont en moyenne significativement plus élevés que ceux *a posteriori* (test non-paramétrique de Wilcoxon, $p < 0{,}001$). Contrairement à ce que nous pouvons remarquer pour les énoncés des apprenants japonais (voir figure 3.11), les deux énoncés ayant reçu les scores moyens les plus faibles ne sont pas exactement les mêmes selon les deux phases d'évaluation. Nous retrouvons les énoncés 557 (« Elle n'est jamais à l'heure »), 568 (« Le cours est dans la salle dix ») et 566 (« J'ai fait le ménage toute la journée »). Concernant les deux énoncés ayant reçu les scores moyens les plus élevés, seul l'énoncé 549 (« Elle est mexicaine ») garde la première place quelle que soit l'évaluation, les autres étant les énoncés 556 (« Ils n'ont pas de problème ») et 561 (« J'ai dix-neuf ans »). De même que pour les énoncés des apprenants japonais, les résultats nous permettent d'observer une certaine variabilité dans les scores moyens des énoncés traduits. Nous pouvons également émettre l'hypothèse que les énoncés qui composent notre corpus amènent à des « degrés » de compréhensibilité différents.

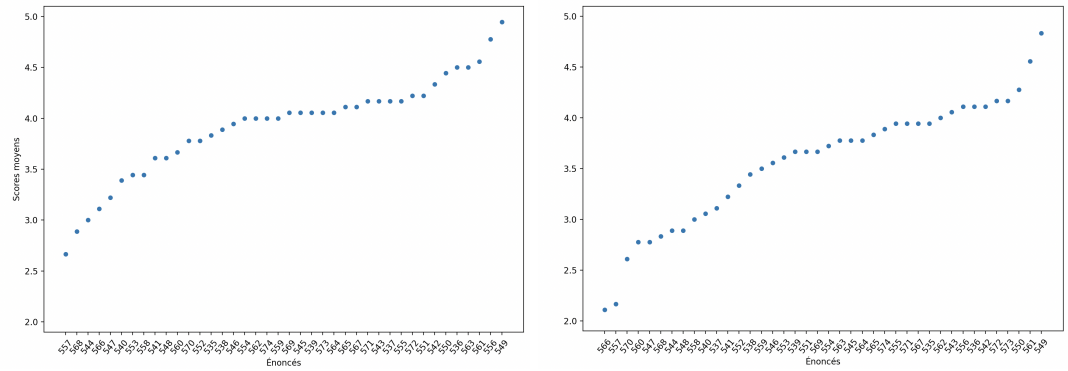


FIGURE 3.16 – Scores moyens de compréhensibilité *a priori* (à gauche) et *a posteriori* (à droite) des énoncés des apprenants allemands. En abscisse, les identifiants des énoncés tels qu'utilisés dans l'interface d'enregistrement des productions orales.

3.5 Conclusion

Nous avons mis en place un protocole afin de réaliser un corpus permettant d'évaluer la compréhensibilité de la parole d'apprenants d'une langue étrangère.

Notre protocole consiste à enregistrer des apprenants lors d'une tâche de traduction orale en L2 d'énoncés cibles écrits en L1. Les énoncés contiennent des difficultés typiques rencontrées par les apprenants lors de la traduction orale L2, aux niveaux

lexical, syntaxique et morphosyntaxique. Une évaluation subjective de la compréhensibilité de la parole a ensuite été menée sur les productions orales selon la définition introduite par Woisard *et al.* (2013).

Lors de ces évaluations, deux scores de compréhensibilité sont attribués à chaque production à l'aide de deux échelles graduées sur cinq points. Le score *a priori* correspond à la compréhensibilité perçue après l'écoute de l'enregistrement audio. Le score *a posteriori* reflète la compréhensibilité du sens du message véhiculé en tenant compte du sens initial du message à transmettre, c'est-à-dire après une prise de connaissance de la traduction attendue.

Nous avons testé la validité de notre protocole sur des apprenants japonais de français. Nous avons créé 40 énoncés japonais qui ont été traduits à l'oral par 40 apprenants japonais. Au total, et hormis les énoncés utilisés pour l'entraînement, 1600 productions ont été collectées et évaluées en termes de compréhensibilité de la parole par 80 évaluateurs français natifs. Ces évaluateurs sont âgés de 19 à 40 ans, n'ont aucune familiarité avec la langue japonaise et ne présentent pas de problème de presbyacousie connu. Le coefficient de corrélation de Spearman moyen entre les évaluations *a priori* s'élève à 0,72, avec un écart inter-annotateurs moyen de 0,43. Pour les évaluations *a posteriori*, nous retrouvons un coefficient de corrélation de Spearman moyen de 0,75, avec un écart inter-annotateurs moyen de 0,48.

Afin de généraliser notre approche à une autre paire de langues, nous l'avons appliquée sur des apprenants allemands de français. Bien que seulement neuf apprenants aient participé à l'élaboration de ce corpus, l'évaluation subjective de la compréhensibilité de la parole n'en est pas moins satisfaisante. Les scores attribués par les mêmes 80 évaluateurs sont eux aussi fortement corrélés, avec un coefficient de corrélation de Spearman moyen de 0,72 pour la première phase (*a priori*) et de 0,77 pour la seconde (*a posteriori*). Nous retrouvons des écarts inter-annotateurs moyens également peu élevés, soit 0,55 pour la première phase et 0,54 pour la seconde.

Nos résultats sont très satisfaisants concernant les évaluations subjectives de la compréhensibilité de la parole des apprenants japonais. De plus, les résultats obtenus concernant les apprenants allemands indiquent que notre protocole peut se généraliser à d'autres paires de langues.

Suite à la création de nos deux corpus CAF-jp et CAF-al, et vis-à-vis de notre définition de la compréhensibilité, nous allons prédire de manière automatique la compréhensibilité de la parole *a posteriori* d'apprenants non-natifs du français dans le chapitre suivant.

4

Prédiction de la compréhensibilité de la parole non-native

Sommaire

Nous avons, dans les chapitres précédents, présenté différentes mesures linguistiques multi-niveaux permettant d'évaluer les productions orales d'apprenants de langue étrangère. Ces évaluations s'effectuent sur les plans phonético-phonologique, lexical, syntaxique et discursif. Nous avons validé ces différents paramètres sur le corpus CLIJAF_18, constitué d'enregistrements de productions orales d'apprenants japonais de français, en démontrant que leur évolution est en adéquation avec les niveaux CECRL des apprenants (voir section 2.2.3). Nous avons également montré qu'il est possible de regrouper les apprenants selon leurs niveaux CECRL, sur la base de ces mesures et par le biais de l'algorithme d'apprentissage automatique non-supervisé *k-means* (voir section 2.2.4). De plus, la prédiction du niveau CECRL des apprenants du corpus CLIJAF_38 sur la base de cinq paramètres a révélé des résultats prometteurs.

À l'issue de ces travaux, et pour se focaliser plus spécifiquement sur la mesure de la compréhensibilité de la parole telle que définie dans le chapitre 1, nous avons réalisé deux corpus composés de productions orales d'apprenants de français, de L1 japonais (CAF-jp) ou allemand (CAF-al). Les productions orales ont été collectées par le biais d'une tâche de traduction orale en L2 d'énoncés présentés sous forme écrite en L1. Ce corpus a ensuite été enrichi par des scores de compréhensibilité de la parole évaluée de manière subjective par un panel de 80 francophones natifs.

Ce dernier chapitre présente la méthodologie proposée, implémentée et testée afin de prédire automatiquement les scores de compréhensibilité de la parole des apprenants issus de nos corpus. Dans un premier temps, nous présentons les nouvelles mesures linguistiques qui sont venues enrichir notre jeu de paramètres, au regard de la nature de la tâche de production réalisée. Dans un second temps, nous analysons les performances de différents algorithmes d'apprentissage automatique lors de la prédiction de la compréhensibilité de la parole. Nous détaillons ensuite une méthode d'optimisation du meilleur modèle de prédiction. Enfin, nous discutons des différents résultats obtenus. Nous appliquons tout d'abord cette méthode sur notre corpus CAF-jp, puis sur notre deuxième corpus CAF-al afin de tester la capacité de généralisation de notre méthode.

4.1 Enrichissement des mesures linguistiques multi-niveaux

Nous avons au préalable, dans le chapitre 2, implémenté différents paramètres linguistiques permettant d'évaluer les compétences phonético-phonologiques, lexicales, syntaxiques et discursives. Dans le cadre de notre étude sur la compréhensibilité, et vis-à-vis de la nature de nos corpus CAF-jp et CAF-al, nous pouvons enrichir notre ensemble de mesures linguistiques en rajoutant notamment des mesures segmentales, lexicales et sémantiques. En effet, nos corpus ont été constitués de façon à connaître le réel sens du message qu'un apprenant devait normalement véhiculer. Cette connaissance nous permet d'extraire des productions orales de nouveaux paramètres lexicaux et sémantiques dont le calcul nécessite une comparaison avec la phrase de référence.

4.1.1 Prononciation au niveau segmental et transcription

Prononciation des phonèmes

Dans le chapitre 2, nous avions utilisé un système de reconnaissance automatique de la parole disponible sur la plateforme PATY (que nous appelons *système PATY* dans la suite du manuscrit) afin de mesurer la qualité de la prononciation des apprenants japonais de français. Ce système est doté d'un modèle de langage permettant de transcrire la parole de manière orthographique. Pour notre étude sur la compréhensibilité de la parole, nous souhaitons utiliser une mesure de prononciation qui soit la plus proche possible du signal de parole, donc plus représentative de la prononciation des apprenants. C'est dans cette optique que nous avons utilisé un système de reconnaissance automatique de la parole présentant une architecture hybride de type TDNNF-HMM (*Factorised Time-Delay Neural Network Hidden Markov Model* ; Gelin, 2022). Ce modèle, également disponible depuis la plateforme PATY, a été entraîné sur le corpus français *Common Voice*[13] (148,9 heures de données) avec l'outil Kaldi (Povey *et al.*, 2011). La sortie du système se présente sous la forme d'une suite phonémique, avec un score de confiance associé à chaque phonème. Celui-ci varie de 0 (confiance la plus faible) à 1 (confiance la plus élevée). Pour chaque fichier audio, nous mesurons un score moyen de prononciation des phonèmes avec la formule suivante :

$$\textit{Prononciation} = \frac{1}{n}\sum_{i=0}^{n} s_{phon_i} \tag{4.1}$$

avec n le nombre de phonèmes décodés dans le fichier audio traité et s_{phon_i} le score de confiance du système pour le phonème i. Le score de prononciation varie de 0 (prononciation nulle) à 1 (prononciation parfaite).

Transcription des productions orales

Notre corpus contient les transcriptions manuelles données par les évaluateurs à l'issue de l'évaluation subjective de la compréhensibilité de la parole. Ces transcriptions ne sont pour autant pas complètes, étant donné que les évaluateurs avaient pour consigne de transcrire par un point d'interrogation chaque mot qu'ils ne parvenaient pas à distinguer à l'issue de l'écoute de la production. De plus, dans l'optique de proposer une méthode complètement automatique, nous devons également automatiser la transcription des productions. Pour aller au-delà de la transcription phonétique, nous avons utilisé un système de reconnaissance automatique de la parole doté d'un modèle de langage afin d'obtenir des transcriptions orthographiques.

Nous avons comparé les transcriptions orthographiques de cinq systèmes de reconnaissance automatique de la parole : le système PATY utilisé dans le chapitre 2, les systèmes CRDNN et Wav2vec2 de Speechbrain (Ravanelli *et al.*, 2021) et les systèmes Base et Medium de Whisper (Radford *et al.*, 2022). Nous comparons les transcriptions

13. https://commonvoice.mozilla.org/fr

orthographiques de ces différents systèmes avec les transcriptions subjectives dont nous disposons en mesurant le WER (*Word Error Rate*) moyen. Cette métrique calculée entre une transcription issue d'un système automatique et une transcription issue de l'évaluation subjective de la compréhensibilité s'obtient avec la formule suivante :

$$WER = \frac{S + O + I}{N} \tag{4.2}$$

avec S le nombre de substitutions, O le nombre d'omissions, I le nombre d'insertions et N le nombre de mots de référence. Plus le WER est faible, plus les transcriptions sont similaires (un WER de 0 implique des transcriptions identiques).

Pour rappel, nous disposons de deux transcriptions subjectives par enregistrement audio. Chacune d'elle peut être comparée à la transcription automatique obtenue avec un des quatre systèmes précédemment cités. Nous pouvons ainsi calculer, sur l'ensemble des enregistrements audio, un WER moyen par système. La table 4.1 ci-dessous présente les différents résultats obtenus.

TABLE 4.1 – Résultats du calcul du WER moyen, en pourcentage, de chaque système de reconnaissance automatique de la parole sur l'ensemble du corpus CAF-jp.

	Système PATY	**Speechbrain**		**Whisper**	
		CRDNN	**Wav2vec2**	**Base**	**Medium**
WER	$24,39\%$	$27,69\%$	$15,55\%$	$32,64\%$	$20,93\%$

Nous remarquons que le WER moyen le plus faible a été obtenu pour le système Wav2vec2 de Speechbrain. Nous appliquerons ainsi toutes nos mesures lexicales, syntaxiques et discursives sur les transcriptions générées par ce système.

4.1.2 Appropriation lexicale

Nous avons vu dans le chapitre 1 consacré à l'état de l'art sur la compréhensibilité de la parole d'apprenants de langues étrangères, qu'une des mesures lexicales généralement employée pour évaluer les productions d'apprenants est l'*appropriation lexicale*. Cette mesure permet de vérifier si un mot est employé de manière appropriée d'après le contexte d'énonciation. Elle se trouve être intéressante et particulièrement adéquate dans notre cas, c'est-à-dire lorsque nous connaissons le réel message que devait véhiculer un apprenant. Ainsi, nous sommes en mesure de savoir si la substitution d'un mot est appropriée. Cette idée d'appropriation du mot de substitution revient à mesurer si le sens de l'énoncé reste similaire ou non. De manière intuitive, un sens différent implique que la substitution n'est pas appropriée, et inversement, un sens similaire implique que la substitution est appropriée.

Il existe différentes études visant à trouver le meilleur substitut d'un mot dans une phrase donnée. Certaines de ces études se concentrent sur l'utilisation de *word embeddings* pour effectuer cette tâche (nous pouvons citer par exemple Melamud *et al.* (2015) et Zhou *et al.* (2019)). Les *word embeddings* sont des représentations vectorielles multidimensionnelles de mots, et sont principalement utilisés en traitement

automatique du langage naturel (TALN) et en apprentissage automatique. Une des particularités de cette représentation est que les mots apparaissant dans un contexte similaire ont tendance à avoir des vecteurs relativement proches. Les relations sémantiques et syntaxiques entre les mots seraient donc préservées lors de leur transformation en vecteurs. Nous nous sommes intéressés aux modèles pré-entraînés de type word2vec (Mikolov *et al.*, 2013b) et BERT (Devlin *et al.*, 2018).

Modèles word2vec

Les modèles word2vec sont des réseaux de neurones à deux couches, entraînés selon deux architectures : l'une appelée CBOW (*Continuous Bag of Words*), l'autre appelée Skip-gram. Les modèles de type CBOW visent à prédire un mot en prenant en compte son contexte (les différents mots l'entourant dans une phrase par exemple), tandis que, dans les modèles de type Skip-gram, le mot est utilisé pour prédire son contexte (Mikolov *et al.*, 2013a) (voir figure 4.1).

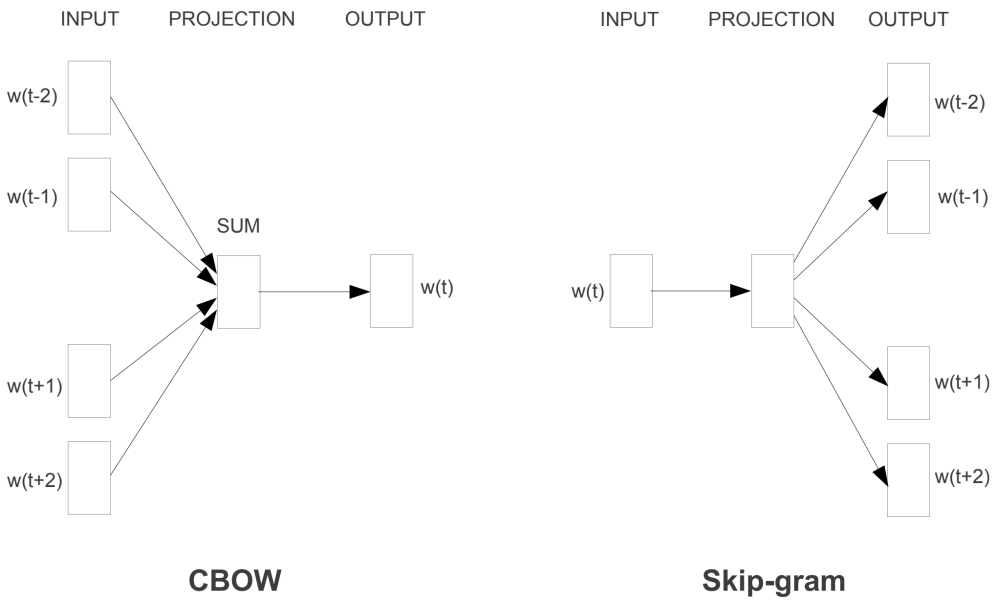


FIGURE 4.1 – Schéma des architectures CBOW et Skip-gram de Mikolov *et al.* (2013a). w(t) correspond au mot courant, w(t+1), w(t+2), w(t-1) et w(t-2) à son contexte.

Les modèles word2vec considérés ici sont pré-entraînés pour le français (Fauconnier, 2015). Il existe deux types de modèles, ceux entraînés sur le corpus frWaC, conçu dans le cadre du programme WaCky (Baroni *et al.*, 2009), et ceux entraînés sur le corpus Wikipédia français (frWiki [14]). Le corpus frWac contient 1,6 milliard de mots lemmatisés et a été construit à partir de pages internet d'extension *.fr*, tandis que le corpus frWiki contient 178,9 millions de mots issus d'environ 504.000 pages encyclopédiques de Wikipédia.

14. https://dumps.wikimedia.org/frwiki/

Modèles BERT

Les modèles de la famille BERT sont entraînés à l'aide d'un *Transformer*, un réseau de neurones utilisant un mécanisme d'attention permettant d'apprendre les relations contextuelles entre les mots. Les modèles considérés ici sont les modèles CamemBERT (Martin *et al.*, 2020) et FlauBERT (Le *et al.*, 2020), pré-entraînés pour le français. Les principales déclinaisons de CamemBERT sont les modèles CamemBERT-large, entraîné sur le corpus CCNet (135 Go de texte) (Wenzek *et al.*, 2019) et CamemBERT-base, entraîné sur le corpus OSCAR (138 Go texte) (Suárez *et al.*, 2019). Ces modèles contiennent respectivement 110 millions et 335 millions de paramètres. Concernant FlauBERT, les principaux modèles sont FlauBERT-large et FlauBERT-base, tous deux entraînés sur 75 Go de données textuelles issues de 24 sous-corpus (Wikipedia, livres [15] et Common Crawl [16], par exemple). Les deux versions se différencient de par le nombre de paramètres, à savoir 238 millions pour FlauBERT-base et 373 millions pour FlauBERT-large. De manière générale, le nombre de paramètres des modèles FlauBERT surpasse celui des modèles CamemBERT.

Mesures basées sur les *word embeddings*

Notre choix de modèle de *word embeddings* s'est porté sur les modèles BERT. En effet, word2vec ne génère qu'un *embedding* unique par mot, contrairement à BERT qui génère des *embeddings* différents selon le contexte du mot. Par exemple, le mot *espèce* désigne de la monnaie (« payer en espèce ») ou des catégories (« espèces animales, … »). Le fait qu'un mot puisse avoir plusieurs *embeddings* selon son contexte d'énonciation nous permet d'avoir plus de précision sur notre mesure. De plus, word2vec ne gère pas les mots *oov* (*out-of-vocabulary*), c'est-à-dire que si un mot ne fait pas partie du corpus utilisé pour l'entraînement, il sera impossible de le convertir en *embedding*. Concernant BERT, les mots *oov* n'amènent aucune difficulté, étant donné que les *embeddings* ne sont pas générés à partir de mots entiers, mais de sous-mots. Ainsi, BERT semble pouvoir générer une quasi-infinité d'*embeddings*, tandis que word2vec ne peut pas générer plus d'*embeddings* que de mots vus pendant l'entraînement du modèle.

Afin de sélectionner le modèle BERT que nous allons utiliser, nous comparons les performances de FlauBERT et CamemBERT sur différentes tâches du *benchmark* FLUE (*French Language Understanding Evaluation*) [17] (Le *et al.*, 2020). Ce *benchmark* regroupe différentes tâches d'évaluation pour le TAL français, telles que :

— la classification de texte (corpus CLS ; Prettenhofer et Stein, 2010),

— l'identification de paraphrases (corpus PAWS-X (Yang *et al.*, 2019), extension du corpus PAWS (Zhang *et al.*, 2019b) à 6 langues),

— l'inférence en langage naturel (corpus XNLI (Conneau *et al.*, 2018), extension du corpus MultiNLI (Williams *et al.*, 2017) à 15 langues),

15. https://www.gutenberg.org/
16. https://data.statmt.org/ngrams/deduped2017/
17. https://github.com/getalp/Flaubert/tree/master/flue

— l'analyse syntaxique et l'étiquetage en *Part Of Speech* (POS, corpus français *French Treebank* ; Abeillé *et al.*, 2003),

— la désambiguïsation du sens des mots (tâche FrenchSemEval (Segonne *et al.*, 2019), ciblant uniquement les verbes, et une version modifiée de la partie française de la tâche de désambiguïsation multilingue de SemEval-2013 (Navigli *et al.*, 2013), ciblant les noms).

Nous nous sommes référés à une étude réalisée par Kamal Eddine *et al.* (2021) afin de comparer les performances des modèles. Les performances du modèle CamemBERT-large sont meilleures que les performances des modèles CamemBERT-base, FlauBERT-base et FlauBERT-large sur les tâches de classification de texte, d'identification de paraphrases et d'inférence en langage naturel. Concernant les tâches d'étiquetage en POS et de désambiguïsation, seules les performances des modèles FlauBERT-base, FlauBERT-large et CamemBERT-base sont, à notre connaissance, comparées (Le *et al.*, 2020). Étant donné que le modèle CamemBERT-large obtient de meilleures performances sur trois tâches du corpus FLUE, et plus particulièrement sur la tâche d'identification de paraphrases qui est très proche de la dimension que nous cherchons à évaluer, nous l'avons choisi pour implémenter nos mesures basées sur les *word embeddings*. CamemBERT-large nous fournira ainsi des vecteurs de taille 1024 pour chaque mot de notre corpus.

Pour répondre aux besoins de notre étude et intégrer à nos mesures linguistiques ce qui relève de l'appropriation lexicale, nous nous sommes inspirés de la mesure développée par Zhang *et al.* (2019a). Ainsi, nous avons intégré quatre mesures basées sur le calcul de la moyenne entre les similarités cosinus maximales de chacun des *word embeddings* qui composent deux phrases (une phrase de référence et une phrase d'hypothèse). Les similarités cosinus maximales sont issues d'une matrice de similarité, obtenue en deux étapes :

1. encodage des phrases d'hypothèse et de référence en *word embeddings* ;
2. calcul de la similarité cosinus entre chaque paire de mots issus de l'hypothèse et de la référence et stockage des résultats dans une matrice.

La figure 4.2 issue de Zhang *et al.* (2019a) illustre ces étapes pour obtenir la matrice de similarité maximale entre les mots des deux phrases. La moyenne des similarités cosinus maximales peut être pondérée avec les fréquences inverses de chaque mot, ce qui permet de donner plus d'importance aux mots moins fréquents dans la langue cible, reconnus comme plus significatifs pour la similarité sémantique que les mots courants (Banerjee et Lavie, 2005; Vedantam *et al.*, 2015).

Sur cette figure, la moyenne des similarités cosinus maximales correspond à :

$$\textit{Moyenne SimCos} = \frac{0,713 + 0,515 + 0,858 + 0,796 + 0,913}{5} \tag{4.3}$$

et la moyenne pondérée à :

$$\textit{Moyenne pondérée SimCos} = \frac{(0,713 \times 1,27) + (0,515 \times 7,94) + ... + (0,913 \times 8,88)}{1,27 + 7,94 + 1,82 + 7,90 + 8,88} \tag{4.4}$$

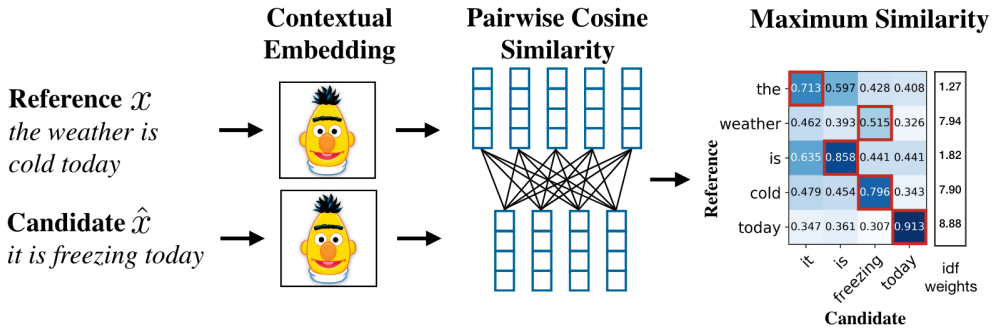


FIGURE 4.2 – Illustration du calcul général effectué pour obtenir la matrice de similarité cosinus entre les mots de deux phrases (Zhang *et al.*, 2019a, p. 4). Les scores entourés en rouge dans la matrice de similarité correspondent aux similarités cosinus maximales obtenues entre deux mots. Le vecteur à droite de la matrice correspond aux fréquences inverses (idf) des mots de la phrase de référence.

La similarité cosinus entre deux *word embeddings* s'obtient avec la formule suivante :

$$SimCos = \frac{A.B}{\|A\|\|B\|} \tag{4.5}$$

avec A et B représentant deux *word embeddings*. La similarité cosinus est comprise dans l'intervalle [-1,1], avec -1 indiquant que les vecteurs sont colinéaires mais de sens opposé, 0 indiquant que les vecteurs sont orthogonaux et 1 indiquant que les vecteurs sont colinéaires. Dans notre cas où les vecteurs représentent des *word embeddings*, plus la similarité cosinus est proche de 1, plus les mots ont le même sens, avec 1 équivalent à des mots identiques.

Contrairement à Zhang *et al.* (2019a), nous mesurons nos scores d'appropriation lexicale entre des phrases lemmatisées. La lemmatisation nous permet de ne pas tenir compte de différents éléments morphologiques, tels que la conjugaison, l'accord sujet-verbe, le genre et le nombre, et de nous concentrer uniquement sur l'appropriation d'une unité lexicale en elle-même. Nos mesures permettent de calculer les quatre scores suivants :

— **appropriation lexicale - tous les mots (TM)** : ce score correspond à la moyenne des similarités cosinus maximales entre tous les mots de deux phrases lemmatisées,

— **appropriation lexicale pondérée - tous les mots (TMP)** : score précédent pondéré par les fréquences inverses des mots. Les fréquences inverses sont obtenues en inversant les fréquences des mots de la BDD Lexique3 (version 3.83) (New *et al.*, 2005),

— **appropriation lexicale - mots lexicaux (ML)** : ce score correspond à la moyenne des similarités cosinus maximales entre les mots lexicaux de deux phrases lemmatisées ;

— **appropriation lexicale pondérée - mots lexicaux (MLP)** : score précédent pondéré par la fréquence inverse des mots, en inversant les fréquences contenues dans la BDD Lexique3.

En plus d'avoir un sens similaire au mot remplacé dans la production de l'apprenant, un bon substitut doit également respecter le sens primaire de la phrase dans laquelle il intervient (Zhou *et al.*, 2019). Par exemple, un bon substitut pour le mot « voiture » dans la phrase « La voiture roule vite » serait le mot « automobile », et un mauvais substitut pour le mot « ordinateur » dans la phrase « J'ai apporté mon ordinateur » serait le mot « tablier ». Une des manières de vérifier la similarité entre les phrases est d'utiliser les *sentence embeddings*.

Mesures basées sur les *sentence embeddings*

Les *sentence embeddings* sont des représentations vectorielles de phrases. Ces représentations permettent, entre autres, de comparer des phrases différentes d'un point de vue sémantique (Reimers et Gurevych, 2019). Il existe différents modèles de *sentence embeddings*. Cependant, pour garder une certaine cohérence entre nos mesures, nous avons choisi d'utiliser le modèle de *sentence embeddings* CamemBERT-large. À partir de ce modèle, nous obtenons deux paramètres linguistiques supplémentaires, à savoir :

— **similarité sémantique** : ce score correspond à la similarité cosinus entre une phrase de référence et une phrase hypothèse. Les mots qui composent les phrases ne sont pas lemmatisés, permettant ainsi de garder les informations morphologiques (la conjugaison par exemple) ;

— **appropriation lexicale - phrase** : ce score correspond à la similarité cosinus entre deux phrases lemmatisées. La lemmatisation permet de faire abstraction de la morphologie (genre, nombre, accord sujet-verbe, conjugaison) et de se concentrer uniquement sur le choix des unités lexicales. La lemmatisation permet ainsi de faire rentrer cette mesure dans la catégorie *appropriation lexicale*.

Pour pouvoir mesurer l'appropriation lexicale et la similarité sémantique des différentes productions orales des apprenants, nous utilisons la traduction attendue (voir table C.1 en annexe, colonne « Énoncé cible ») comme phrase de référence et la transcription fournie par le système Speechbrain Wav2vec2 sélectionné au préalable (voir section 4.1.1) comme phrase hypothèse. Nous avons ajouté ces six nouveaux paramètres linguistiques à notre jeu de paramètres et pouvons maintenant les utiliser dans notre processus de prédiction du score de compréhensibilité.

4.2 Prédiction et analyse quantitative des performances

Cette section présente les différentes méthodes proposées afin de prédire de manière automatique la compréhensibilité des productions orales des apprenants du corpus CAF-jp. Les prédictions s'effectuent sur la base des différents paramètres linguistiques décrits précédemment et extraits automatiquement pour chaque enregistrement audio traité. Les productions orales des apprenants sont ainsi représentées par un total de 18 paramètres extraits grâce aux mesures présentées dans la table 4.2.

TABLE 4.2 – Paramètres extraits des productions des apprenants et caractéristiques des différents niveaux linguistiques décrits dans ce manuscrit

Niveaux	Paramètres
Prononciation	Score moyen d'identification des phonèmes
Fluence	Débit de parole Pourcentage de parole Écart-type de la durée des pseudo-syllabes Nombre normalisé de pauses silencieuses
Lexique	Diversité lexicale Densité lexicale Sophistication lexicale Appropriation lexicale (TM) Appropriation lexicale (TMP) Appropriation lexicale (ML) Appropriation lexicale (MLP) Appropriation lexicale (phrase)
Syntaxe	Longueur moyenne des productions Profondeur moyenne des arbres syntaxiques
Discours	Proportion de connecteurs du discours Diversité des connecteurs du discours
Sémantique	Similarité sémantique

Scores de compréhensibilité à prédire

Les productions orales de notre corpus CAF-jp ont été évaluées de manière subjective par deux évaluateurs en termes de compréhensibilité (voir chapitre 3). Les évaluations ont porté sur la compréhensibilité *a priori* et *a posteriori*, résultant en deux scores de compréhensibilité par production orale. Nous nous concentrons ici sur la prédiction de la compréhensibilité *a posteriori*, étant donné qu'elle représente le plus la compréhensibilité d'un apprenant lorsque l'auditeur a connaissance du message à véhiculer. L'accord inter-annotateurs obtenu entre les différentes évaluations ($\rho = 0{,}75$) nous permet de prendre en compte la moyenne des scores de compréhensibilité *a posteriori* afin que chaque production orale possède un seul et unique score de compréhensibilité compris entre 1 (compréhensibilité la plus faible) et 5 (compréhensibilité la plus élevée).

Modèles de prédiction

Nous utilisons et comparons les performances de deux modèles de prédiction différents. Le premier correspond à un modèle de régression linéaire LASSO, également utilisé dans le chapitre 2. Contrairement à une régression linéaire simple, ce type de modèle permet d'avoir un aperçu des paramètres ayant contribué aux prédictions grâce à son hyper-paramètre de régularisation α, permettant ainsi une certaine interprétation des résultats. Pour le deuxième modèle, nous avons opté pour un modèle de régression non linéaire. Étant donné que l'extraction des paramètres linguistiques s'opère sur la transcription obtenue à l'issue de la reconnaissance de la parole, il est possible qu'une mauvaise prononciation puisse entraîner une transcription erronée, affectant par la suite les scores des paramètres de plus haut niveau (lexicaux, syntaxiques et discursifs). Par conséquent, la combinaison de nos paramètres ne serait pas linéaire. Nous avons donc choisi d'utiliser l'algorithme de régression Random Forest (voir annexe D), qui appartient aux méthodes d'ensemble, et notée RF dans la suite du manuscrit.

Méthodologie de prédiction

Plusieurs stratégies de fusion sont explorées, notamment la fusion précoce, la fusion intermédiaire et la fusion tardive. La fusion précoce équivaut à la construction d'un seul système de régression pour prédire les scores de compréhensibilité de la parole à partir des paramètres dont nous disposons. Pour la fusion intermédiaire, un système de régression par « famille » de paramètres est construit, et enfin, concernant la fusion tardive, un système de régression par paramètre est construit.

Afin de nous assurer que les enregistrements audio d'un même locuteur ne figurent pas en même temps dans le jeu d'entraînement et dans le jeu de test, nous utilisons une stratégie de validation croisée appelée *leave-one-speaker-out* pour évaluer les performances des systèmes de prédiction. Contrairement à la stratégie *leave-one-out* (voir chapitre 2, section 2.3.3), qui sépare le jeu de données de taille N en un jeu d'entraînement de taille $N-1$ et un jeu de test de taille 1, le *leave-one-speaker-out* sépare le jeu de données en un jeu d'entraînement de taille $N-L$ et un jeu de test de taille L, avec L le nombre d'enregistrements d'un même locuteur. Nous adoptons également une stratégie de validation croisée imbriquée afin de trouver les meilleurs hyper-paramètres qui optimisent les performances des systèmes de prédiction (voir chapitre 2, figure 2.7).

Les scores de compréhensibilité *a posteriori* de chaque enregistrement sont prédits à l'aide de nos différentes méthodes de prédiction et de stratégies de fusion. Comme chaque apprenant a produit 40 enregistrements différents, nous prédisons 40 scores de compréhensibilité par apprenant, soit un score par enregistrement. Nous calculons ensuite la moyenne des 40 scores pour obtenir un unique score moyen par apprenant. Afin d'évaluer les performances des modèles, nous mesurons également un score moyen de la vérité terrain, c'est-à-dire un score moyen de compréhensibilité

a posteriori à partir des 40 scores obtenus pour chaque apprenant (voir figure 4.3 pour une illustration schématisée).

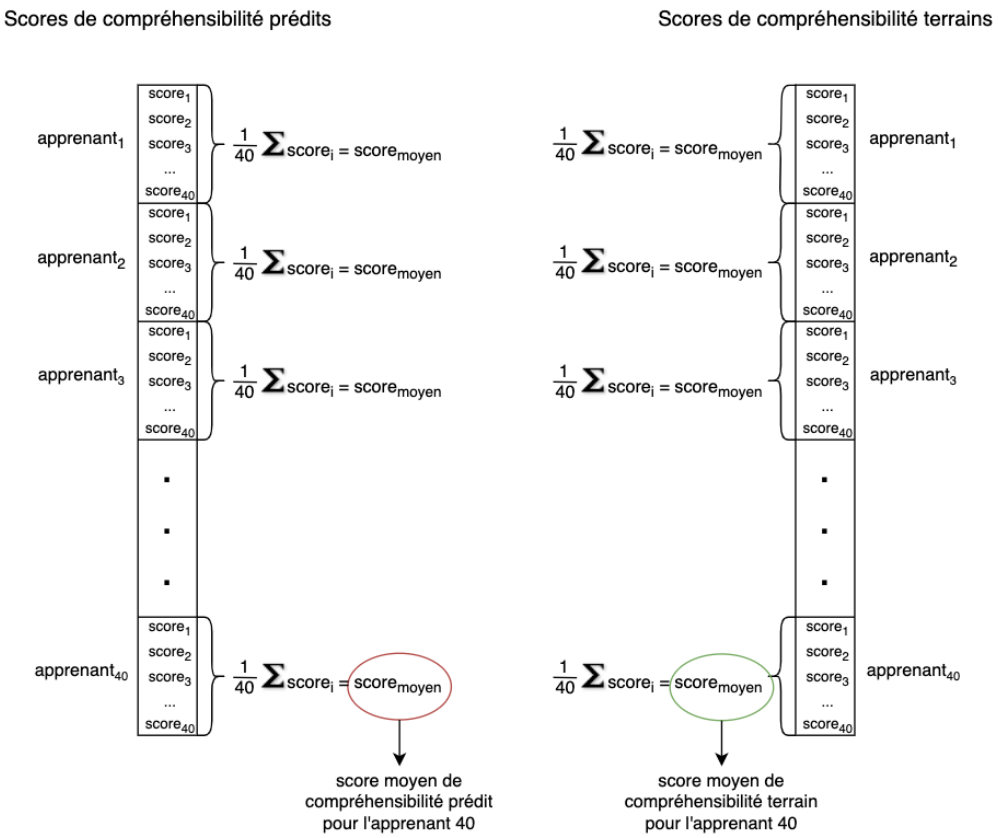


FIGURE 4.3 – Illustration schématisée du calcul des scores moyens représentant chaque apprenant. À gauche, les scores moyens obtenus à l'issue de la prédiction automatique, à droite les scores moyens issus de l'annotation manuelle humaine.

Les performances des systèmes sont mesurées en termes de corrélation de Pearson (r), de MAE et de coefficient de détermination R^2 entre les prédictions et les vérités terrain (voir équations 4.6 et 2.12). Le coefficient de détermination indique le pourcentage de variance de la variable dépendante, ici le score de compréhensibilité, expliquée par les variables indépendantes, ici nos paramètres. Il correspond au carré du coefficient de corrélation de Pearson et varie de 0 à 1, 0 indiquant que le modèle utilisé n'explique pas la variabilité des données, et 1 indiquant que le modèle explique toute la variabilité des données. Le système idéal pour la prédiction de la compréhensibilité de la parole serait alors un système qui maximise le coefficient de corrélation de Pearson et qui minimise la MAE.

$$r = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{4.6}$$

avec $cov(X,Y)$ la covariance des variables X et Y et σ_X et σ_Y leurs écarts-types.

4.2.1 Fusion précoce

Cette stratégie de fusion consiste à prédire les scores de compréhensibilité sur la base de l'ensemble des paramètres dont nous disposons, à savoir 18 (voir figure 4.4). Dans cette première phase de prédiction, nous comparons les résultats obtenus *via* l'utilisation de la régression LASSO et de la régression RF.

La table 4.3 présente les résultats obtenus lors de la prédiction de la compréhensibilité par le biais de la régression LASSO et de la régression RF. La figure 4.5 présente la droite de régression de la meilleure méthode.

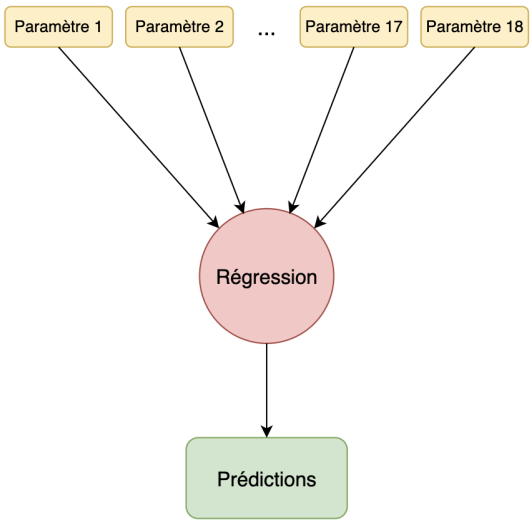


FIGURE 4.4 – Illustration schématisée de la stratégie de fusion précoce des scores de prédiction.

TABLE 4.3 – Performances de prédiction de la compréhensibilité en fusion précoce : corrélation de Pearson (r), significativité (p-*value*), MAE ($\pm$std), R^2.

Algorithme	***r***	***p-value***	**MAE**	$\boldsymbol{R^2}$
LASSO	$0,95$	$< 0,001$	$0,18(\pm 0,16)$	$0,91$
Random Forest	$0,97$	$< 0,001$	$0,16(\pm 0,13)$	$0,94$

Comme nous pouvons le constater dans la table 4.3, nous obtenons de meilleurs résultats en termes de corrélation et d'erreur moyenne avec la régression RF. Nous comparons également les deux coefficients de corrélation en appliquant une transformation de Fisher afin d'estimer la significativité de la différence de performances en utilisant l'outil de Lee et Preacher (2013). Cette transformation résulte en un Z score égal à -2,69 (p < 0,01). Par convention, les Z scores ayant une valeur absolue supérieure à 1,96 sont considérés comme significatifs. Nous pouvons donc en conclure que les tendances sont significativement mieux prédites par la régression RF que par la régression LASSO.

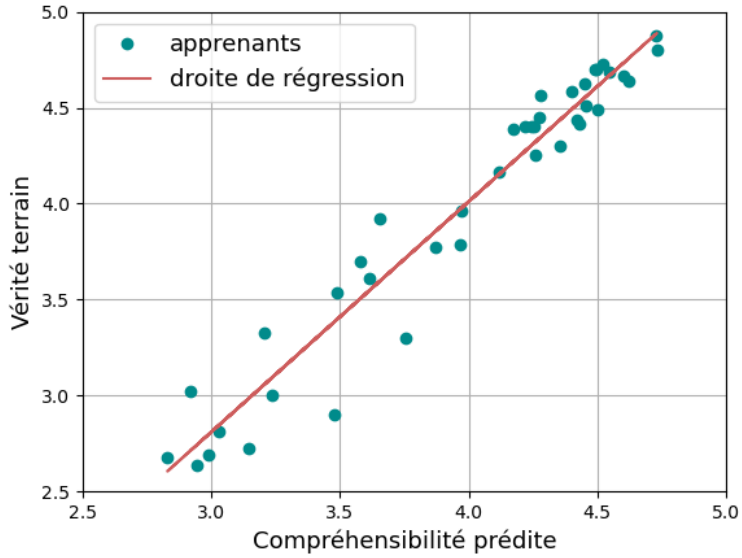


FIGURE 4.5 – Nuage de points représentant les scores moyens prédits par la régression RF par rapport à la vérité terrain (fusion précoce).

4.2.2 Fusion intermédiaire

Pour cette stratégie de fusion, nous construisons sept systèmes de régression, un par famille de mesures utilisées (voir table 4.4 et figure 4.6). Nous calculons ensuite la moyenne des sept scores de compréhensibilité afin d'obtenir un unique score par apprenant.

La table 4.5 présente les résultats obtenus lors de la prédiction de la compréhensibilité en fusion intermédiaire avec la régression LASSO et la régression RF. La droite de régression obtenue avec la meilleure méthode est visible sur la figure 4.7.

Nous obtenons une nouvelle fois de meilleurs résultats avec l'algorithme Random Forest, et ce de manière significative (Z score de -8,31, $p < 0{,}001$). Les performances de cet algorithme, bien que moins élevées qu'avec la fusion précoce, sont tout de même très élevées.

4.2.3 Fusion tardive

Nous construisons ici un système de prédiction par paramètre (voir figure 4.8). Nous prédisons au total 18 scores de compréhensibilité par apprenant, puis calculons la moyenne afin que chaque apprenant ait un seul score. La table 4.6 présente les résultats obtenus à l'issue de la prédiction. Nous obtenons une nouvelle fois de très bonnes performances avec la régression RF (voir figure 4.9), légèrement supérieures à celles obtenues avec la régression LASSO (Z score de 0,6 donc non significatif), mais inférieures à celles obtenues avec la stratégie de fusion précoce.

TABLE 4.4 – Familles et mesures dont sont issus les paramètres utilisés pour la fusion intermédiaire.

Familles	Paramètres
Prononciation	Score moyen d'identification des phonèmes
Fluence	Débit de parole Pourcentage de parole Écart-type de la durée des pseudo-syllabes Nombre normalisé de pauses silencieuses
Richesse lexicale	Diversité lexicale Densité lexicale Sophistication lexicale
Appropriation lexicale	Appropriation lexicale (TM) Appropriation lexicale (TMP) Appropriation lexicale (ML) Appropriation lexicale (MLP) Appropriation lexicale (phrase)
Complexité syntaxique	Longueur moyenne des productions Profondeur moyenne des arbres syntaxiques
Cohésion du discours	Proportion de connecteurs du discours Diversité des connecteurs du discours
Sémantique	Similarité sémantique

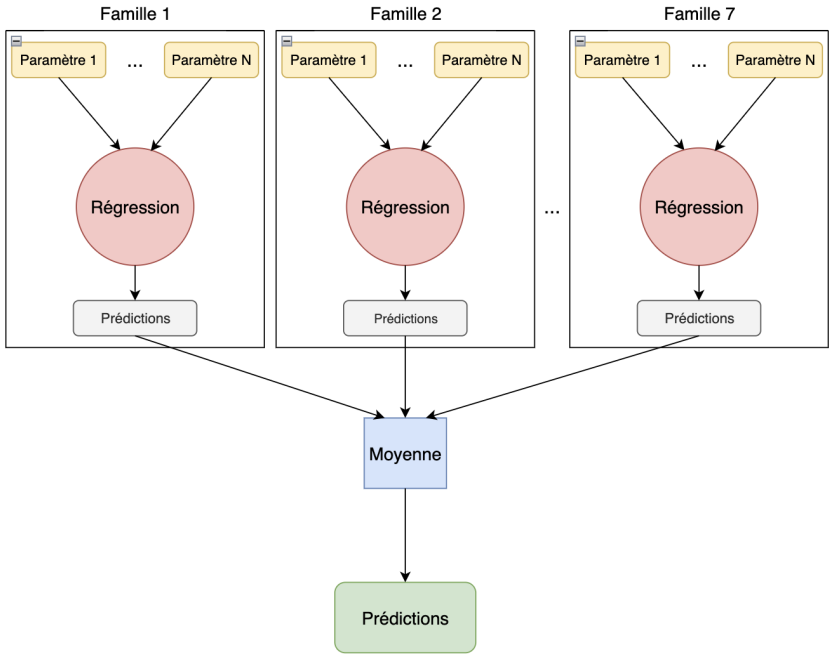


FIGURE 4.6 – Illustration schématisée de la stratégie de fusion intermédiaire des scores de prédiction.

4.2.4 Bilan des prédictions

L'utilisation de deux algorithmes d'apprentissage automatique et de trois stratégies de fusion a permis d'obtenir de très bons résultats. Bien que la régression LASSO soit

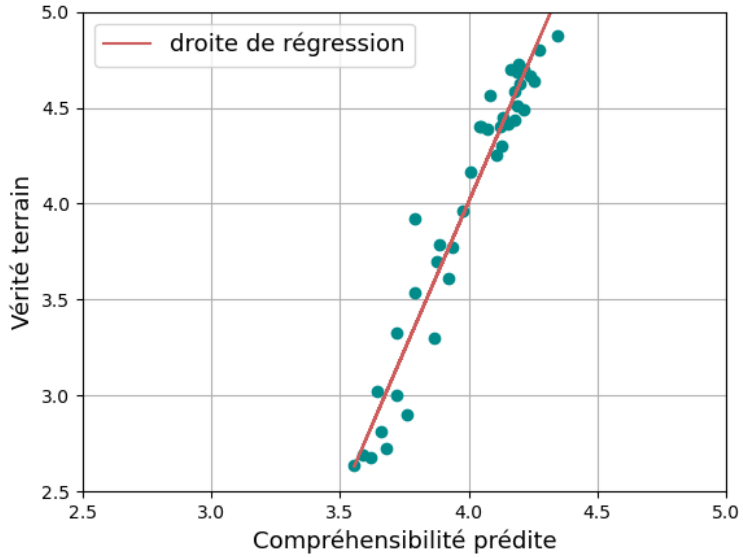


FIGURE 4.7 – Nuage de points représentant les scores moyens prédits par la régression RF par rapport à la vérité terrain (fusion intermédiaire).

TABLE 4.5 – Performances de prédiction de la compréhensibilité en fusion intermédiaire : corrélation de Pearson (r), significativité (*p-value*), MAE ($\pm$std), R^2.

Algorithme	***r***	***p-value***	**MAE**	$\boldsymbol{R^2}$
LASSO	$0,39$	$< 0,05$	$0,60(\pm 0,33)$	$0,15$
Random Forest	$0,96$	$< 0,001$	$0,43(\pm 0,25)$	$0,93$

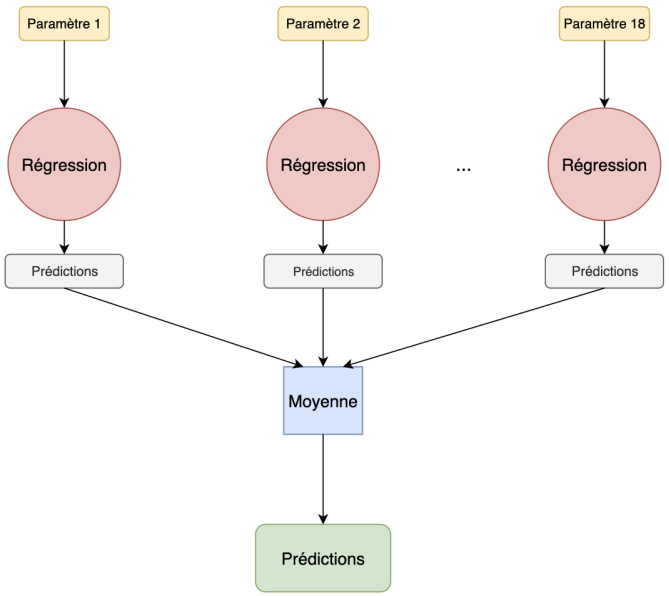


FIGURE 4.8 – Illustration schématisée de la stratégie de fusion tardive des scores de prédiction.

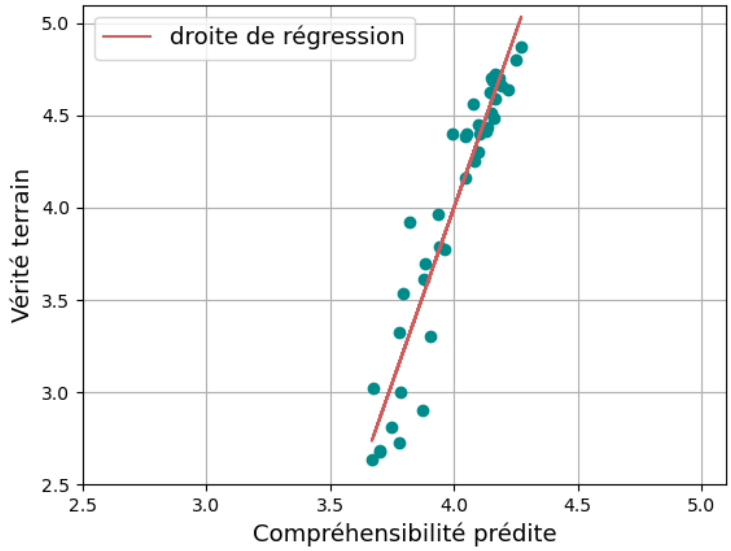


FIGURE 4.9 – Nuage de points représentant les scores moyens prédits par la régression RF par rapport à la vérité terrain (fusion tardive).

TABLE 4.6 – Performances de prédiction de la compréhensibilité en fusion tardive : corrélation de Pearson (r), significativité (*p-value*), MAE ($\pm$std), R^2.

Algorithme	***r***	***p-value***	**MAE**	$\boldsymbol{R^2}$
LASSO	$0,95$	$< 0,001$	$0,54(\pm 0,30)$	$0,90$
Random Forest	$0,94$	$< 0,001$	$0,46(\pm 0,27)$	$0,90$

performante (r = 0,95 (p < 0,001), MAE = 0,18 ($\pm$0,16), R^2 = 0,91), elle est tout de même significativement inférieure à la méthode basée sur l'algoritme Random Forest en fusion précoce (r = 0,97 (p < 0,001), MAE = 0,16 ($\pm$0,13), R^2 = 0,94). Nous constatons également que la fusion précoce permet d'obtenir de meilleurs résultats de prédiction que la fusion intermédiaire ou tardive. Dans notre cas, cette première stratégie est donc à privilégier. De plus, l'erreur moyenne de prédiction que nous obtenons (MAE = 0,16 ($\pm$0,13)) est inférieure à l'écart moyen inter-annotateurs, qui est de 0,48 (voir chapitre 3, table 3.4). Notre système de prédiction se rapproche donc d'une évaluation humaine.

Pour rappel, notre système de prédiction se base sur 18 paramètres issus de différents niveaux linguistiques. Afin de pouvoir interpréter au mieux les résultats en termes de contribution des paramètres, il est nécessaire d'étudier l'apport d'une phase de sélection de ces derniers.

4.3 Interprétabilité et analyse qualitative des prédictions

Maintenant que nous avons obtenu de bonnes performances en termes de prédiction de la compréhensibilité de la parole des apprenants du corpus CAF-jp, nous allons explorer la possibilité d'améliorer nos prédictions et de les rendre interprétables. Nous analyserons ensuite les résultats obtenus de manière qualitative.

4.3.1 Réduction du jeu de paramètres

Lorsque nous utilisons un large ensemble de paramètres pour entraîner un modèle, nous prenons le risque d'utiliser des informations inutiles pouvant faire baisser les performances. Certains paramètres pouvant être corrélés entre eux ou redondants, il est important de procéder à une phase de sélection en amont de l'entraînement, permettant ainsi de réduire cet ensemble au minimum nécessaire et suffisant afin d'améliorer les performances d'un modèle. En plus d'apporter une potentielle amélioration au niveau des performances, la sélection de paramètres aiderait également à réduire la complexité du modèle et à rendre l'interprétation des résultats plus simple. Pour mener à bien cette sélection, nous utilisons une approche statistique appelée BorutaPy (Kursa et Rudnicki, 2010). Les paramètres statistiquement moins pertinents sont éliminés de manière itérative, jusqu'à aboutir à un ensemble réduit de paramètres pertinents pour les prédictions. Les différentes étapes de l'algorithme BorutaPy sont :

1. création du duplicata : à partir des données existantes $\{x_1 ... x_n\}$, BorutaPy crée un duplicata et mélange les observations pour chaque paramètre (voir figure 4.10). Les nouveaux paramètres notés $\{x'_1 ... x'_n\}$ sont appelés des *paramètres fantômes* (*shadow features*),
2. création d'un nouveau jeu de données : le duplicata créé à l'étape précédente est concaténé au jeu de données initial,
3. entraînement : le jeu de données est utilisé pour entraîner un modèle basé sur la RF,
4. calcul de l'importance : l'importance de chaque paramètre est mesurée en utilisant une fonction propre à la méthode Random Forest (*Mean Decrease Accuracy*, Breiman (2001) ou *Mean Decrease Impurity*, Breiman (2002)). L'importance d'un paramètre représente la perte de précision entre le paramètre d'origine et son paramètre fantôme,
5. calcul des Z scores : les Z scores des paramètres sont mesurés sur la base de l'importance mesurée au point 4 ;
6. comparaison et décision : à chaque itération, l'algorithme compare les Z scores entre un paramètre d'origine et les paramètres fantômes. Si le Z score du paramètre d'origine est meilleur que ceux des paramètres fantômes, alors le paramètre

est marqué comme étant important. Par exemple, si le Z score de x_1 est supérieur au maximum des Z scores de $\{x'_1 ... x'_n\}$, alors x_1 est important.

Le schéma de la figure 4.10 illustre les deux premières étapes de cette méthode.

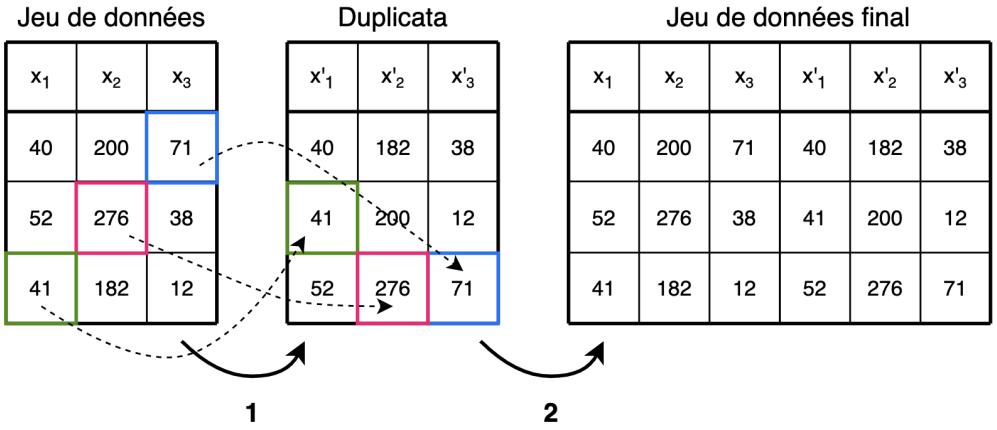


FIGURE 4.10 – Illustration schématisée de l'algorithme BorutaPy. Chaque ligne correspond à une donnée et chaque colonne à un paramètre.

L'application de l'algorithme BorutaPy sur nos données nous a permis d'identifier trois paramètres pertinents, à savoir le pourcentage de parole, le score de similarité sémantique et le score d'appropriation lexicale au niveau de la phrase. Les résultats de la prédiction de la compréhensibilité de la parole en utilisant uniquement ces trois paramètres avec l'algorithme Random Forest et la stratégie de fusion précoce sont présentés sur la figure 4.11. La table 4.7 présente une comparaison des performances obtenues, avant et après la mise en place de la sélection de paramètres.

TABLE 4.7 – Performances de prédiction de la compréhensibilité avec et sans sélection de paramètres : corrélation de Pearson (r), MAE ($\pm$std), R^2.

Algorithme	***r***	***p-value***	**MAE**	$\boldsymbol{R^2}$
Random Forest	$0,97$	$< 0,001$	$0,16(\pm 0,13)$	$0,94$
Sélection paramètres + Random Forest	$0,97$	$< 0,001$	$0,15(\pm 0,12)$	$0,94$

Nous constatons une légère hausse des performances en termes d'erreur moyenne de prédiction. Même si la seule amélioration concerne l'erreur moyenne, ce système de prédiction nous permet tout de même de mieux interpréter les résultats. Étant donné que seulement trois paramètres ont été utilisés au lieu de 18, nous pouvons conclure que chaque score prédit reflète les performances de fluence phonétiques, lexicales et sémantiques d'un apprenant.

4.3.2 Analyse des scores des apprenants

Pour compléter cette analyse, nous nous sommes focalisés sur les scores terrains dans le but de mieux comprendre la répartition des apprenants en termes de compréhensibilité. Nous mesurons pour cela l'écart-type des 40 scores de compréhensibilité

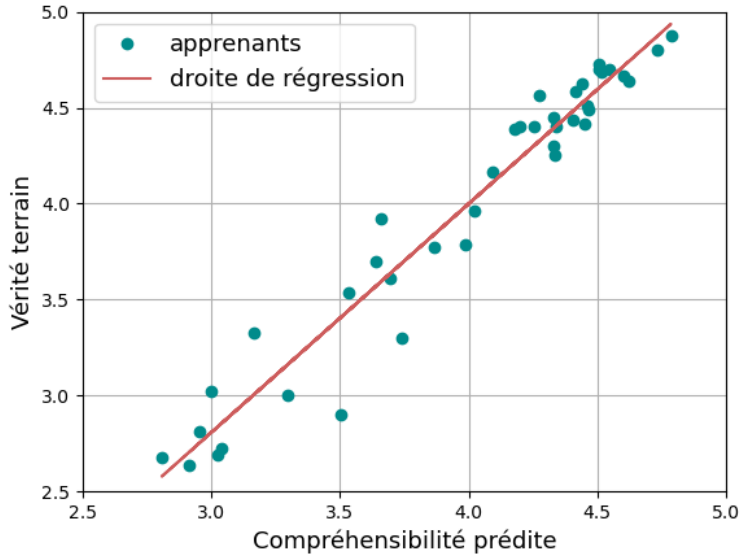


FIGURE 4.11 – Nuage de points représentant les scores moyens prédits par l'algorithme Random Forest à l'issue de la sélection de paramètres par rapport à la vérité terrain (fusion précoce).

terrain reçus par chaque apprenant. Un score de compréhensibilité moyen égal à 5 et un écart-type très faible signifieraient que les annotateurs ont en moyenne attribué presque systématiquement un score de 5 à un apprenant. Au contraire, un écart-type très élevé signifierait que les annotateurs n'ont pas été consistants dans les scores attribués. La figure 4.12 présente ces différents résultats.

Nous pouvons observer que les écarts-types des scores terrains de compréhensibilité des apprenants ayant des scores moyens supérieurs à 4,5 sont très faibles (presque tous < 0,8). Ces apprenants ont donc reçu des scores plutôt consistants pour tous leurs enregistrements audio. Nous observons aussi que, plus les scores sont faibles, plus les écarts-types sont élevés, signifiant une plus grande diversité dans les scores donnés par les annotateurs.

4.3.3 Réduction du jeu de données

Actuellement, nous prédisons la compréhensibilité de la parole d'un apprenant à partir de 40 enregistrements audio et de trois paramètres. L'analyse que nous présentons ici a pour but d'estimer s'il est possible de réduire le nombre d'enregistrements audio tout en gardant des performances satisfaisantes. Nous rappelons qu'un enregistrement audio correspond à la traduction d'un énoncé distinct. Dans l'optique d'industrialiser notre méthodologie de prédiction, il serait intéressant d'enregistrer un ensemble limité d'énoncés qui puisse tout de même nous fournir une bonne approximation de la compréhensibilité d'un apprenant.

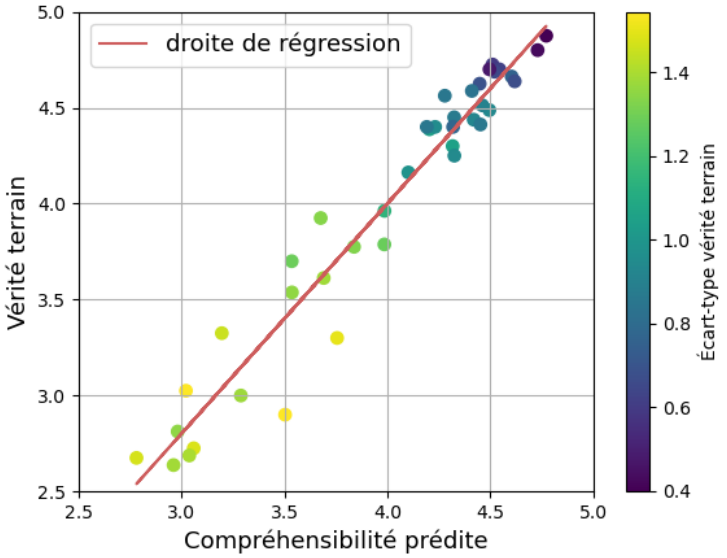


FIGURE 4.12 – Nuage de points représentant les scores moyens prédits par la régression RF à l'issue de la sélection de paramètres par rapport à la vérité terrain, colorés selon l'écart-type de la vérité terrain.

Nous avons mesuré les coefficients de corrélation et les erreurs moyennes des prédictions de la compréhensibilité de la parole en faisant varier le nombre d'énoncés qui constituent notre jeu de test. Nous avons entraîné notre système de prédiction de la même manière que précédemment, avec la stratégie de fusion précoce et une validation croisée imbriquée de type *leave-one-speaker-out*. Nous effectuons les prédictions uniquement sur un ensemble réduit de données. Nous avons testé toutes les combinaisons de N énoncés possibles, avec N allant de 1 à 6 (nous n'avons pas dépassé $N = 6$ par souci de temps de calcul). La figure 4.13 présente l'évolution des coefficients de corrélation moyens et des erreurs moyennes.

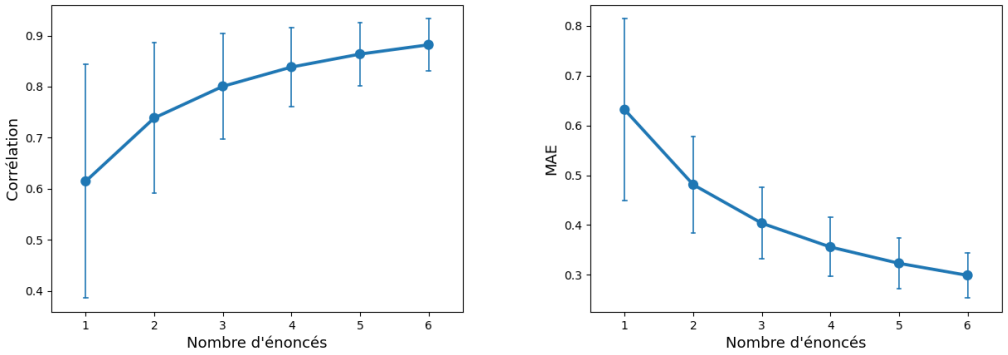


FIGURE 4.13 – Évolution de la corrélation (à gauche) et de la MAE (à droite) selon le nombre d'énoncés (moyenne et écart-type).

Bien évidemment, nous remarquons que l'augmentation du nombre d'énoncés a tendance à améliorer la corrélation moyenne et à diminuer les MAE moyennes. De même, les écarts-types, que ce soit pour les coefficients de corrélation ou les MAE, ont tendance à être de moins en moins élevés. La table 4.8 présente les résultats obtenus pour les meilleures combinaisons (minimisant la MAE) selon le nombre d'énoncés.

TABLE 4.8 – Résultats obtenus par les meilleures combinaisons selon le nombre d'énoncés : corrélation de Pearson (r) et MAE.

Nombre d'énoncés	**1**	**2**	**3**	**4**	**5**	**6**
r	$0,22$	$0,76$	$0,67$	$0,87$	$0,84$	$0,94$
MAE	$0,18$	$0,20$	$0,17$	$0,15$	$0,14$	$0,13$

En moyenne, plus le jeu de données contient d'énoncés, plus les performances sont meilleures, avec des coefficients de corrélation moyens allant de $r = 0,61$ (un énoncé) à $r = 0,88$ (six énoncés) et des MAE allant de 0,63 (un énoncé) à 0,29 (six énoncés) si on se réfère aux résultats de la figure 4.13. Nous remarquons également sur la table 4.8 que les performances obtenues en prenant la meilleure combinaison de six énoncés sont équivalentes aux performances obtenues avec 40 énoncés (jeu de données complet). Ces différents résultats nous montrent qu'il est tout à fait possible d'obtenir une très bonne approximation du niveau de compréhensibilité des apprenants japonais en leur demandant de traduire seulement six énoncés.

Une analyse encore plus détaillée nous permet de vérifier le type d'énoncés présents dans les meilleures combinaisons. Pour rappel, les énoncés ont été construits par des enseignants de FLE afin de contenir un type de difficulté précis. Les difficultés peuvent ainsi être d'ordre lexical, syntaxique ou morphosyntaxique. La table 4.9 permet de visualiser la diversité des difficultés présentes dans les meilleures combinaisons d'énoncés et la table 4.10 présente les énoncés des meilleures combinaisons.

TABLE 4.9 – Diversité des difficultés (type et nombre) dans les meilleures combinaisons d'énoncés.

Combinaisons	**Lexicale**	**Syntaxique**	**Morphosyntaxique**
1 énoncé	0	0	1
2 énoncés	0	0	2
3 énoncés	0	0	3
4 énoncés	1	0	3
5 énoncés	1	0	4
6 énoncés	2	2	2

Nous remarquons dans un premier temps sur la table 4.10 que la combinaison de trois énoncés équivaut aux énoncés des combinaisons d'un et deux énoncés. Les combinaisons de quatre et cinq énoncés sont aussi équivalentes aux combinaisons de $N - 1$ énoncés, avec seulement l'ajout d'un énoncé. Enfin, la combinaison de six énoncés contient seulement deux énoncés présents dans les autres combinaisons et quatre nouveaux énoncés. Dans un second temps, nous pouvons observer que les meilleures combinaisons de un, deux et trois énoncés contiennent uniquement des

Table 4.10 – Énoncés présents dans les meilleures combinaisons.

Combinaisons	Énoncés
1 énoncé	« Il est japonais »
2 énoncés	« Elle est étudiante » « Ma mère est italienne »
3 énoncés	« Elle est étudiante » « Il est japonais » « Ma mère est italienne »
4 énoncés	« Le professeur est dans la salle dix » « Elle est étudiante » « Il est japonais » « Ma mère est italienne »
5 énoncés	« Le professeur est dans la salle dix » « Elle est étudiante » « Il est japonais » «Elle est japonaise » « Ma mère est italienne »
6 énoncés	« J'ai vu un film » « Demain je vais rendre visite à ma mère » « J'ai apporté mon ordinateur » « Elle est étudiante » « Ma mère est italienne » « J'aime la musique »

énoncés avec une difficulté morphosyntaxique. À partir de quatre énoncés, au moins un énoncé contient une difficulté lexicale. Enfin, la difficulté ciblant la syntaxe n'est présente que pour la meilleure combinaison de six énoncés. Il est intéressant de noter que la combinaison de six énoncés, qui donne les meilleures performances, contient exactement deux énoncés avec une difficulté lexicale, deux énoncés avec une difficulté syntaxique et deux énoncés avec une difficulté morphosyntaxique. Cette combinaison permet ainsi de prédire la compréhensibilité de la parole sur la base d'une diversité des difficultés équilibrée.

4.4 Ouverture à une autre langue maternelle

Afin de vérifier si notre méthodologie de prédiction de la compréhensibilité de la parole non-native reste cohérente pour des apprenants de langue maternelle différente, nous l'avons appliquée à notre corpus CAF-al.

4.4.1 Premiers résultats de prédiction

Nous utilisons la stratégie de prédiction ayant donné les meilleurs résultats avec les apprenants japonais, à savoir la fusion précoce avec l'algorithme Random Forest, pour créer un système de prédiction des scores de compréhensibilité de la parole des apprenants allemands. Nous appliquons également une validation croisée imbriquée de type *leave-one-speaker-out*. La figure 4.14 présente les résultats de la prédiction.

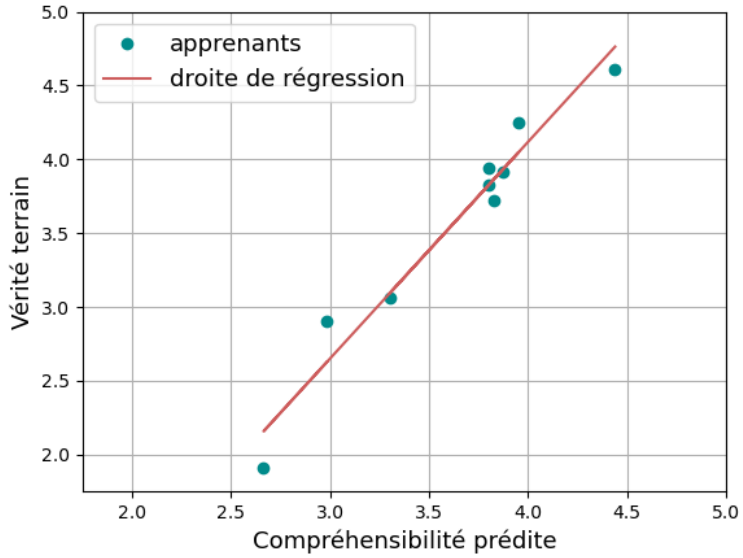


FIGURE 4.14 – Nuage de points représentant les scores moyens prédits par l'algorithme Random Forest par rapport à la vérité terrain pour les apprenants allemands (fusion précoce).

Même si notre jeu de données ne contient que neuf apprenants allemands, nous obtenons tout de même de très bons résultats, avec un coefficient de corrélation de $r = 0{,}97$ ($p < 0{,}001$), une MAE de 0,20 ($\pm 0{,}21$) et un R^2 de 0,95.

4.4.2 Sélection de paramètres

Nous mettons également en place une phase de sélection de paramètres afin de pouvoir rendre nos résultats plus interprétables. Nous utilisons l'algorithme BorutaPy pour sélectionner les paramètres les plus pertinents pour la prédiction. Étonnamment, ce sont les trois mêmes paramètres que pour les apprenants japonais qui ont été identifiés, à savoir le pourcentage de parole, le score de similarité sémantique et le score d'appropriation lexicale au niveau de la phrase. Nous observons une hausse des performances, avec une corrélation de $r = 0{,}98$ ($p < 0{,}001$), une MAE de 0,18 ($\pm 0{,}17$) et un R^2 de 0,96 (voir figure 4.15).

4.4.3 Vers une application industrielle : apprentissage et inférence sur deux L1 différentes

La stratégie de prédiction la plus intéressante d'un point de vue industriel consiste à ne devoir entraîner qu'un seul système de prédiction qui puisse être mis en production pour faire de l'inférence sur de nouvelles données. Dans notre cas, et étant donné que notre corpus CAF-jp contient le plus de données (1600 fichiers audio contre 360 pour

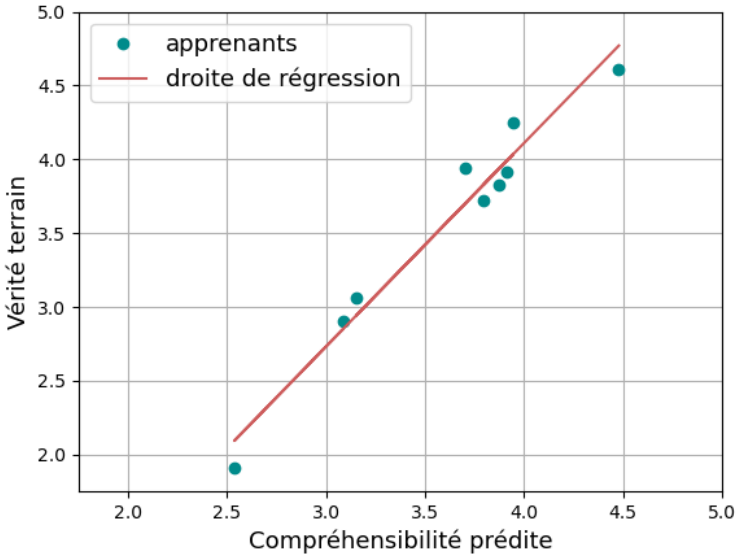


FIGURE 4.15 – Nuage de points représentant les scores moyens prédits par l'algorithme Random Forest à l'issue de la sélection de paramètres par rapport à la vérité terrain pour les apprenants allemands (fusion précoce).

CAF-al), cela consiste à réaliser l'entraînement sur toutes les données du corpus CAF-jp et de tester sur les données du corpus CAF-al.

Nous entraînons une régression RF, avec la stratégie de fusion précoce, sur toutes les données des apprenants japonais. Étant donné que, dans la section 4.3.1, nous avons obtenu de meilleures performances en n'utilisant que trois paramètres sur les 18, nous n'utilisons ici que ces trois mêmes paramètres afin d'avoir, dès le début, des résultats interprétables. Nous faisons ensuite une inférence sur les données des apprenants allemands afin de prédire leurs scores de compréhensibilité. La figure 4.16 présente le nuage de points et la droite de régression obtenue.

Les performances sont, une fois de plus, très élevées, avec un coefficient de corrélation de $r = 0{,}98$ ($p < 0{,}001$), une MAE de 0,34 ($\pm 0{,}25$) et un R^2 de 0,97. Bien que la MAE soit un peu plus élevée que lorsque nous faisons l'entraînement sur les données des apprenants allemands (voir section 4.4.2), elle reste tout de même très acceptable. Ces résultats nous permettent d'affirmer qu'il est tout à fait possible d'utiliser un système entraîné sur une L1 spécifique pour prédire les scores de compréhensibilité d'apprenants ayant une L1 différente.

4.5 Conclusion

Nous avons enrichi notre jeu de paramètres en implémentant de nouvelles mesures linguistiques d'appropriation lexicale et de similarité sémantique. Sur la base de 18 paramètres, nous avons comparé les résultats de la prédiction de la compréhensibilité

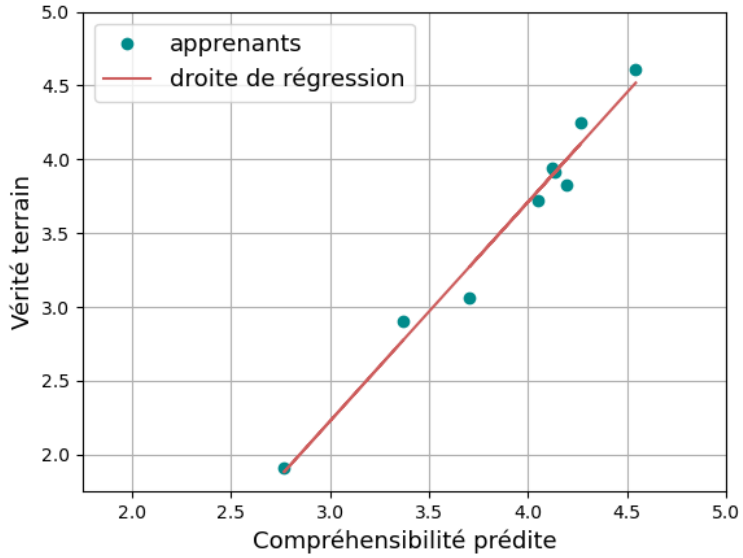


FIGURE 4.16 – Nuage de points représentant les scores moyens prédits par la régression RF à l'issue de la sélection de paramètres par rapport à la vérité terrain pour les apprenants allemands (fusion précoce, entraînement sur les apprenants japonais).

de la parole en utilisant trois stratégies de fusion différentes (fusion précoce, fusion intermédiaire et fusion tardive) et deux algorithmes de prédiction, la régression linéaire de type LASSO et la régression Random Forest. Les meilleures performances pour le corpus CAF-jp ont été obtenues avec l'algorithme Random Forest en fusion précoce : $r = 0{,}97$, $p < 0{,}001$, MAE = 0,16 $\pm$0,13 et $R^2 = 0{,}94$.

Ces performances étant atteintes avec l'utilisation de 18 paramètres, cela rend difficile l'interprétation des résultats. Nous avons donc mis en place une phase de sélection de paramètres en utilisant la méthode appelée BorutaPY. Cette méthode, basée sur des techniques statistiques, nous a permis d'identifier de manière itérative les trois paramètres les plus pertinents à la prédiction, à savoir le pourcentage de parole, le score de similarité sémantique et le score d'appropriation lexicale au niveau de la phrase. Sur la base de ces trois paramètres, nous avons pu atteindre de meilleures performances de prédiction : $r = 0{,}97$, $p < 0{,}001$, MAE = 0,15 ($\pm$0,12) et $R^2 = 0{,}94$. Étant donné que l'écart inter-annotateurs moyen obtenu au chapitre 3 section 3.3.4 est plus élevé que notre erreur moyenne de prédiction, nous pouvons considérer que notre système se rapproche d'une évaluation humaine de la compréhensibilité de la parole non-native. Une analyse qualitative de nos données a également révélé que les apprenants ayant un score de compréhensibilité supérieur à 4,5 ont reçu des scores de compréhensibilité peu variables par les évaluateurs, alors que les scores sont plus hétérogènes lorsque la compréhensibilité diminue. De plus, nous avons remarqué que nous pouvons atteindre une très bonne approximation de la compréhensibilité de la parole en réduisant notre jeu de données à seulement six énoncés. La combinaison de six énoncés qui obtient les meilleures performances de prédiction ($r = 0{,}94$, MAE = 0,13) est constituée d'une

diversité pédagogiquement équilibrée, avec deux énoncés contenant une difficulté au niveau lexical, deux au niveau morphosyntaxique et deux au niveau syntaxique. Cette approche permet donc aux apprenants de traduire seulement six énoncés au lieu des 40 initiaux.

Concernant le corpus CAF-al, nous obtenons également de très bonnes performances de prédiction avec l'algorithme Random Forest et la stratégie de fusion précoce : $r = 0{,}97$, $p < 0{,}001$, MAE = 0,20 ($\pm$0,21) et $R^2 = 0{,}95$. En réduisant l'ensemble de paramètres utilisés à seulement trois paramètres à la suite de l'application de la méthode BorutaPy, nous avons pu légèrement améliorer nos performances de prédiction : $r = 0{,}98$, $p < 0{,}001$, MAE = 0,18 ($\pm$0,17) et $R^2 = 0{,}96$.

Enfin, nous avons montré qu'il est tout à fait possible d'entraîner un modèle sur des données d'une L1 spécifique et d'arriver à de bonnes performances en testant sur des données d'une autre L1. En effet, en réalisant l'entraînement sur les données des apprenants japonais, nous prédisons de manière efficace la compréhensibilité de la parole des apprenants allemands : $r = 0{,}98$, $p < 0{,}001$, MAE = 0,34 ($\pm$0,25) et $R^2 = 0{,}97$. Ces résultats nous confortent dans l'idée qu'il est tout à fait envisageable, d'un point de vue industriel, de ne devoir entraîner qu'un seul modèle afin de prédire la compréhensibilité de la parole d'apprenants de français, et ce de manière totalement indépendante de leur L1.

Conclusions et perspectives

Les travaux de recherche présentés dans ce mémoire ont été réalisés dans le cadre d'une collaboration entre l'entreprise Archean Technologies et le laboratoire IRIT, au travers du LabCom ALAIA (financement ANR-18-LCV3-001 (FR)) et d'une thèse CIFRE obtenue dans ce contexte. L'objectif était de caractériser la compréhensibilité de la parole d'apprenants d'une langue étrangère, et de proposer une méthode permettant de l'évaluer automatiquement en utilisant des technologies d'intelligence artificielle.

Conclusions générales

Comment caractériser la compréhensibilité de la parole de personnes apprenant une langue étrangère ?

Dans le premier chapitre, nous avons étudié le concept de compréhensibilité telle qu'utilisée dans les domaines de la didactique des langues. Ce travail nous a permis de proposer une définition de la compréhensibilité inspirée de la littérature. La compréhensibilité a ainsi pu être définie comme étant la « Capacité de l'auditeur à interpréter le sens du message oral produit par un locuteur, sans tenir compte de la précision ou de la justesse phonétique ou lexicale » (Woisard *et al.*, 2013). Contrairement à l'intelligibilité de la parole, qui repose essentiellement sur des critères acoustico-phonétiques, la compréhensibilité fait également intervenir les niveaux lexicaux, morphosyntaxiques et discursifs, et agit au niveau sémantico-discursif. De ce fait, la compréhensibilité fait référence à la compréhension du sens du message véhiculé par un locuteur. Nous pouvons ainsi la représenter comme une construction de dimensions linguistiques, allant de la phonétique-phonologie à la sémantique.

Néanmoins, l'aspect linguistique ne permet pas de caractériser complètement la compréhensibilité. En effet, nous avons pu observer dans la littérature que celle-ci pouvait également être influencée par d'autres facteurs. En premier lieu, la tâche de production orale utilisée afin de susciter la parole d'apprenants. En effet, la nature de cette tâche amène à une utilisation variée des ressources linguistiques de l'apprenant, en partie liée au temps de planification, à la familiarité avec le sujet traité et aux objectifs de communication. En deuxième lieu, nous avons pu observer des différences de compréhensibilité selon le profil de l'auditeur. Nous avons vu dans la littérature qu'un auditeur familier avec la L1 d'un apprenant a plus de facilités à comprendre

la production de ce dernier et la jugera comme plus compréhensible qu'un auditeur non familier. Il en est de même pour des auditeurs ne partageant pas la même langue maternelle. Selon la similarité linguistique entre la L1 d'un auditeur et la L2 d'un locuteur, le locuteur sera reconnu comme étant plus ou moins compréhensible. Nous avons également vu que l'expérience d'un auditeur, d'un point de vue linguistique ou pédagogique/didactique, exerce une influence sur la compréhensibilité d'un apprenant. En effet, les connaissances linguistiques en L2 d'un auditeur lui permettent de mieux appréhender les productions orales des apprenants de cette même langue. Le dernier facteur non linguistique pouvant impacter la compréhensibilité est le profil de l'apprenant. D'un point de vue phonético-phonologique, la langue maternelle d'un apprenant a une influence sur sa compréhensibilité, et ce indépendamment de la potentielle familiarité d'un auditeur avec son accent L1.

Quel protocole faut-il définir et mettre en place pour collecter des données pertinentes ?

Une contribution importante de ce travail de thèse a été de constituer un corpus qui nous permette d'évaluer la compréhensibilité de la parole non-native. Le protocole a été co-construit avec des experts en apprentissage du français à l'université de Waseda au Japon. Afin de rendre compte de la compréhensibilité d'un apprenant vis-à-vis du réel sens qu'il souhaite véhiculer, et non du message effectivement produit, nous avons créé une tâche de traduction orale. Cette méthode de collecte de productions orales nous permet d'avoir une référence (la traduction cible) et ainsi de pouvoir évaluer la compréhensibilité que nous ciblons. Les énoncés à traduire sont spécialement conçus pour contenir chacun une difficulté typique de traduction rencontrée par les apprenants en L2. Cette difficulté peut intervenir au niveau lexical, morphosyntaxique ou syntaxique.

Avec l'aide d'experts en linguistique nous avons créé 45 énoncés à traduire par des apprenants japonais de français, dont 5 ont servi pour l'entraînement. Nous avons pu récolter les productions orales de 40 apprenants, traduisant chacun les 40 énoncés proposés. Notre premier corpus CAF-jp se compose ainsi de 1600 productions orales d'apprenants japonais de français. Nous avons appliqué le même protocole en soumettant 45 énoncés spécifiquement conçus par des enseignants de FLE pour un public d'apprenants allemands de français. La collecte de nouvelles productions réalisées par neuf apprenants allemands nous a permis de constituer un second corpus, CAF-al, contenant 360 enregistrements audio.

Quel protocole faut-il mettre en place pour annoter ces données et obtenir une vérité terrain en termes de compréhensibilité ?

La compréhensibilité de la parole des productions orales collectées a été évaluée de manière subjective par des locuteurs natifs du français, afin de recueillir des scores de

compréhensibilité dits *scores de terrain*. Le protocole défini tient compte des différents facteurs pouvant affecter la compréhensibilité et qui ne sont pas liés à la compétence ou à la performance de l'apprenant en L2. Ces facteurs pouvant constituer des biais pour la prédiction de la compréhensibilité des apprenants, nous nous sommes efforcés de les contrôler en recrutant 80 évaluateurs francophones natifs âgés de 18 à 40 ans, ne présentant aucun problème de presbyacousie et aucune familiarité avec la langue ou l'accent L1. Sur la base de la définition de Woisard *et al.* (2013), deux scores de compréhensibilité ont été attribués aux productions orales *via* une interface développée spécifiquement. Le premier score, que nous avons désigné par *score a priori* tient compte de la compréhensibilité sans connaissance du message à véhiculer. Le second score, appelé *score a posteriori*, correspond à la mesure de la compréhensibilité après prise de connaissance du réel sens que devait véhiculer l'apprenant. L'étude de l'accord inter-annotateurs que nous avons menée a révélé de bons résultats. Pour les enregistrements du corpus CAF-jp, nous obtenons un coefficient de corrélation de Spearman de $\rho = 0{,}72$ pour les scores *a priori* et un coefficient de corrélation de Spearman de $\rho = 0{,}75$ pour les scores *a posteriori*. Nous avons ainsi pu vérifier que les évaluateurs ont eu tendance à interpréter la définition de la compréhensibilité de la parole de la même manière, bien qu'une part de subjectivité réside toutefois dans l'interprétation et l'évaluation du fait de *comprendre*. Cette première campagne d'annotation du corpus CAF-jp nous a permis de collecter 3200 scores de compréhensibilité, dont 1600 *a priori* et 1600 *a posteriori*.

Pour notre second corpus CAF-al, les évaluations ont été réalisées avec le même groupe d'évaluateurs. Nous avons également pu obtenir des accords inter-annotateurs satisfaisants, avec un coefficient de corrélation de $\rho = 0{,}72$ pour les scores *a priori* et de $\rho = 0{,}77$ pour les scores *a posteriori*. Pour ce corpus, 720 scores de compréhensibilité ont été collectés, dont 360 *a priori* et 360 *a posteriori*.

Ces résultats d'évaluation subjective de la compréhensibilité d'apprenants japonais et allemands de français indiquent que notre protocole peut s'appliquer à d'autres paires de langues L1/français. En utilisant les scores collectés lors de la réalisation de ces deux corpus, nous pouvons explorer la possibilité de prédire de manière automatique la compréhensibilité de la parole d'apprenants non natifs de français.

Quelle méthode proposer pour mesurer de manière automatique la compréhensibilité des apprenants ?

Paramètres multi-niveaux

Pour valider expérimentalement le rôle des mesures linguistiques identifiées lors de notre état de l'art, une première étude a été menée à partir du corpus CLIJAF. Ce corpus contient des enregistrements d'apprenants japonais de français. Bien que la tâche sous-jacente n'ait pas été directement liée à la compréhensibilité, nous avons pu exploiter ce corpus grâce à la collaboration avec le Professeur Sylvain Detey, chercheur en didactique des langues qui a dirigé la collecte de ces données. En premier lieu,

l'objectif a été de prédire le niveau CECRL des apprenants, à partir de différentes mesures linguistiques.

Les mesures identifiées comme jouant un rôle dans la compréhensibilité se trouvent l'être également pour l'évaluation du niveau de compétences linguistiques. Elles correspondent à des jeux de paramètres extraits aux niveaux phonético-phonologique, lexical, syntaxique et discursif. Nous avons montré, sur la base de seulement six de ces paramètres, qu'un regroupement des apprenants selon leurs niveaux CECRL était possible à l'aide de l'algorithme d'apprentissage automatique non supervisé *k-means* (score de pureté de 0,88, indice de Rand de 0,85). Ces travaux ont donné lieu à une publication dans la conférence francophone JEP (Journées d'Études sur la Parole) (De Fino *et al.*, 2022a). Étant donné que le niveau CECRL rend essentiellement compte des compétences globales des apprenants (compétences orales et, principalement, écrites), nous avons enrichi le corpus en faisant évaluer le niveau en production orale par trois enseignants experts de FLE. Sur la base de ces évaluations collectées *via* une interface développée spécifiquement, et de cinq paramètres linguistiques, nous avons appliqué l'algorithme de régression linéaire LASSO selon le principe de la validation croisée imbriquée *leave-one-out*. Nous avons obtenu de bonnes performances de prédiction du niveau des apprenants en production orale, avec un coefficient de corrélation de $r = 0{,}71$ et une MAE de 0,53. Ce travail a donné lieu à une publication dans la conférence internationale Interspeech 2022 (De Fino *et al.*, 2022b). Nous avons ainsi pu démontrer la pertinence de la combinaison de paramètres linguistiques pour évaluer des productions orales d'apprenants L2.

Notre protocole de collecte de données nous permettant de disposer d'une référence quant au sens du message censé être véhiculé par les apprenants, nous avons pu enrichir notre jeu de paramètres lors du dernier chapitre. Ces paramètres sont d'ordre lexical et sémantique. Nous avons ainsi utilisé 18 paramètres au total, issus des niveaux phonético-phonologique, lexical, syntaxique, discursif et sémantique, afin de prédire la compréhensibilité de la parole d'apprenants non-natifs du français.

Prédiction de la compréhensibilité

Nous nous sommes concentrés sur la prédiction des scores *a posteriori* des apprenants japonais collectés comme décrit dans le troisième chapitre. En effet, les scores *a posteriori* correspondent à la compréhensibilité des apprenants compte tenu du sens initial du message à transmettre, à l'inverse du score *a priori* qui ne rend compte que de la compréhensibilité perçue par l'évaluateur. L'étude sur les accords inter-annotateurs nous a permis d'agréger les scores et de prendre en compte la moyenne des scores de chaque production. Ainsi, notre méthodologie a consisté à prédire un score de compréhensibilité par production orale. Comme le volume des données collectées ne permettait pas d'organiser les données en corpus d'apprentissage et de test, nous avons procédé à une validation croisée imbriquée de type *leave-one-speaker-out*. Nous avons ensuite fait une agrégation des scores obtenus afin de n'avoir qu'un seul score de compréhensibilité par apprenant. Les algorithmes de régression linéaire LASSO et

de régression Random Forest ont été utilisés dans nos expériences afin de comparer les résultats obtenus avec un modèle linéaire et un modèle non-linéaire. Nous avons également exploré différentes stratégies de fusion afin de comparer les résultats : fusion précoce, intermédiaire et tardive. Nous avons obtenu les meilleures performances avec l'algorithme Random Forest en fusion précoce. Afin d'introduire une certaine interprétabilité dans notre système de prédiction, nous avons procédé à une phase de sélection de paramètres pour effectuer des prédictions sur la base d'un ensemble restreint de paramètres. Les meilleurs résultats ont été obtenus avec un jeu de trois paramètres : le pourcentage de parole, l'appropriation lexicale au niveau de la phrase et la similarité sémantique. Nous avons amélioré les performances de prédiction et avons obtenu un coefficient de corrélation de Pearson de $r = 0{,}97$ et une MAE de 0,15.

D'un point de vue industriel, une réduction du temps et des ressources nécessaires pour enregistrer des apprenants L2 et obtenir une bonne approximation de leur compréhensibilité à un instant donné est primordiale. Après une analyse ayant pour objectif la réduction du nombre d'énoncés à traduire tout en permettant de bonnes performances, les résultats obtenus avec seulement six énoncés au lieu de 40 sont très prometteurs ($r = 0{,}94$, MAE = 0,13). Nous avons ensuite mis en place la même méthodologie de prédiction pour prédire les scores *a posteriori* des apprenants allemands. Les performances obtenues ($r = 0{,}98$, MAE = 0,18) nous confortent dans l'idée que notre protocole est pertinent lorsqu'il est appliqué à une autre paire de langue L1/français. Enfin, nous avons procédé à une analyse croisée entre deux L1 pour avoir une idée de l'indépendance vis-à-vis de la langue maternelle. En entraînant une Random Forest uniquement sur les données de CAF-jp, la prédiction de la compréhensibilité des apprenants de CAF-al se voit être, encore une fois, très satisfaisante ($r = 0{,}98$, MAE = 0,34). Ces résultats ouvrent sur le fait qu'il est possible d'entraîner un modèle sur une L1 spécifique et de prédire la compréhensibilité de la parole de manière satisfaisante d'apprenants d'une L1 différente. L'application, d'un point de vue industriel, se voit être très intéressante, car il suffirait d'un seul système de prédiction par langue cible pour pouvoir prédire les scores de compréhensibilité d'apprenants de L1 différentes.

Discussion et perspectives

Valorisation industrielle

Le protocole de prédiction de la compréhensibilité d'apprenants d'une langue étrangère a fait l'objet d'une demande provisoire de brevet (De Fino *et al.*, 2024). Le travail présenté dans cette thèse pourra être utilisé par la société Archean Technologies dans le cadre d'une application d'aide à l'apprentissage d'une langue étrangère. Pour de nouveaux apprenants japonais et allemands, un exercice de traduction pourrait être

proposé, conformément au protocole que nous avons défini, afin d'évaluer la compréhensibilité de la parole en utilisant nos systèmes appris respectivement sur de la parole japonaise ou allemande. Pour de nouvelles paires L1/français, une collaboration pourrait être entreprise avec des enseignants de FLE travaillant auprès de populations d'apprenants afin de constituer les énoncés spécifiques à traduire. L'évaluation de la compréhensibilité pourra ainsi s'opérer en utilisant notre système de prédiction entraîné sur notre corpus CAF-jp.

Vers l'évaluation de la compréhensibilité de la parole spontanée

Lors de notre état de l'art, nous avons vu qu'il n'y a pas une définition unique du concept de compréhensibilité de la parole. Celle-ci peut varier selon le type de tâche de production orale utilisée. Bien que nous ayons obtenu de bons scores de prédiction, nous devons garder à l'esprit que la compréhensibilité que nous prédisons reflète uniquement la compréhensibilité des apprenants dans le contexte de traduction, et non leur compréhensibilité globale.

Lors de ce travail de thèse, nous avons utilisé une tâche de production orale relativement cadrée afin de prédire la compréhensibilité de la parole. Il serait intéressant de faire évoluer notre recherche et de l'étendre à des tâches différentes. Nous pourrions commencer par des tâches de productions orales semi-spontanées, comme par exemple la description d'histoires imagées, la réponse à des questions ouvertes liées à des situations d'interaction de la vie quotidienne, et finir par des tâches de productions orales totalement spontanées. Un nouveau protocole d'évaluation subjective de la compréhensibilité devra être défini pour chaque type de tâche. Également, même si notre méthodologie de prédiction de la compréhensibilité pourra être utilisée quel que soit le type de productions orales traité, des énoncés plus spontanés vont poser des contraintes sur les paramètres et les modèles de prédiction utilisés, qu'il faudra adapter.

Un choix différent de paramètres

À l'issue de la sélection de paramètres, nous avons pu observer que seuls trois d'entre eux jouent un rôle clé pour prédire de manière satisfaisante la compréhensibilité de la parole non-native, à savoir le pourcentage de parole, le score d'appropriation lexicale et le score de similarité sémantique. Ces deux derniers sont particulièrement dépendants de notre tâche de collecte de données, car ils nécessitent d'avoir connaissance du message à censé être émis par le locuteur afin d'être mesurés. Dans le cas d'une parole plus spontanée, une telle référence quant au sens du message initial ne sera pas forcément disponible. De même, bien que nous ayons utilisé tous les paramètres implémentés durant ce travail de recherche pour effectuer nos premières prédictions, certains ne s'appliquent pas au type de productions relativement cadrées que nous avons utilisées. En effet, les énoncés étant assez courts, les paramètres issus de la complexité syntaxique et de la cohésion du discours ne paraissent pas pertinents.

De même, le vocabulaire utilisé dans les énoncés à traduire était relativement courant, ce qui n'apporte pas de grande variabilité sur les paramètres de la richesse lexicale. Nous pouvons observer des différences entre ces paramètres, mais elles sont seulement liées à la diversité des traductions réalisées. Lors de la définition d'un autre type de tâche de production orale, la compréhensibilité ne serait ainsi plus fortement liée aux paramètres d'appropriation lexicale et de similarité sémantique, mais à des paramètres issus de différents niveaux linguistiques, comme la fluence ou la cohésion du discours lors de productions complètement spontanées. Il serait également bénéfique d'enrichir notre jeu de paramètres et d'en implémenter de nouveaux. Par exemple, les paramètres de précision syntaxique pourraient se montrer particulièrement efficaces pour l'évaluation de la compréhensibilité de la parole non-native. Le fait de pouvoir mesurer la justesse grammaticale d'une phrase, en termes d'ordre des mots et d'accord entre le sujet et le verbe par exemple, paraît cohérent et approprié pour cette problématique. Il a d'ailleurs été démontré par Varonis et Gass (1982) que les phrases agrammaticales avaient un impact négatif sur la compréhensibilité.

Un choix différent de modèles de prédiction

De même, dans notre contexte de traduction, nous avons testé les performances de prédiction de deux algorithmes, la régression linéaire LASSO et la méthode basée sur les Random Forest. Bien que nous ayons obtenu de meilleures performances de prédiction en utilisant cette dernière, il est possible que ce ne soit pas la meilleure méthode pour d'autres jeux de données et de paramètres. Il faudrait ainsi également envisager de tester différents algorithmes de prédiction dans une situation de parole spontanée, comme par exemple la régression logistique (Cox, 1958) ou l'algorithme SVR (*Support Vector Regression* ; Cortes et Vapnik, 1995).

Explicabilité des prédictions

La sélection de paramètres que nous avons réalisée dans le chapitre 4 nous a permis d'identifier les paramètres jouant un rôle prépondérant dans la prédiction de la compréhensibilité dans le contexte d'une tâche de traduction. Notre démarche permet ainsi d'apporter une *interprétation* aux résultats. Cette interprétation nous donne d'ailleurs l'opportunité de fournir, d'un point de vue didactique, des retours quant aux aspects de la parole sur lesquels les apprenants doivent se concentrer pour améliorer leur compréhensibilité. Nos résultats interprétables ne sont toutefois pas *explicables*. L'explicabilité des prédictions permettrait de connaître les influences respectives de chacun des paramètres sur les prédictions individuelles. En outre, la compréhension détaillée de chaque prédiction nous donnerait la possibilité de fournir des retours personnalisés à chaque apprenant. Actuellement, pour une prédiction donnée, nous savons que la fluence, l'appropriation lexicale et la similarité sémantique ont mené à un score faible, par exemple (*interprétabilité*). Nous ne connaissons cependant pas l'influence exacte de chacun de ces paramètres pour parvenir à ce score (*explicabilité*).

Il serait alors intéressant de développer une analyse de résultats basée sur les valeurs SHAP (SHapley Additive exPlanations ; Lundberg et Lee, 2017). Les valeurs SHAP sont issues d'une approche de théorie des jeux et permettent d'expliquer les résultats de tout modèle d'apprentissage automatique. La mise en place d'une analyse des valeurs SHAP pourrait nous faire bénéficier d'informations sur les influences des paramètres sur chaque prédiction individuelle. Ainsi, pour un enregistrement audio donné, nous pourrions observer l'apport positif ou négatif de chaque paramètre et le quantifier. En reprenant exemple sur notre système, nous pourrions savoir si le score de similarité sémantique (ou tout autre paramètre) a contribué fortement à la diminution du score de compréhensibilité, et à quel point. Le fait de savoir quel paramètre a exercé une influence positive ou négative pour une prédiction donnée représente une information précieuse pour l'amélioration des compétences d'un apprenant, et un enjeu pour les plateformes d'aide à l'apprentissage des langues.

A

Apprenants du corpus CLIJAF

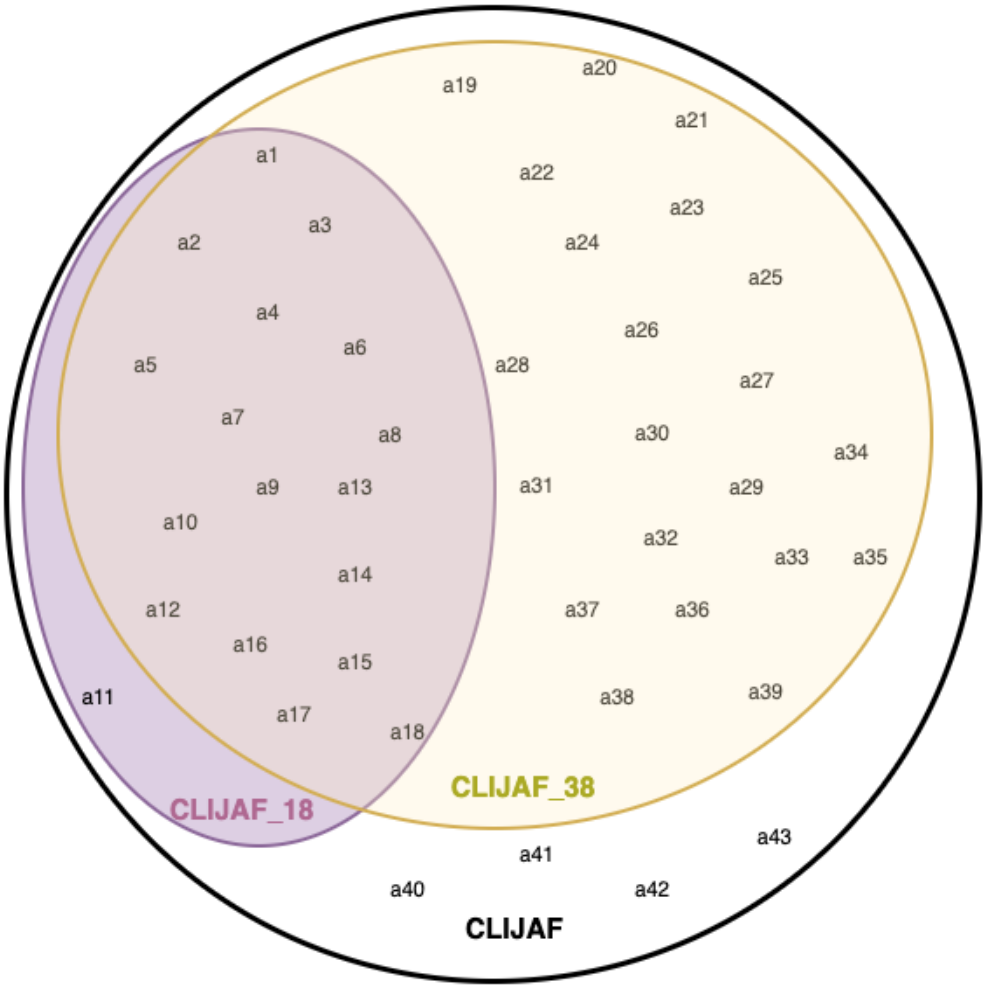


FIGURE A.1 – Répartition des apprenants du corpus CLIJAF en sous-ensemble CLIJAF_18 et CLIJAF_38. Les apprenants sont représentés par le texte aN, où N correspond au numéro de l'apprenant.

B

Interface d'enregistrement des productions orales d'apprenants L2

Deuxième étape d'enregistrement

La traduction attendue implique qu'il ne reste qu'un bloc non utilisé lors de cette étape, idéalement le bloc contenant l'erreur typique. Chaque apprenant peut néanmoins utiliser autant de blocs souhaités dans l'ordre supposé correct. Cette deuxième étape peut être complètement indépendante de la première, dite de *parole spontanée*, car il n'existe pas qu'une seule traduction correcte par énoncé, mais plusieurs. De ce fait, il est possible qu'un apprenant doive construire un énoncé totalement différent de celui qu'il aurait pu produire lors de la première étape. Par exemple, pour l'énoncé attendu « J'ai de l'argent » (voir figures B.1 et B.2), il est possible qu'un apprenant produise l'énoncé « je suis riche » lors de la première étape, dont le sens littéral reste tout de même cohérent avec la traduction attendue.

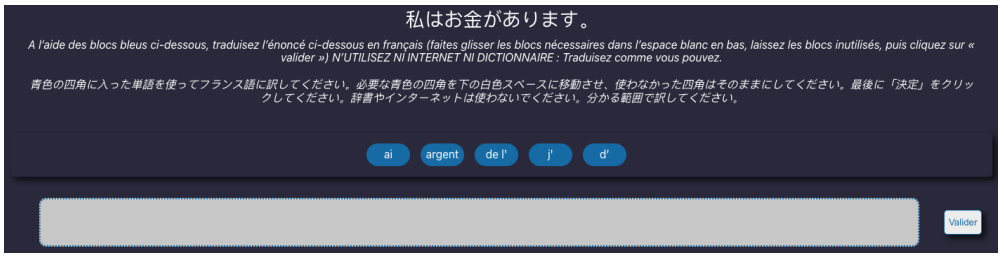


FIGURE B.1 – Interface d'enregistrement de l'apprenant - étape 2.1 : construction de la traduction à l'aide des blocs pour l'énoncé « *J'ai de l'argent* ». De haut en bas : l'instruction en français, l'instruction en japonais, les différents blocs et l'espace réservé pour construire la traduction avec les blocs. La difficulté ciblée est « *de l'* » et l'erreur courante est « *d'* » (« *J'ai d'argent* »).

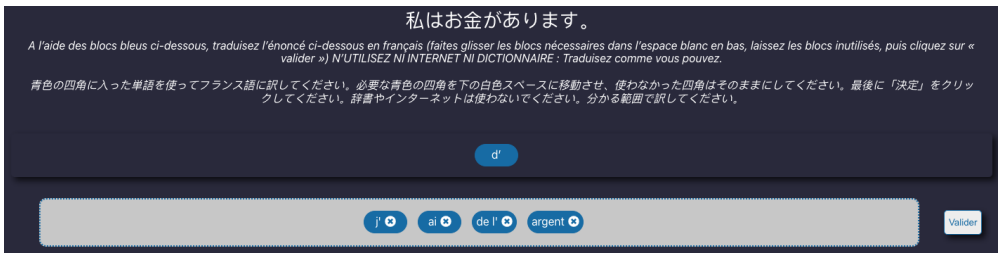


FIGURE B.2 – Interface d'enregistrement de l'apprenant - étape 2.2 : construction de la traduction à l'aide des blocs. Ici, la traduction « *J'ai de l'argent* » a été construite.

Troisième et quatrième étapes d'enregistrement

FIGURE B.3 – Interface d'enregistrement de l'apprenant - étape 3 : lecture et enregistrement de la traduction construite à l'étape précédente.

FIGURE B.4 – Interface d'enregistrement de l'apprenant - étape 4 : lecture et enregistrement de la traduction attendue. Ici, l'apprenant a construit « *Elle est japonais* », qui ne correspond pas à la traduction attendue « *Elle est japonaise* ».

Étant donné que certains énoncés contiennent des adverbes de temps, il n'est pas clairement indiqué à l'apprenant si la traduction qu'il vient de construire à la troisième étape est correcte ou non (par exemple, si « *demain je vais voir ma mère* » est la traduction attendue, « *je vais voir ma mère demain* » ne peut pas être qualifiée d'incorrecte, même si le mot « *demain* » n'est pas placé en début de phrase, comme attendu).

C

Matériel de traduction

TABLE C.1 – Énoncés du corpus de traduction pour les apprenants japonais.

Identifiant	Énoncé cible	Traduction japonaise	Difficulté ciblée	Erreur typique
304	j'ai fini le travail en une heure	私は1時間で仕事を終わらせました。	en	dans
305	j'aime la cuisine italienne	私はイタリア料理が好きです。	la cuisine italienne	le plat italien
306	j'ai vu un film	私は映画を見ました。	film	cinéma
307	samedi dernier je suis allé au café	この前の土曜日に私はカフェに行きました。	samedi	le samedi
308	j'aime le chocolat	私はチョコレートが好きです。	le chocolat	les chocolats
309	j'ai rendez-vous mercredi	私は水曜日に約束があります。	mercredi	à mercredi
310	j'ai acheté un manuel	私は教科書を買いました。	manuel	texte
311	j'ai de l'argent	私はお金があります。	de l'	d'
312	je n'ai pas d'argent	私はお金がありません。	d'	de l'
313	j'ai mangé des pâtes	私はパスタを食べました。	des	de
314	j'ai habité en France pendant un an	私はフランスに１年間住んでいました。	j'ai habité	j'habité
315	je vais en France	私はフランスに行きます。	en	à la
316	nous mangeons ensemble	私達は一緒に食べています。	mangeons	mangent
318	le professeur est dans la salle dix	先生は10番教室にいます。	salle	chambre
319	je bois beaucoup de café	私はたくさんコーヒーを飲みます。	de	du
320	j'aime le café	私はコーヒーが好きです。	le	du
321	je dois répondre avant la semaine prochaine	私は来週までに返事しなければなりません。	avant	jusqu'à
322	demain je vais rendre visite à ma mère	私は明日母を訪ねます。	rendre visite à	visiter
323	demain je vais voir ma mère	私は明日母に会います。	voir	rencontrer
324	le film était bien	映画は良かったです。	bien	bon
325	il porte un chapeau	彼は帽子を被っています。	porte	met
326	il y a deux assiettes sur la table	テーブルの上にお皿が２枚あります。	assiettes	plats
327	je viens avec toi	私は君と一緒に行きます。	viens	vais
328	j'ai un cours de français vendredi	私は金曜日にフランス語の授業があります。	un cours	une course
329	j'ai réussi l'examen	私は試験に受かりました。	réussi	passé
330	j'ai raté l'examen	私は試験に落ちました。	raté	tombé
331	j'ai lu une partie de ce livre	私はこの本の一部を読みました。	partie	part
332	j'ai apporté mon ordinateur	私はパソコンを持ってきました。	apporté	porté
333	hier il est allé à l'école primaire	昨日彼は小学校に行きました。	primaire	première
334	j'ai acheté un billet d'avion	私は航空券を買いました。	billet	ticket
335	j'aime beaucoup ce peintre	私はこの画家がとても好きです。	peintre	peinture
336	mes amis peuvent venir demain	私の友人たちは明日来ることができます。	peuvent	peut
337	elle est étudiante	彼女は学生です。	étudiante	étudiant
338	il est japonais	彼は日本人です。	japonais	japonaise
339	elle est japonaise	彼女は日本人です。	japonaise	japonais
340	ma mère est italienne	私の母はイタリア人です。	italienne	italien
341	je suis rentré à la maison	私は家に帰りました。	je suis	j'ai
342	j'aime la musique	私は音楽が好きです。	la	de la
343	je joue du piano	私はピアノを弾きます。	du	le
344	ce livre est intéressant	この本は面白いです。	intéressant	intéressé

TABLE C.2 – Énoncés du corpus de traduction pour les apprenants allemands.

Identifiant	Énoncé cible	Traduction allemande	Difficulté ciblée	Erreur typique
535	je n'ai pas le temps	ich habe keine Zeit	le	du
536	nous sommes allés au Portugal	wir sind nach Portugal gefahren	au	à le
537	il fait froid	es ist kalt	fait	est
538	il fait froid	ich fahre mit dem Zug	en	avec le
539	il boit du café au petit-déjeuner	er trinkt Kaffee zum Frühstuck	du	le
540	je rends visite à ma tante	ich besuche meine Tante	rends visite à	visite
541	son mari s'appelle Pierre	ihr Mann heißt Pierre	mari	homme
542	la boulangerie est dans cette rue	die Bäckerei ist in dieser Straße	rue	route
543	elle mange toujours trop de chocolat	sie isst immer zu viel Schokolade	de	le
544	il a plus de deux heures de retard	er hat mehr als zwei Stunden Verspätung	de	que
545	nous devons répondre à cette question	wir müssen diese Frage beantworten	répondre à	répondre
546	les touristes ont attendu le bus pendant des heures	die Touristen haben stundenlang auf den Bus gewartet	le	sur le
547	le cours va bientôt commencer	die Vorlesung fängt gleich an	le cours	la cour
548	ils regardent trop la télévision	sie schauen zuviel fern	regardent	voient
549	elle est mexicaine	sie ist Mexikanerin	mexicaine	mexicain
550	elle est allemande	sie ist Deutsche	allemande	allemand
551	son amie est riche	seine Freundin ist reich	son	s'
552	j'ai toujours été d'accord avec toi	ich bin immer einverstanden mit dir gewesen	j'ai	je suis
553	c'est un vieil homme	das ist ein alter Mann	vieil	vieux
554	tu achètes des pommes	du kaufst Äpfel	des pommes	pommes
555	cette année je vais en Suisse	dieses Jahr fahre ich in die Schweiz	en	à la
556	ils n'ont pas de problème	sie haben keine Probleme	de problème	des problèmes
557	elle n'est jamais à l'heure	sie ist niemals pünktlich	jamais	pas jamais
558	ce roman est moins intéressant que le premier	dieser Roman ist weniger interessant als der erste	que	comme
559	je veux vous montrer un livre	ich möchte Ihnen ein Buch zeigen	veux	vous veux
560	il sait nager depuis longtemps	er kann seit lange Zeit schwimmen	sait	peut
561	j'ai dix-neuf ans	ich bin 19	dix-neuf ans	dix-neuf
562	c'est un bel homme	das ist ein schöner Mann	bel	beau
563	il a connu sa femme à l'université	er hat seine Frau an der Universität kennengelernt	connu	su
564	voilà mes amis de Paris	hier sind meine Freunde aus Paris	voilà	ici
565	j'ai acheté une voiture chère	ich habe einen teuren Wagen gekauft	voiture chère	chère voiture
566	j'ai fait le ménage toute la journée	ich habe den ganzen Tag geputzt	toute la journée	tout le jour
567	le film était bien	der Film war gut	bien	bon
568	le cours est dans la salle dix	die Vorlesung findet im Raum zehn statt	salle	chambre
569	j'ai acheté un billet de train	ich habe einen Bahn ticket gekauft	billet	ticket
570	j'aime beaucoup cette peinture	ich mag sehr gern dieses Bild	peinture	image
571	j'ai amené les enfants à l'école	ich habe die Kinder zur Schule gebracht	amené	apporté
572	j'ai obtenu mon diplôme	ich habe mein Diplom erhalten	obtenu	reçu
573	je t'écoute	ich höre dir zu	écoute	entends
574	ils ont trois enfants	sie haben drei Kinder	ont	sont

D

Description du fonctionnement de l'algorithme de Random Forest

La random forest représente une forêt composée de plusieurs arbres de décision. Un arbre de décision est un graphe non orienté, acyclique et connexe. Il contient différents nœuds, dont un nœud racine, un ou des nœuds internes et un ou des nœuds terminaux (également appelé *feuilles*). Les nœuds internes représentent les différentes séquences de décision à prendre, tandis que les feuilles représentent les prédictions (voir figure D.1).

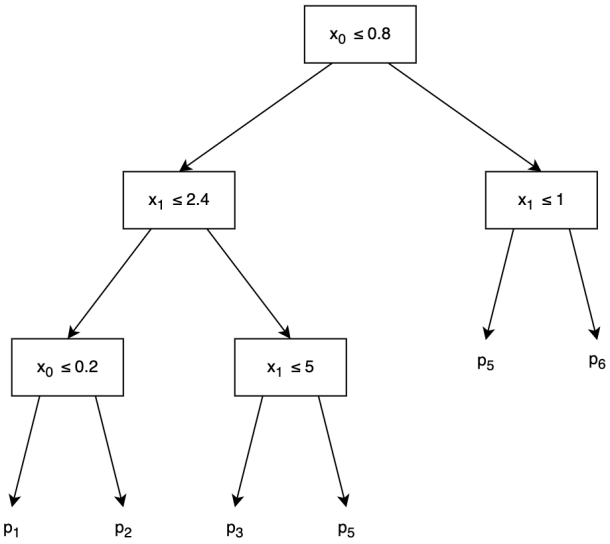


FIGURE D.1 – Exemple d'arbre de décision où x_0 et x_1 représentent deux paramètres et p_1, p_2, p_3, p_4, p_5 et p_6 les six prédictions différentes que peut renvoyer l'arbre.

Chaque arbre appartenant à la forêt est construit à partir d'un sous-ensemble aléatoire de données et de caractéristiques. Ces deux sous-ensembles sont légèrement différents entre chaque arbre, ce qui permet d'introduire une variabilité et une diversité des arbres au sein de la forêt. Une fois la forêt complètement construite, les arbres sont utilisés collectivement pour effectuer des prédictions. Dans le cas de la classification, les prédictions des différents arbres sont soumises à un vote majoritaire pour obtenir la prédiction finale. Dans le cas de la régression, cas dans lequel nous nous trouvons actuellement, les prédictions de chaque arbre sont moyennées afin d'obtenir une prédiction finale (voir figure D.2).

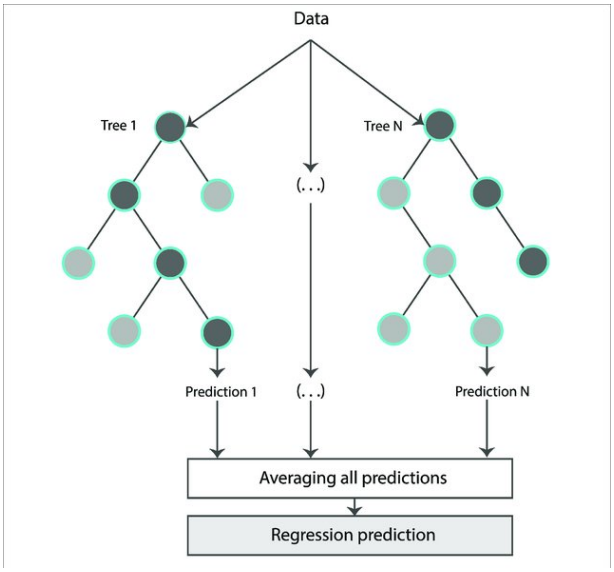


FIGURE D.2 – Exemple de random forest (Bikia *et al.*, 2021).

Bibliographie

Abdi, H. *et al.* (2007). Bonferroni and šidák corrections for multiple comparisons. *Encyclopedia of measurement and statistics*, 3(01):2007.

Abeillé, A., Clément, L. et Toussenel, F. (2003). Building a treebank for french. *Treebanks : Building and using parsed corpora*, pages 165–187.

André-Obrecht, R. (1988). A new statistical approach for the automatic segmentation of continuous speech signals. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(1):29–40.

Banerjee, S. et Lavie, A. (2005). METEOR : An automatic metric for MT evaluation with improved correlation with human judgments. *In* Goldstein, J., Lavie, A., Lin, C.-Y. et Voss, C., éditeurs : *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Baroni, M., Bernardini, S., Ferraresi, A. et Zanchetta, E. (2009). The wacky wide web : a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43:209–226.

Barrett, W. L., Gluckman, J. L., Wilson, K. M. et Gleich, L. L. (2004). A comparison of treatments of squamous cell carcinoma of the base of tongue : surgical resection combined with external radiation therapy, external radiation therapy alone, and external radiation therapy combined with interstitial radiation. *Brachytherapy*, 3(4):240–245.

Beacco, J. C., Bouquet, S. et Porquier, R. (2004). *Niveau B2 pour le français*. Didier, Paris.

Bergeron, A. et Trofimovich, P. (2017). Linguistic dimensions of accentedness and comprehensibility : Exploring task and listener effects in second language french. *Foreign Language Annals*, 50(3):547–566.

Bikia, V., Rovas, G., Pagoulatou, S. et Stergiopulos, N. (2021). Determination of aortic characteristic impedance and total arterial compliance from regional pulse wave velocities using machine learning : An in-silico study. *Frontiers in Bioengineering and Biotechnology*, 9.

Blache, P. (2010). Un modèle de caractérisation de la complexité syntaxique. *In Actes de la 17e conférence sur le Traitement Automatique des Langues Naturelles*, pages 81–90, Montréal, Canada. ATALA.

Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8:3–62.

Bongaerts, T., Mennen, S. et Slik, F. v. d. (2000). Authenticity of pronunciation in naturalistic second language acquisition : The case of very advanced late learners of dutch as a second language. *Studia linguistica*, 54(2):298–308.

Bonvin, A. et Lambelet, A. (2019). Exploration empirique de la richesse lexicale : la perception humaine. *Linguistik Online*, 100(77):65–94.

Bradlow, A. R. et Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, 106(2):707–729.

Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.

Breiman, L. (2002). Manual on setting up, using, and understanding random forests v3. 1. *Statistics Department University of California Berkeley, CA, USA*, 1(58):3–42.

Chen, L.-Y. et Jang, J.-S. R. (2012). Improvement in automatic pronunciation scoring using additional basic scores and learning to rank. *In Thirteenth Annual Conference of the International Speech Communication Association*.

Conneau, A., Lample, G., Rinott, R., Williams, A., Bowman, S. R., Schwenk, H. et Stoyanov, V. (2018). Xnli : Evaluating cross-lingual sentence representations. *arXiv preprint arXiv :1809.05053*.

Conseil De l'Europe (2001). *Cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer (CECR)*. Didier, Paris.

Cortes, C. et Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3): 273–297.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society : Series B (Methodological)*, 20(2):215–232.

Crowther, D., Trofimovich, P. et Isaacs, T. (2016). Linguistic dimensions of second language accent and comprehensibility : Nonnative listeners' perspectives. *Journal of Second Language Pronunciation*, 2(2):160–182.

Crowther, D., Trofimovich, P., Isaacs, T. et Saito, K. (2015a). Does a speaking task affect second language comprehensibility ? *The Modern Language Journal*, 99(1):80–95.

Crowther, D., Trofimovich, P., Saito, K. et Isaacs, T. (2015b). Second language comprehensibility revisited : Investigating the effects of learner background. *TESOL quarterly*, 49(4):814–837.

Crowther, D., Trofimovich, P., Saito, K. et Isaacs, T. (2018). Linguistic dimensions of l2 accentedness and comprehensibility vary across speaking tasks. *Studies in Second Language Acquisition*, 40(2):443–457.

Cruickshanks, K. J., Wiley, T. L., Tweed, T. S., Klein, B. E., Klein, R., Mares-Perlman, J. A. et Nondahl, D. M. (1998). Prevalence of Hearing Loss in Older Adults in Beaver Dam, Wisconsin : The Epidemiology of Hearing Loss Study. *American Journal of Epidemiology*, 148(9):879–886.

Cucchiarini, C., Strik, H. et 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*, 107:989–99.

Daller, M., Hout, R. et Treffers-Daller, J. (2003). Lexical richness in the spontaneous speech of bilinguals. *Applied Linguistics*, 24:197–222.

De Fino, V., Fontan, L., Ferrané, I. et Pinquier, J. (2024). Procédé de mesure de la compréhensibilité d'échantillons de parole. Dépôt de brevet n°FR2401263. Institut National de la Propriété Intellectuelle.

De Fino, V., Fontan, L., Pinquier, J., Barcat, C., Ferrané, I. et Detey, S. (2022a). Mesures automatiques de parole non-native : exploration pilote d'un corpus d'apprenants japonais de français et différenciation de niveaux. *In 34èmes Journées d'Études sur la Parole (JEP 2022)*, pages 1–10, Noirmoutier, France.

De Fino, V., Fontan, L., Pinquier, J., Ferrané, I. et Detey, S. (2022b). Prediction of L2 speech proficiency based on multi-level linguistic features. *In 23rd INTERSPEECH Conference : Human and Humanizing Speech Technology (INTERSPEECH 2022)*, Incheon, South Korea. The Acoustical Society of Korea.

Derwing, T. (2003). What do esl students say about their accents ? *Canadian Modern Language Review*, 59(4):547–567.

Derwing, T. M. et Munro, M. J. (1997). Accent, intelligibility, and comprehensibility : Evidence from four l1s. *Studies in second language acquisition*, 19(1):1–16.

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

Derwing, T. M. et Munro, M. J. (2015). *Pronunciation Fundamentals : Evidence-based perspectives for L2 teaching and research*, volume 42. John Benjamins, Amsterdam.

Derwing, T. M., Munro, M. J. et Thomson, R. I. (2008). A longitudinal study of esl learners' fluency and comprehensibility development. *Applied linguistics*, 29(3):359–380.

Derwing, T. M., Munro, M. J., Thomson, R. I. et Rossiter, M. J. (2009). The relationship between l1 fluency and l2 fluency development. *Studies in Second Language Acquisition*, 31(4):533–557.

Derwing, T. M., Rossiter, M. J., Munro, M. J. et Thomson, R. I. (2004). Second language fluency : Judgments on different tasks. *Language learning*, 54(4):655–679.

Detey, S., Fontan, L., Le Coz, M. et Jmel, S. (2020). Computer-assisted assessment of phonetic fluency in a second language : a longitudinal study of Japanese learners of French. *Speech Communication*, 125:69–79.

Detey, S., Fontan, L. et Pellegrini, T. (2016). Traitement de la prononciation en langue étrangère : approches didactiques, méthodes automatiques et enjeux pour l'apprentissage. *Revue TAL*, 57(3):15–39.

Detey, S. et Kawaguchi, Y. (2008). Interphonologie du Français Contemporain (IPFC) : récolte automatisée des données et apprenants japonais. *In Journées PFC : Phonologie du français contemporain : variation, interfaces, cognition*, Paris : MSH.

Detey, S., Le Coz, M., Fontan, L., Barcat, C., Kawaguchi, Y., Akihiro, H., Sugiyama, K. et Kondo, N. (2018). Annotations minimales multi-niveaux d'un corpus de parole spontanée d'apprenants japonais de FLE et traitement automatique : perspectives didactiques. *In FLORAL-IPFC2018 : contact de langues et (inter)phonologie de corpus*, Paris : MSH.

Devlin, J., Chang, M.-W., Lee, K. et Toutanova, K. (2018). Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*.

Eskenazi, M. (2009). An overview of spoken language technology for education. *Speech communication*, 51(10):832–844.

Estève, Y., Bazillon, T., Antoine, J.-Y., Béchet, F. et Farinas, J. (2010). The EPAC corpus : Manual and automatic annotations of conversational speech in French broadcast news. *In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Farinas, J. et Pellegrino, F. (2001). Automatic rhythm modeling for language identification. *In Proc. 7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, pages 2539–2542.

Fauconnier, J.-P. (2015). French word embeddings.

Flege, J. E. (1988). Factors affecting degree of perceived foreign accent in english sentences. *The Journal of the Acoustical Society of America*, 84(1):70–79.

Fontan, L. (2012). *De la mesure de l'intelligibilité à l'évaluation de la compréhension de la parole pathologique en situation de communication*. Thèse de doctorat, Université Toulouse le Mirail-Toulouse II.

Fontan, L., Kim, S., De Fino, V. et Detey, S. (2022). Predicting speech fluency in children using automatic acoustic features. *In Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2022)*, Chiang Mai, Thailand.

Fontan, L., Le Coz, M. et Alazard, C. (2020). Using the forward-backward divergence segmentation algorithm and a neural network to predict L2 speech fluency. *In Proc. 10th International Conference on Speech Prosody 2020*, pages 925–929.

Fontan, L., Le Coz, M. et Detey, S. (2018). Automatically Measuring L2 Speech Fluency without the Need of ASR : A Proof-of-concept Study with Japanese Learners of French. *In Proc. Interspeech 2018*, pages 2544–2548.

Foster, P. et Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second language acquisition*, 18(3):299–323.

Galliano, S., Geoffrois, E., Mostefa, D., Choukri, K., Bonastre, J.-F. et Gravier, G. (2005). The ester phase ii evaluation campaign for the rich transcription of french broadcast news. *In Interspeech*, pages 1149–1152.

Gardner, R. C. (1985). Social psychology and second language learning : The role of attitudes and motivation. *(No Title)*.

Gardner, R. C. (2010). *Motivation and second language acquisition : The socio-educational model*, volume 10. Peter Lang.

Gass, S. et Varonis, E. M. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Language learning*, 34(1):65–87.

Gelin, L. (2022). *Reconnaissance automatique de la parole d'enfants apprenant×e×s lecteur×ice×s en salle de classe : modélisation acoustique de phonèmes*. Theses, Université Paul Sabatier - Toulouse III.

Guiraud, P. (1959). *Problèmes et méthodes de la statistique linguistique*, volume 2. D. Reidel, Dodrecht.

Hahn, L. D. (2004). Primary stress and intelligibility : Research to motivate the teaching of suprasegmentals. *TESOL quarterly*, 38(2):201–223.

Heba, A. (2021). *Reconnaissance automatique de la parole à large vocabulaire : des approches hybrides aux approches End-to-End*. Thèse de doctorat, Université Toulouse 3 Paul Sabatier.

Hu, W., Qian, Y. et Soong, F. K. (2013). A new dnn-based high quality pronunciation evaluation for computer-aided language learning (call). *In Interspeech*, pages 1886–1890.

Isaacs, T. et Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of l2 pronunciation : Revisiting research conventions. *Language Assessment Quarterly*, 10(2):135–159.

Isaacs, T. et Thomson, R. I. (2020). Reactions to second language speech : Influences of discrete speech characteristics, rater experience, and speaker first language background. *Journal of Second Language Pronunciation*, 6(3):402–429.

Isaacs, T. et Trofimovich, P. (2012). Deconstructing comprehensibility : Identifying the linguistic influences on listeners' l2 comprehensibility ratings. *Studies in Second Language Acquisition*, 34(3):475–505.

Isaacs, T., Trofimovich, P. et Foote, J. A. (2018). Developing a user-oriented second language comprehensibility scale for english-medium universities. *Language Testing*, 35(2):193–216.

Jenkins, J. (2000). *The phonology of English as an international language*. Oxford university press.

Jones, J. et Hunter, D. (1995). Qualitative research : consensus methods for medical and health services research. *Bmj*, 311(7001):376–380.

Kamal Eddine, M., Tixier, A. et Vazirgiannis, M. (2021). BARThez : a skilled pretrained French sequence-to-sequence model. *In* Moens, M.-F., Huang, X., Specia, L. et Yih, S. W.-t., éditeurs : *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9369–9390, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kearns, J. (2014). Librivox : Free public domain audiobooks. *Reference Reviews*, 28(1):7–8.

Kennedy, S. et Trofimovich, P. (2008). Intelligibility, comprehensibility, and accentedness of l2 speech : The role of listener experience and semantic context. *Canadian Modern Language Review*, 64(3):459–489.

Kruskal, W. H. et Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621.

Kursa, M. B. et Rudnicki, W. R. (2010). Feature selection with the boruta package. *Journal of statistical software*, 36:1–13.

Laborde, V., Pellegrini, T., Fontan, L., Mauclair, J., Sahraoui, H. et Farinas, J. (2016). Pronunciation assessment of japanese learners of french with gop scores and phonetic information. *In Annual conference Interspeech (INTERSPEECH 2016)*, pages 2686–2690, San Francisco, CA, US. International Speech Communication Association (ISCA).

Lahuerta Martínez, A. C. (2018). Analysis of syntactic complexity in secondary education EFL writers at different proficiency levels. *Assessing Writing*, 35:1–11.

Lamel, L. F., Gauvain, J.-L., Eskénazi, M. *et al.* (1991). Bref, a large vocabulary spoken corpus for french1. *training*, 22(28):50.

Laufer, B. et Nation, P. (1995). Vocabulary size and use : Lexical richness in L2 written production. *Applied linguistics*, 16(3):307–322.

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L. et Schwab, D. (2020). Flaubert : Unsupervised language model pre-training for french. *In Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Lee, I. A. et Preacher, K. J. (2013). Calculation for the test of the difference between two dependent correlations with one variable in common [computer software].

Levis, J. M. (2006). Pronunciation and the assessment of spoken language. *In Spoken English, TESOL and applied linguistics : Challenges for theory and practice*, pages 245–270. Springer.

Lindqvist, C., Bardel, C. et Gudmundson, A. (2011). Lexical richness in the advanced learner's oral production of french and italian l2. *IRAL*, 49(3):221–240.

Lloyd, S. (1982). Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137.

Loewen, S., Crowther, D., Isbell, D. R., Kim, K. M., Maloney, J., Miller, Z. F. et Rawal, H. (2019). Mobile-assisted language learning : A duolingo case study. *ReCALL*, 31(3):293–311.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4):474–496.

Lundberg, S. M. et Lee, S.-I. (2017). A unified approach to interpreting model predictions. *In* Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. et Garnett, R., éditeurs : *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Lyster, R. et Saito, K. (2010). Oral feedback in classroom sla : A meta-analysis. *Studies in second language acquisition*, 32(2):265–302.

Mann, H. B. et Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60.

Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., de la Clergerie, É. V., Seddah, D. et Sagot, B. (2020). Camembert : a tasty french language model. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

McMillan, S. S., King, M. et Tully, M. P. (2016). How to use the nominal group and delphi techniques. *International journal of clinical pharmacy*, 38:655–662.

Melamud, O., Levy, O. et Dagan, I. (2015). A simple word embedding model for lexical substitution. *In Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 1–7.

Mikolov, T., Le, Q. V. et Sutskever, I. (2013a). Exploiting similarities among languages for machine translation. *arXiv preprint arXiv :1309.4168*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. et Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Muñoz, C. (2014). Exploring young learners' foreign language learning awareness. *Language awareness*, 23(1-2):24–40.

Munro, K. et Stone, M. (2021). The challenges of facemasks for people with hearing loss. *ENT & audiology news*.

Munro, M. J. et Derwing, T. M. (1995a). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language learning*, 45(1):73–97.

Munro, M. J. et Derwing, T. M. (1995b). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and speech*, 38(3):289–306.

Munro, M. J. et Derwing, T. M. (1999). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language learning*, 49:285–310.

Munro, M. J. et Derwing, T. M. (2011). The foundations of accent and intelligibility in pronunciation research. *Language Teaching*, 44(3):316–327.

Navigli, R., Jurgens, D. et Vannella, D. (2013). Semeval-2013 task 12 : Multilingual word sense disambiguation. *In Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2 : Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 222–231.

New, B., Pallier, C. et Ferrand, L. (2005). Manuel de Lexique 3. *Behavior Research Methods, Instruments, & Computers*, 36(3):516–524.

Ovtcharov, V., Cobb, T. et Halter, R. (2006). La richesse lexicale des productions orales : Mesure fiable du niveau de compétence langagière. *Canadian Modern Language Review-revue Canadienne Des Langues Vivantes*, 63:107–125.

O'Brien, M. G., Derwing, T. M., Cucchiarini, C., Hardison, D. M., Mixdorff, H., Thomson, R. I., Strik, H., Levis, J. M., Munro, M. J., Foote, J. A. *et al.* (2018). Directions for the future of technology in pronunciation research and teaching. *Journal of Second Language Pronunciation*, 4(2):182–207.

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D. et Le, Q. V. (2019). Specaugment : A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv :1904.08779*.

Pennington, M. C. (2021). Teaching pronunciation : The state of the art 2021. *RELC Journal*, 52(1):3–21.

Pommée, T., Balaguer, M., Mauclair, J., Pinquier, J. et Woisard, V. (2022). Intelligibility and comprehensibility : A delphi consensus study. *International Journal of Language & Communication Disorders*, 57(1):21–41.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G. et Vesely, K. (2011). The kaldi speech recognition toolkit. *In IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society. IEEE Catalog No. : CFP11SRW-USB.

Prettenhofer, P. et Stein, B. (2010). Cross-language text classification using structural correspondence learning. *In Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 1118–1127.

Racine, I., Zay, F., Detey, S. et Kawaguchi, Y. (2012). Des atouts d'un corpus multitâches pour l'étude de la phonologie en L2 : l'exemple du projet "Interphonologie du français contemporain" (IPFC). *In* Kamber, A. et Skupien, C., éditeurs : *Recherches récentes en FLE*, pages 1–19. Peter Lang, Bern.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C. et Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision.

Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J., Chou, J.-C., Yeh, S.-L., Fu, S.-W., Liao, C.-F., Rastorgueva, E., Grondin, F., Aris, W., Na, H., Gao, Y., Mori, R. D. et Bengio, Y. (2021). SpeechBrain : A general-purpose speech toolkit. arXiv :2106.04624.

Reimers, N. et Gurevych, I. (2019). Sentence-bert : Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv :1908.10084*.

Robinson, P. (2005). Cognitive complexity and task sequencing : Studies in a componential framework for second language task design. *International Review of Applied Linguistics in Language Teaching*, 43(1):1–32.

Rossi-Gensane, N. (2010). Oralité, syntaxe et discours. *In* Detey, S., Durand, J., Laks, B. et Lyche, C., éditeurs : *Les variétés du français parlé dans l'espace francophone : ressources pour l'enseignement*, pages 83–106. Ophrys, Paris.

Roze, C., Danlos, L. et Muller, P. (2012). LEXCONN : a French lexicon of discourse connectives. *Discours - Revue de linguistique, psycholinguistique et informatique*.

Saito, K. (2015). Experience effects on the development of late second language learners' oral proficiency. *Language Learning*, 65(3):563–595.

Saito, K. et Akiyama, Y. (2017). Linguistic correlates of comprehensibility in second language japanese speech. *Journal of Second Language Pronunciation*, 3(2):199–217.

Saito, K., Macmillan, K., Kachlicka, M., Kunihara, T. et Minematsu, N. (2023). Automated assessment of second language comprehensibility : Review, training, validation, and generalization studies. *Studies in Second Language Acquisition*, 45(1):234–263.

Saito, K. et Shintani, N. (2016). Do native speakers of north american and singapore english differentially perceive comprehensibility in second language speech ? *Tesol Quarterly*, 50(2):421–446.

Saito, K., Tran, M., Suzukida, Y., Sun, H., Magne, V. et Ilkan, M. (2019). How do second language listeners perceive the comprehensibility of foreign-accented speech ? : Roles of first language profiles, second language proficiency, age, experience, familiarity, and metacognition. *Studies in Second Language Acquisition*, 41(5):1133–1149.

Saito, K., Trofimovich, P. et Isaacs, T. (2016a). Second language speech production : Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics*, 37(2):217–240.

Saito, K., Trofimovich, P. et Isaacs, T. (2017). Using listener judgments to investigate linguistic influences on l2 comprehensibility and accentedness : A validation and generalization study. *Applied Linguistics*, 38(4):439–462.

Saito, K., Webb, S., Trofimovich, P. et Isaacs, T. (2016b). Lexical correlates of comprehensibility versus accentedness in second language speech. *Bilingualism : Language and Cognition*, 19(3):597–609.

Saito, K., Webb, S., Trofimovich, P. et Isaacs, T. (2016c). Lexical profiles of comprehensible second language speech : The role of appropriateness, fluency, variation, sophistication, abstractness, and sense relations. *Studies in Second Language Acquisition*, 38(4):677–701.

Segonne, V., Candito, M. et Crabbé, B. (2019). Using wiktionary as a resource for wsd : the case of french verbs. *In Proceedings of the 13th International Conference on Computational Semantics-Long Papers*, pages 259–270. Association for Computational Linguistics.

Skehan, P. (2009). Modelling second language performance : Integrating complexity, accuracy, fluency, and lexis. *Applied linguistics*, 30(4):510–532.

Suárez, P. J. O., Sagot, B. et Romary, L. (2019). Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. *In 7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache.

Suzuki, S. et 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.

Suzukida, Y. et Saito, K. (2021). Which segmental features matter for successful l2 comprehensibility ? revisiting and generalizing the pedagogical value of the functional load principle. *Language Teaching Research*, 25(3):431–450.

Tarone, E. (1983). On the variability of interlanguage systems. *Applied linguistics*, 4(2):142–164.

Tauroza, S. et Luk, J. (1997). Accent and second language listening comprehension. *RELC journal*, 28(1):54–71.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B : Statistical Methodology*, 58(1):267–288.

Trofimovich, P. et Isaacs, T. (2012). Disentangling accent from comprehensibility. *Bilingualism : Language and Cognition*, 15(4):905–916.

Trofimovich, P., Isaacs, T., Kennedy, S., Saito, K. et Crowther, D. (2016). Flawed self-assessment : Investigating self-and other-perception of second language speech. *Bilingualism : Language and Cognition*, 19(1):122–140.

van Hout, R. et Vermeer, A. (2007). Comparing measures of lexical richness. *In* Daller, H., Milton, J. et Treffers-Daller, J., éditeurs : *Modelling and assessing vocabulary knowledge*, pages 93–116. Cambridge University Press, Cambridge.

Varonis, E. M. et Gass, S. (1982). The comprehensibility of non-native speech. *Studies in second language acquisition*, 4(2):114–136.

Vedantam, R., Zitnick, C. L. et Parikh, D. (2015). Cider : Consensus-based image description evaluation.

Wenzek, G., Lachaux, M.-A., Conneau, A., Chaudhary, V., Guzmán, F., Joulin, A. et Grave, E. (2019). Ccnet : Extracting high quality monolingual datasets from web crawl data. *arXiv preprint arXiv :1911.00359*.

Wetzel, M., Zufferey, S. et Gygax, P. (2020). Second language acquisition and the mastery of discourse connectives : Assessing the factors that hinder l2-learners from mastering french connectives. *Languages*, 5(33):35.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. biom. bull., 1, 80–83.

Williams, A., Nangia, N. et Bowman, S. R. (2017). A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv :1704.05426*.

Winke, P., Gass, S. et Myford, C. (2013). Raters' l2 background as a potential source of bias in rating oral performance. *Language Testing*, 30(2):231–252.

Witt, S. M. (2012). Automatic error detection in pronunciation training : Where we are and where we need to go. *In International Symposium on automatic detection on errors in pronunciation training*, volume 1.

Woisard, V., Espesser, R., Ghio, A. et Duez, D. (2013). De l'intelligibilité à la compréhensibilité de la parole, quelles mesures en pratique clinique ? *Revue de Laryngologie Otologie Rhinologie*, 134(1):27–33.

Wold, S., Ruhe, A., Wold, H. et Dunn, Iii, W. (1984). The collinearity problem in linear regression. the partial least squares (pls) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, 5(3):735–743.

Xue, W., Cucchiarini, C., van Hout, R. et Strik, H. (2019). Acoustic correlates of speech intelligibility : the usability of the eGeMAPS feature set for atypical speech. *In Proc. 8th ISCA Workshop on Speech and Language Technology in Education (SLaTE 2019)*, pages 48–52.

Yang, Y., Zhang, Y., Tar, C. et Baldridge, J. (2019). Paws-x : A cross-lingual adversarial dataset for paraphrase identification. *arXiv preprint arXiv :1908.11828*.

Yuan, F. et Ellis, R. (2003). The effects of pre-task planning and on-line planning on fluency, complexity and accuracy in l2 monologic oral production. *Applied linguistics*, 24(1):1–27.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. et Artzi, Y. (2019a). Bertscore : Evaluating text generation with bert. *arXiv preprint arXiv :1904.09675*.

Zhang, Y., Baldridge, J. et He, L. (2019b). Paws : Paraphrase adversaries from word scrambling. *arXiv preprint arXiv :1904.01130*.

Zhou, W., Ge, T., Xu, K., Wei, F. et Zhou, M. (2019). Bert-based lexical substitution. *In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3368–3373.

Résumé

Se faire comprendre en situation de communication, voire d'interaction orale, est essentiel au quotidien. La compréhensibilité est ainsi devenue un objectif important dans le domaine de l'apprentissage des langues, plus encore que d'avoir une parole sans accent étranger, proche d'un locuteur natif. Cependant, les enseignants et apprenants d'une langue étrangère (L2) ne disposent pas d'outils automatiques leur permettant d'évaluer de manière objective la compréhensibilité des productions orales.

La compréhensibilité représente un concept linguistique influencé par des dimensions comme la phonologie/phonétique, la fluence, le lexique, la syntaxe et le discours. En plus de ces dimensions, elle peut également être influencée par le profil d'un apprenant (sa langue maternelle, ou L1, plus ou moins proche de la langue cible), le profil d'un auditeur (familiarisé ou non avec l'accent de l'apprenant) et la tâche de production orale pour les mettre en situation et collecter la parole des apprenants.

Dans nos travaux de recherche nous nous sommes intéressés à la description de ces différentes dimensions dans la littérature. Nous avons ensuite implémenté différents paramètres considérés comme ayant une influence sur la compréhensibilité de la parole. Une première étape a été de valider leur adéquation lors d'une tâche de prédiction du niveau CECRL (Cadre Européen Commun de Référence pour les Langues) des apprenants du corpus CLIJAF. En se fondant sur ces paramètres linguistiques multi-niveaux, nous avons pu aborder la contribution principale de ce travail de thèse en proposant une méthode permettant de mesurer de manière automatique la compréhensibilité des apprenants.

Afin d'évaluer la compréhensibilité, nous avons réalisé deux corpus : CAF-jp (Compréhensibilité d'Apprenants du Français - Japonais) et CAF-al (Compréhensibilité d'Apprenants du Français - Allemands). Ces corpus contiennent respectivement des productions orales de 40 apprenants japonais et 9 apprenants allemands de français. La mise en place d'un protocole de collecte a permis de collecter des productions orales. Ce protocole est basé sur une tâche de traduction orale, en L2, d'énoncés écrits en L1. Les énoncés ont été spécifiquement construits par des experts de FLE (Français Langue Étrangère) afin de contenir des difficultés typiques de traduction propres à chaque paire L1/français. Une fois la collecte des données effectuée, nous avons créé un protocole d'annotations nous permettant d'obtenir des évaluations subjectives de la compréhensibilité de la parole. Nous avons mené une campagne d'annotation auprès de 80 Français natifs et avons collecté 3920 scores de compréhensibilité, dont la moitié correspondent à la compréhensibilité *a priori* (compréhensibilité perçue) et l'autre moitié à la compréhensibilité *a posteriori* (compréhensibilité du sens du message véhiculé après prise en compte du réel sens du message à véhiculer).

Afin de prédire automatiquement la compréhensibilité de la parole des apprenants, nous mettons en place une phase d'extraction de paramètres sur les productions

orales. Ces paramètres sont d'ordre phonético-phonologique, lexical, syntaxique, discursif et sémantique. Nous obtenons d'excellents résultats de prédiction, aussi bien pour le corpus CAF-jp (r=0,97, MAE=0,15) que pour le corpus CAF-al (r=0,98, MAE=0,18), en utilisant l'algorithme *Random Forest*, une stratégie de fusion précoce et une validation croisée imbriquée de type *leave-one-out*.

De plus, en entraînant un modèle sur la totalité des données du corpus CAF-jp et en testant sur les données du corpus CAF-al, nous obtenons également de bonnes performances (r=0,98, MAE=0,34), montrant ainsi la généricité de notre approche.

Nos différents résultats montrent que notre méthodologie de prédiction de la compréhensibilité est tout à fait adaptée à l'évaluation de l'apprentissage du français L2, et pourrait même être appliquée à d'autres paires de langues L1/L2.

Mots-clés: compréhensibilité de la parole, apprentissage des langues, paramètres linguistiques, prédiction automatique

Abstract

Being understood in communication situations, even in oral interactions, is essential in everyday life. Comprehensibility has thus become a significant goal in the field of language learning, even more than having a speech without a foreign accent, close to that of a native speaker. However, teachers and learners of a foreign language (L2) lack automatic tools to objectively assess the comprehensibility of oral productions.

Comprehensibility is a linguistic concept influenced by dimensions such as phonology/phonetics, fluency, lexis, syntax, and discourse. In addition to these dimensions, it can also be influenced by a learner's profile (native language, or L1, more or less similar to the target language), a listener's profile (familiarized or not with the learner's accent), and the oral production task to contextualize and collect learners' speech.

In our research, we focused on describing these different dimensions in the literature. We then implemented various features considered to have an influence on the speech comprehensibility. A first step was to validate their adequacy during a task predicting the CEFRL (Common European Framework of Reference for Languages) level of learners in the CLIJAF corpus. Based on these multi-level linguistic features, we approached the main contribution of this thesis by proposing a method to automatically measure learners' comprehensibility.

To assess comprehensibility, we created two corpora : CAF-jp (Comprehensibility of French Learners - Japanese) and CAF-al (Comprehensibility of French Learners - Germans). These corpora respectively contain oral productions from 40 Japanese learners and 9 German learners of French. The implementation of a collection protocol allowed us to gather oral productions. This protocol is based on an oral translation task, in L2, of statements written in L1. The statements were specifically constructed by experts in FFL (French as a Foreign Language) to contain typical translation difficulties for each

L1/French pair. Once data collection was completed, we created an annotation protocol to obtain subjective evaluations of speech comprehensibility. We conducted an annotation campaign with 80 native French speakers and collected 3920 comprehensibility scores, half of which correspond to *a priori* comprehensibility (perceived comprehensibility) and the other half to *a posteriori* comprehensibility (comprehensibility of the message's meaning after considering the actual message's intended meaning).

To automatically predict learners' speech comprehensibility, we implemented a feature extraction phase on oral productions. These features are phonetic-phonological, lexical, syntactic, discursive, and semantic in nature. We achieved excellent prediction results for both the CAF-jp corpus (r=0.97, MAE=0.15) and the CAF-al corpus (r=0.98, MAE=0.18), using the *Random Forest* algorithm, an early fusion strategy, and a nested *leave-one-out* cross-validation.

Furthermore, by training a model on the entire CAF-jp corpus and testing it on the CAF-al corpus data, we also obtained good performance (r=0.98, MAE=0.34), demonstrating the generality of our approach.

Our various results show that our methodology for predicting comprehensibility is well-suited for evaluating French L2 learning and could even be applied to other L1/L2 language pairs.

Keywords: speech comprehensibility, language learning, linguistic features, automatic prediction