Compréhensibilité de contenus audiovisuels : quelles approches pour une mesure objective ?

Estelle Randria

Université Fédérale
Toulouse Midi-Pyrénées

THÈSE

En vue de l'obtention du
DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par l'Université Toulouse 3 - Paul Sabatier

Présentée et soutenue par

Estelle RANDRIA

Le 17 octobre 2022

Compréhensibilité de contenus audiovisuels : quelles approches pour une mesure objective ?

Ecole 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, Isabelle FERRANé et Lionel FONTAN

Jury

Mme Martine ADDA-DECKER, Rapporteure
M. Freiderikos VALETOPOULOS, Rapporteur
M. Julien PINQUIER, Directeur de thèse
Mme Isabelle FERRANé, Co-directrice de thèse
M. Lionel FONTAN, Co-directeur de thèse du monde socio-économique
M. Sylvain DETEY, Président

Table des matières

Remerciements

Merci à toute ma famille et à mes amis qui me soutiennent dans toutes mes aventures, même celles les plus incongrues, comme cette thèse !

Merci à mes directeurs de thèse Isabelle Ferrané et Julien Pinquier, des chercheurs, des encadrants, des enseignants mais surtout des personnes remarquables ! Ils m'ont accompagnée dans toutes mes études supérieures, du diplôme d'ingénieur au doctorat, je garderai un souvenir incroyable de toutes ces années grâce à eux !

Merci à mon encadrant industriel Lionel Fontan et à Archean Technologies pour les trois ans au sein de votre entreprise ! Merci à tous les Maxime qui m'ont appris énormément en programmation ! Sans vous je ne serai pas le docteur que je suis aujourd'hui.

Merci à toute l'équipe SAMoVA et à tous les doctorants (qui sont aujourd'hui docteurs) dont j'ai croisé la route ! Des esprits brillants et drôles !

Et enfin, merci à mon président de jury et à mes rapporteurs : leur temps et leur implication, et surtout leur approbation finale, m'ont permis d'obtenir ce doctorat !

Introduction générale

L'utilisation de documents audiovisuels est commune dans les méthodologies modernes d'enseignement de langue étrangère. De plus, avec la popularité de plateformes de vidéos telles que Netflix®, Prime Video®, Disney+®, Salto®ou Arte.tv®, les documents audiovisuels sont de plus en plus disponibles et accessibles. Il s'agit de ressources illimitées et intéressantes pour l'enseignement et l'apprentissage des langues. Cependant, les documents tels quels ne sont pas conçus à des fins pédagogiques. Quand les professeurs de langue étrangère souhaitent exploiter des documents audiovisuels pour leurs classes, il est nécessaire de passer par une phase de didactisation qui peut se définir par une « opération consistant à transformer ou à exploiter un document langagier brut pour en faire un objet d'enseignement » [Cuq et al., 2003].

Cette phase est non seulement fonction du niveau de langue des apprenants qui est nécessaire pour comprendre le contenu présenté, mais aussi de la thématique que l'enseignant souhaite aborder pendant son cours. S'il existe des plateformes, comme TV5 Monde[1] pour le français, qui mettent à disposition des documents didactisés, la didactisation est toujours faite de façon manuelle par les enseignants de langue étrangère. Il s'agit d'un investissement considérable de la part du professeur et cela peut avoir un impact sur la diversité des documents utilisés à des fins pédagogiques.

Problématique de la thèse

Partant de constat, le pôle de Recherche et Développement Archean LABS de la société Archean Technologies, qui collabore étroitement avec des professeurs de Français Langue Étrangère (FLE) exerçant au Japon, et l'Institut de Recherche Informatique de Toulouse (IRIT) ont proposé une thèse CIFRE dont l'objectif principal est de proposer des outils qui permettront d'alléger le processus de didactisation. En effet, dans cette thèse nous cherchons à étudier et développer une solution pour mesurer la capacité d'un contenu audiovisuel à être compris par un des apprenants. Dans ce contexte, nous avons voulu étudier en premier lieu quels sont les facteurs qui peuvent entrer en jeu dans la constitution de cette mesure et ensuite étudier quel procédé permettrait d'obtenir

1. https ://www.tv5monde.com/

une mesure objective qui soit corrélée avec une mesure similaire établie par des spécialistes en enseignement des langues.

Notion de compréhensibilité

La notion de compréhensibilité est au centre de ce travail. En fonction du domaine dans lequel elle est utilisée, sa définition varie. Le terme *compréhensible* est communément définit par « qui peut être compris, intelligible, clair ; compris sans difficulté[2] » et la *compréhensibilité*[3] comme « la qualité de ce qui est compréhensible ». Dans des domaines plus spécialisés comme le domaine clinique, la *compréhensibilité* est définie comme 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]. Dans le domaine de l'apprentissage des langues secondes [Kennedy and Trofimovich, 2008] la *compréhensibilité* correspond à la perception qu'ont les auditeurs de leur facilité à comprendre un énoncé produit par un locuteur non-natif. Dans ce travail de thèse, nous nous positionnons du point de vue l'apprenant et nous cherchons à mesurer la **capacité d'un document audiovisuel à être compris par un apprenant en fonction de son niveau dans la langue apprise, ou langue cible**. Nous parlerons également dans la suite de **niveau de compréhensibilité** d'un contenu audiovisuel qui pourra varier de **facile** à **difficile**, en considérant soit le document dans sa globalité, soit en considérant les différentes modalités qui le composent (audio et visuel).

Objectifs

Cette définition nous a permis d'identifier trois objectfis. Le premier est d'**identifier ce qui, au sein du document audiovisuel, va avoir un impact sur la capacité du contenu à être facilement compris**. Le deuxième objectif est de se servir de ces connaissances pour **mettre en place une mesure objective du niveau de compréhensibilité en explorant différentes approches et différentes modalités**, sans prendre en compte la dimension humaine susceptible d'affecter la compréhension. Le dernier objectif **consiste à comparer les approches et valider leurs pertinences vis à vis de leur intégration dans l'outil industriel** développé par la société Archean. Ainsi, le niveau de compréhensibilité associé à chaque contenu audiovisuel traité servirait d'indicateur pour faciliter le travail de didactisation réalisé par les professeurs de FLE.

2. https://www.larousse.fr/dictionnaires/francais/comprehensible/17770
3. https://www.cnrtl.fr/definition/comprÃ¯hensibilitÃ¯

Contributions

Cette thèse vise à construire un modèle de prédiction automatique du niveau de compréhensibilité de contenus audiovisuels. Pour mener ce travail à bien, il était nécessaire de considérer la question du niveau de compréhensibilité sous différents angles et de mettre en relation plusieurs disciplines : de la didactique des langues à l'apprentissage automatique, en passant par le traitement automatique des différentes modalités (audio, images et texte). Nous avons donc opté pour une approche multidisciplinaire. Elle est ainsi basée à la fois sur un **état de l'art** relatif aux phénomènes qui sont identifiés en **didactique des langues** comme affectant la compréhensibilité des contenus soumis aux apprenants, mais aussi sur un état de l'art des méthodes du domaine du **traitement automatique** qui permettent de modéliser et prédire le niveau de compréhensibilité de documents en fonction des modalités ou combinaison de modalités qui composent ces documents.

Dans notre étude, nous mettons en avant, grâce à cet état de l'art, l'ensemble des phénomènes qui ont été identifiés dans la littérature comme ayant une influence sur le niveau de compréhensibilité de documents en tenant compte des différentes modalités : texte, audio, images et une combinaison de modalités (audio+images, audio+images+texte). Pour étudier plus en avant ces divers phénomènes et pouvoir proposer par la suite notre mesure objectives, nous avons construit un corpus dédié à notre étude, le **corpus ESCAL** (Étude Subjective de la Compréhensibilité pour l'Apprentissage des Langues) qui regroupe des documents audiovisuels, extraits de documents de fiction. À ces documents ont été associées des évaluations réalisées par des experts en didactique *via* une interface dédiée et un protocole bien défini. Ce corpus sert, d'une part, à démontrer l'impact des phénomènes sur le niveau de compréhensibilité et d'autre part à obtenir des évaluations subjectives des enseignants de FLE. Ces évaluations subjectives serviront ensuite à la construction de nos modèles de prédiction du niveau de compréhensibilité.

Pour la construction de notre **mesure objective**, nous faisons le choix d'explorer deux approches. Dans la première approche, nous voulons obtenir une mesure qui ait une signification pour une personne qui y serait confrontée : notre but est d'apporter de la transparence sur la façon dont la mesure a été obtenue en explicitant les éléments qui y contribuent tout en quantifiant leur influence. Cette approche est fondée sur la sélection d'un ensemble de paramètres issus des différentes modalités prise en compte, à savoir les modalités audio, vidéo (succession d'images du document) et texte (transcription manuelle de la parole). Chaque modalité apporte des informations différentes. Ces paramètres bas niveaux sont tous en lien avec un phénomène qui va avoir une influence sur la compréhensibilité des contenus audiovisuels traités et ce qu'ils mesurent peut être facilement expliqué. Il s'agit d'une approche que nous qualifierons dans la suite de ce manuscrit d' ***approche interprétable***. La seconde approche fait appel à l'utilisation des réseaux de neurones profonds : à partir de réseaux pré-entraînés, capables d'extraire des informations des modalités audio, vidéo et texte, nous obtenons des représentations qui vont servir à alimenter différents

types de modèles de régression. Cette approche s'oppose à la première, car en faisant appel à des réseaux de neurones utilisés comme des « boîtes noires », nous perdons la capacité à expliquer les phénomènes qui ont été pris en compte pour aboutir à la mesure produite. En opposition à la première approche, nous qualifierons par la suite cette approche d'*textbfapproche neuronale*. En comparant les modèles issus de ces deux approches, nous identifions celle qui est la plus pertinente pour répondre à notre problématique et nous pouvons ensuite mettre en place le protocole qui permettra d'intégrer notre mesure objective du niveau de compréhensibilité dans notre **application industrielle à destination des professeurs de FLE**.

Organisation du mémoire

Ce mémoire de thèse se compose de deux parties, chacune composée de plusieurs chapitres.

La première partie porte sur l'étude des facteurs qui interviennent dans la mesure de la compréhensibilité des documents audiovisuels. Le premier chapitre est consacré au rôle des « documents authentiques » et de l'interaction dans la didactique des langues étrangères. Nous étudions ensuite, dans un deuxième chapitre, les facteurs qui affectent la compréhensibilité de manière générale, en nous basant sur la littérature du domaine de la didactique des langues. Nous présentons le corpus ESCAL dans le troisième chapitre, en décrivant sa mise en place, la phase d'annotation par des experts et l'analyse des résultats qui en découle. Enfin, dans un quatrième chapitre, nous identifions à partir de l'état de l'art et des analyses du corpus quels sont les phénomènes et facteurs pertinents pour estimer le niveau de compréhensibilité d'un document audiovisuel.

La seconde partie de ce mansucrit est centrée sur l'élaboration de notre mesure objective du niveau de compréhensibilité de documents audiovisuels. Le premier chapitre de cette partie porte sur l'ensemble des paramètres pertinents qui vont être la base de de notre mesure objective. Ensuite nous détaillons dans un deuxième chapitre la démarche et les résultats de l'*approche interprétable* avant de détailler dans un troisième chapitre la méthodologie et les résultats de l'*approche neuronale*. Le troisième chapitre présente un cas d'application sur un nouveau jeu de données, sur lequel nous comparons les évaluations liées à la perception humaine du niveau de compréhensibilité au comportement du modèle de prédiction présentant les meilleurs résultats.

Enfin, nous concluons sur l'ensemble des travaux qui ont été réalisés et les perspectives qui se présentent à l'issue de cette thèse. Nous présentons notamment l'ensemble des applications industrielles envisageables à partir du meilleur indicateur que nous avons pu obtenir.

Première partie

Étude des facteurs intervenant dans la mesure de compréhensibilité de documents audiovisuels

Introduction

Dans le cadre de l'enseignement des langues étrangères, un enseignant qui souhaite exploiter un document audiovisuel doit s'assurer que ce document est adapté au niveau des apprenants. Les **documents pédagogiques** (conçus spécifiquement pour les cours de langue) sont fournis avec une indication sur le niveau requis dans la langue cible pour que le document soit compréhensible par un apprenant. Cette indication peut être apparentée à un niveau de compréhensibilité du document, la compréhensibilité étant la difficulté du document à être compris. Ainsi, le niveau de compréhensibilité sert d'indicateur à l'enseignant pour choisir des documents adaptés au niveau de connaissance de la langue de ses apprenants. Cependant, un tel indicateur n'est pas défini pour les documents qui n'ont pas été spécifiquement conçus pour les cours de langue. Dans ce cas on parle de **documents authentiques**, dont une définition détaillée sera donnée dans la section 1.3.1). Or, avec l'évolution des méthodologies d'apprentissage et des objectifs pédagogiques, l'usage de documents authentiques, et notamment de documents audiovisuels de fiction, est maintenant entrée dans les habitudes d'enseignement. Les enseignants en FLE ont effectivement besoin d'enseigner également à leurs apprenants comment interagir en situation réelle, au quotidien. Les interactions, d'après la définition de Véronique Traverso, sont des situations de communication qui sont cohérentes en termes d'objectif, de participants (personnes qui prennent part à la situation de communication) et de cadre spatio-temporel [Traverso, 1996]. Les documents authentiques constituent une source pertinente pour trouver des interactions auxquelles il est intéressant d'exposer les apprenants. Cependant, en l'absence d'indicateurs concernant le niveau de langue requis, les enseignants doivent analyser le document pour savoir s'il est approprié dans le contexte de l'apprentissage. L'investissement que cela exige tend à limiter la diversité des documents authentiques proposés par les enseignants. Il existe des organismes qui se consacrent à la création d'exercices de difficulté variable en lien avec des documents authentiques (c'est le cas de TV5 Monde[4] par exemple) mais, le niveau de connaissance de la langue nécessaire pour réaliser les exercices est lié directement aux exercices créés et il n'y a pas d'information sur le niveau de compréhensibilité du document authentique en lui-même. Il y a donc un besoin autour de l'attribution de niveau de compréhensibilité de documents authentiques. L'objectif de cette thèse est de répondre

4. http://www.tv5monde.com/

à ce besoin et de proposer une mesure qui permette d'évaluer automatiquement le niveau de compréhensibilité de documents audiovisuels, en exploitant des paramètres issus des composantes audio, vidéo et textuelle et éventuellement en les combinant (paramètres multimodaux). Pour réussir à construire une telle mesure, il est nécessaire d'identifier les éléments qui ont un impact sur la compréhensibilité, ce sera la problématique de cette première partie.

Avant toute chose, il est important de comprendre comment et pourquoi l'exploitation des documents authentiques contenant des interactions a pris une part importante dans l'enseignement des langues étrangères, c'est pourquoi le premier chapitre sera consacré à un bref historique de l'évolution des méthodologies d'enseignement qui a conduit à mettre l'apprentissage de l'interaction au centre de l'enseignement de langues étrangères et a ainsi favorisé l'exploitation de documents authentiques. Ensuite, je présenterai l'état de l'art réalisé autour de la didactique des langues et du traitement automatique pour identifier l'ensemble de facteurs susceptibles de rentrer en jeu dans la mesure du niveau de compréhensibilité des documents selon leur types, qu'ils soient représentés par des textes, des enregistrements audio ou audiovisuels. Cet état de l'art servira de point de départ pour la mise en place de l'étude réalisée ensuite, menée auprès d'enseignants de FLE afin de déterminer quels sont les éléments et les phénomènes qui jouent un rôle dans la compréhensibilité de documents audiovisuels.

Chapitre 1

Apprentissage des langues étrangères : rôles des documents authentiques et de l'interaction

1.1 Introduction

Dans le contexte de l'apprentissage des langues étrangères, il est important de faire la distinction entre les termes « **méthode** » et « **méthodologie** ». Pour la suite, nous utilisons les définitions apportées par Christian Puren [Puren, 1988] dans son introduction intitulée « Histoire des méthodologies de l'enseignement des langues » publiée en 1988. Il y définit la méthode comme « un ensemble de procédés et de techniques de classe qui ont pour objectif de susciter un comportement ou une activité déterminés » chez l'apprenant et la méthodologie, comme un « ensemble cohérent de procédés, techniques et méthodes qui s'est révélé capable [...] de générer des cours relativement originaux par rapport aux cours antérieurs et équivalents entre eux quant aux pratiques d'enseignement / apprentissages induites ». C'est donc l'originalité d'une méthodologie qui crée une rupture vis-à-vis des autres méthodologies existantes.

Si nous tentons de résumer ce qui fait la distinction entre méthode et méthodologie, la méthode englobe le matériel, les supports utilisés pour l'enseignement, tandis que la méthodologie a une dimension plus théorique. Celle-ci elle relève de la façon d'appréhender et de réaliser l'enseignement, en fonction des objectifs pédagogiques par exemple. La méthode va découler d'une méthodologie donnée et permettra l'exécution de ses principes.

Depuis la première méthodologie reconnue, qui était utilisée pour l'enseignement des langues dites « mortes », de nombreuses déclinaisons de méthodologies sont apparues [Puren, 1989], notamment pour s'adapter aux changements amenés par le contexte historique. Les supports utilisés pour l'enseignement ont évolué en même temps que les méthodologies, dans leur type (texte, audio, vidéo) et leur contenu [Riquois, 2010].

1.2 Méthodologies pour l'apprentissage des langues étrangères : bref historique

1.2.1 Méthodologie traditionnelle

La méthodologie la plus ancienne est celle que nous désignons comme méthodologie traditionnelle (ou classique) : elle est très utilisée pendant la seconde moitié du $19^{\text{ème}}$ siècle. À l'origine, elle servait pour enseigner le grec et le latin. Accordant une grande importance aux œuvres littéraires, mais aussi aux œuvres artistiques, elle est principalement axée sur l'écrit et repose sur l'apprentissage de la grammaire et sur la traduction de textes littéraires. L'oral tient un rôle secondaire. En termes de supports, l'enseignement avec cette méthodologie nécessite surtout l'utilisation de manuels de grammaire et de textes traduits, donc la méthodologie traditionnelle fait plus appel à du contenu textuel, voire parfois à des images (faisant plutôt office d'illustrations). La méthodologie traditionnelle induisait, à ses origines, une certaine passivité de l'étudiant dans son propre

apprentissage, les cours étant effectivement « dominés » par l'enseignant faisant figure de « savoir [et d']autorité » [Seara, 2001]. Par la suite, même si la méthodologie traditionnelle n'a pas été abandonnée et est encore utilisée actuellement (dans le cadre de l'enseignement supérieur notamment), des méthodologies plus centrées sur l'intuition et la motivation de l'étudiant ont été mises en place. Nous pouvons citer ici les méthodologies directe et active, qui laissent plus la place à la production orale et encouragent plus l'implication de l'apprenant dans son apprentissage.

1.2.2 Méthodologie directe

La méthodologie directe est considérée par Christophe Puren comme étant « la première méthodologie spécifique à l'apprentissage des langues vivantes étrangères » [Puren, 1988]. Elle s'inspire de la manière dont est acquise la langue maternelle et fait appel à un ensemble de procédés et de techniques qui permettent d'éviter d'utiliser la langue maternelle comme intermédiaire durant l'apprentissage, la traduction n'a pas de place dans l'utilisation de cette méthodologie. L'écoute et la production orale tiennent une place centrale dans cette méthodologie, où cette fois l'écrit est relégué au second plan. À l'inverse de la méthodologie traditionnelle, la méthodologie directe fait beaucoup appel à la réactivité des élèves, ce qui induit l'utilisation de ce qui est désigné comme méthode active. La méthode active consiste à encourager les apprenants à participer activement pendant leurs classes : en répondant à des questions, en jouant des saynètes ou en répétant ce qui est entendu... La méthodologie directe est la première à avoir pris en considération la motivation de l'étudiant. L'enseignant s'adapte aux besoins et aux intérêts de l'élève, et fait varier la difficulté des contenus ainsi que ses thématiques. Le matériel utilisé est plus diversifié, car tout ce qui peut permettre de montrer, nommer et expliquer sans avoir recours à la traduction peut être utilisé. Les supports écrits restent utilisés avec cette méthodologie ; ils comportent des leçons de grammaire, des listes de vocabulaire, mais aussi des textes. Des tableaux muraux sont aussi proposés, pour permettre de montrer le plus d'objets et de concepts possible aux apprenants.

1.2.3 Méthodologie active

La méthodologie active mêle méthodologie traditionnelle et méthodologie directe. Elle garde les principes de la méthode directe : la motivation de l'élève reste centrale dans l'enseignement, et l'oral continue de tenir une place importante, mais elle est plus flexible puisque l'utilisation de la langue maternelle en cours est autorisée, alors que la méthodologie directe était parfois trop rigide en l'interdisant en classe. Cependant, la méthodologie active réintègre certains procédés et techniques de la méthodologie traditionnelle, notamment en redonnant sa place à l'écrit en cours de langue étrangère. C'est avec la méthodologie active que l'utilisation de matériels et supports plus novateurs est apparue : les textes sont utilisés comme supports didactiques, puis les supports audio, notamment pour l'étude de la phonétique, vont commencer à être plus présents

dans les salles de classe (radio, magnétophone...). Avec l'apparition de nouvelles technologies, et la mise à disposition de nouveaux matériels pédagogiques (magnétophones, projecteurs...), les enseignants vont peu à peu se détacher de la méthodologie active, pour se tourner vers des méthodologies tirant parti le plus possible des innovations techniques, c'est ce qui va les mener à l'utilisation de méthodes audiovisuelles.

1.2.4 Méthodologie audio-orale

Il est intéressant de s'attarder sur la méthodologie audio-orale, qui est née pendant la Seconde Guerre mondiale aux États-Unis et qui est, par exemple, encore utilisée aujourd'hui par la Légion étrangère. Cette méthodologie s'inspire d'une méthode conçue par le linguiste Bloomfield à la demande de l'armée (la « méthode de l'armée ») [Bloomfield, 1942], qui avait besoin de former rapidement des personnes parlant d'autres langues que l'anglais. L'apprentissage passe par l'audition et la compréhension, l'expression orale, la lecture puis la rédaction [Coste, 1970]. La méthodologie audio-orale s'appuie sur la psychologie béhavioriste [Watson, 1916], et suppose que l'apprentissage d'une langue repose sur un réflexe conditionné : stimulus, réponse, renforcement. Les apprenants sont soumis à des écoutes intensives pour assimiler les prononciations, avant de s'exercer eux-mêmes à la production. Les structures morphosyntaxiques sont enseignées à partir de phrases modèles, qui sont ensuite manipulées (répétitions, substitutions...) par les apprenants pour se familiariser avec la structure à intégrer. La « méthode de l'armée » comporte majoritairement des exercices d'écoute, de manipulation de répétition et d'imitation. Le fait qu'il y ait une grande partie du travail autour de l'oral a favorisé l'utilisation des supports d'écoute modernes : des films, des extraits radio, des disques... L'utilisation de textes littéraires arrive tardivement dans la formation des élèves, mais garde néanmoins sa place dans l'apprentissage.

1.2.5 Méthodologie audiovisuelle

En même temps que la méthodologie audio-orale, des méthodes dites audiovisuelles, utilisant des documents audiovisuels, sont apparues. La méthodologie SGAV (Structuro-Globale Audio-Visuelle) [Ivan et al., 2006] provient de cette nouvelle mouvance et proposait de combiner l'utilisation des manuels, d'images fixes (des diapositives ou des reproductions d'œuvres d'art par exemple) et du son (il s'agissait de supports magnétiques), en suivant l'idée que l'apprentissage de la langue passe par l'oreille et la vue. Bien qu'étant apparue dans la même période que la méthodologie audio-orale, Puren fait remarquer que la méthodologie SGAV compte plus de points communs avec la méthodologie directe [Puren, 1988], notamment :

- la prédominance de l'apprentissage de l'oral sur l'apprentissage de l'écrit,
- l'utilisation de l'image comme vecteur d'explication directe sans passer par la langue maternelle,

— l'exploitation du dialogue (ici oral et non plus écrit) pour l'acquisition du vocabulaire et des structures,
— la sollicitation des élèves et le maintien de leur motivation, qui passe par exemple par la mise en situation dans les dialogues de personnages dont ils peuvent se sentir proches.

La méthodologie SGAV a comme particularité d'accorder tout autant d'importance à l'apprentissage d'éléments linguistiques (grammaire, lexique, phonétique) qu'à l'apprentissage d'éléments extralinguistiques (gestes, mimiques...). Tout doit être assimilé ensemble, comme cela se fait pour l'apprentissage de la langue maternelle. C'est ce qui est désigné comme du structuroglobalisme. L'exploitation de bandes magnétiques conjointement à des supports visuels est un moyen d'enseigner ces éléments simultanément aux apprenants. En effet, cela permet aux élèves de percevoir les sons, mais aussi les gestes, les rythmes et l'intonation, tout cela en étant exposés à la présentation de dialogues ancrés dans une situation de communication tenant compte des lieux, des circonstances du dialogue, mais aussi des participants et de la situation interactionnelle. Le vocabulaire et les structures à apprendre sont ainsi inscrits dans un contexte cadré, pour faciliter la compréhension de leur utilisation. La méthodologie SGAV cherche à développer à la fois la compréhension orale, la production orale, la compréhension écrite et la production écrite de l'élève, bien que la production orale tienne une place prépondérante (l'écrit étant travaillé très tardivement). Il faut donc coordonner tous les supports (écrits, audio, visuels...) pour « développer les compétences orales en relation avec les compétences écrites et visuelles » [Riquois, 2010]. Cependant, dans la pratique, les supports peuvent être exclusivement oraux, écrits ou visuels tout comme ils peuvent combiner image et texte (bandes dessinées...). Des difficultés liées aux coûts et aux problèmes techniques ont mis à mal l'application des méthodes audiovisuelles. En effet, comme il pouvait être coûteux d'investir dans cette méthodologie, certains manuels étaient conçus de telle façon à ce que les enseignants puissent remplacer les enregistrements en faisant eux-mêmes la lecture et, les manuels mettaient à disposition les images directement, au cas où les diapositives ne seraient pas disponibles. D'autres manuels ont également été conçus pour ne plus avoir recours qu'à eux pendant les classes, laissant ainsi de côté l'utilisation de matériel complémentaire. La méthodologie SGAV a décliné, et a été remplacée par l'approche communicative.

1.2.6 Approche communicative

En 1975, le conseil de l'Europe met en place le « niveau seuil » (*threshold level*) pour l'anglais [Van Ek, 1975]. Le « niveau seuil » sert à répertorier l'ensemble des compétences à acquérir pour pouvoir communiquer efficacement dans un pays étranger. Ce niveau va servir de modèle pour les autres langues, dont le français [Coste and et al, 1976] et il va grandement modifier les objectifs pédagogiques des enseignants de langue étrangère. Cette approche se veut adaptative en fonction des différents contextes d'apprentissage. En effet, elle ne rejette pas

les autres méthodologies de façon catégorique, puisque certaines peuvent s'avérer plus adaptées en fonction du contexte. Ce qui explique que cette méthodologie soit désignée comme une « approche », pour souligner sa souplesse et son ouverture. Cette approche se recentre sur l'apprentissage de la communication et sur les besoins et les attentes de l'apprenant. La langue est vue comme **un instrument de communication et d'interaction sociale**, l'oral demeure important, mais l'écrit n'est plus mis de côté, puisque les besoins de l'élève peuvent relever de l'écrit (savoir rédiger une lettre, écrire des indications...). Tous les moyens sont mis en œuvre pour que l'apprenant puisse s'adapter et surmonter les situations de communication qu'il peut être amené à rencontrer : l'apprenant est actif dans son apprentissage et l'enseignant prend en considération à la fois le groupe-classe et l'individu. Il articule son cours en fonction de qui sont ses apprenants, des objectifs à leur faire atteindre et il réfléchit à la façon dont il va leur permettre d'atteindre ces objectifs. L'enseignement évolue en fonction des progrès effectués par le(s) apprenant(s), ce qui oblige les manuels à être construits différemment : l'enseignant ne suit plus une progression linéaire et doit donc pouvoir sélectionner comme il le souhaite des rubriques du manuel en fonction des besoins de ses élèves. Les manuels sont construits à partir de tâches d'apprentissage communicatives et **l'interaction sociale est au cœur du dispositif**. Le vocabulaire et les structures sont présentés et intégrés par l'élève grâce à des situations de communication données. Les manuels favorisent l'implication des élèves dans des situations d'énonciation (jeux de rôle, simulations...), que ce soit entre eux (improvisation de dialogues) ou avec leurs enseignants. Les documents proposés dans les manuels conçus pour l'approche communicative peuvent être fabriqués à des fins pédagogiques ou être authentiques (c'est-à-dire non conçus pour une classe de langue étrangère). Pour pouvoir présenter des concepts issus du quotidien ou du milieu où l'apprenant veut employer la langue, il est fortement conseillé aux enseignants d'utiliser des documents authentiques variés pour atteindre leurs objectifs pédagogiques. Le type de documents exploités par les enseignants peut varier, allant de la photo aux vidéos (extraits de films, journal télévisé...) en passant par des émissions radiophoniques ou des publicités. Les documents authentiques sont intéressants pour susciter la curiosité de l'apprenant, entretenir sa motivation et l'aider à développer leur lexique. L'approche communicative n'est pas toujours perçue par les enseignants comme une méthodologie en tant que telle, mais elle est bien implantée dans les classes.

Actuellement, les enseignants ne s'accordent pas sur « une méthodologie unique [et] universelle » [Seara, 2001], ce qui amène une certaine diversité dans les cours en fonction des besoins pédagogiques. Les méthodologies qu'ils décident d'utiliser varient selon les besoins des enseignants et/ou des étudiants, et il y a une grande diversification dans les approches proposées et le matériel utilisé. Les enseignants n'utilisent plus forcément un seul manuel pour leurs classes, et peuvent plutôt tendre à construire leur propre méthode en associant des éléments de plusieurs manuels.

1.3 Documents authentiques et interaction pour l'apprentissage des langues étrangères

Si nous constatons que les méthodologies ont évolué, et qu'il n'en existe pas une unique actuellement pour l'enseignement, ce qui demeure constant est le fait de placer l'apprenant au cœur de l'enseignement, et nous pouvons aussi noter que le document authentique a gardé une place de choix dans l'enseignement des langues étrangères.

1.3.1 Documents : authentiques *vs* pédagogiques

Comme nous l'avons dit, les documents vont du contenu littéraire au contenu vidéo. Les professeurs peuvent utiliser deux types de documents. Ils peuvent utiliser des documents pédagogiques (ou « fabriqués ») ou des documents dits « authentiques ».

Les documents pédagogiques sont créés de toute pièce pour les cours de langue, et sont pensés entièrement pour les besoins de l'enseignement de la langue étrangère, en d'autres termes ils ont été conçus autour de critères linguistiques et pédagogiques. Un exemple de documents pédagogiques est le CD fourni avec les manuels pédagogiques où nous trouvons des dialogues joués : la prononciation et le débit sont modulés pour que la compréhension par les apprenants soit plus aisée, et l'environnement sonore est contrôlé pour que le bruit ou la musique de fond éventuelle ne gêne pas l'écoute.

Le document authentique est conçu pour des natifs de la langue enseignée, il peut s'agir d'un document écrit, audio ou audiovisuel [Aslim-Yetis, 2010]. Le document authentique est un plus pour l'enseignement des langues étrangères, car la langue entendue est plus proche de ce à quoi un apprenant pourra être confronté en situation réelle : « il met en jeu les diverses composantes d'une réelle compétence de communication » [Assaad, 2005].

Dans un document audio ou audiovisuel pédagogique, le document est modulé pour être adapté au niveau des étudiants. Le document authentique, lui, est moins normé et permet à l'apprenant d'entendre une version plus « vraie » de la langue, avec ses variations, son rythme, son jargon... En effet, dans un document audio ou audiovisuel pédagogique, le document est conçu pour être adapté au niveau des étudiants. L'ensemble des avantages apportés par l'usage de documents authentiques est notamment listé par Hedaywa et Sourak dans [Hedaywa and Sourak, 2013].

Utiliser des documents authentiques permet d'entretenir l'intérêt et la motivation des apprenants. Il existe une infinité de documents authentiques, ce qui permet de maintenir de la diversité dans les documents présentés. Et avoir la capacité de comprendre un document qui est initialement destiné aux natifs de la langue peut être une source de satisfaction et de confiance pour les étudiants en langue étrangère. S'il y a une grande variété de documents authentiques qui

peuvent être exploités pour les cours de langue, ce sont les documents audiovisuels qui nous intéressent dans cette thèse.

1.3.2 Intérêt de la modalité audiovisuelle

Je me suis intéressée à l'étude des interactions comme unité de segmentation de contenu audiovisuel, notamment des films, et donc, aux documents de type audiovisuels authentiques. Si l'utilisation des vidéos nous a intéressés, au-delà de l'aspect motivant évoqué un peu plus tôt, c'est en grande partie parce que les films, notamment les films de fiction, sont la plupart du temps constitués d'interactions cadrées. Les caractéristiques et le déroulement des interactions sont déjà définis à l'aide d'un script, ce qui permet de les identifier de façon plus simple que dans d'autres types de documents audiovisuels. Comme nous l'avons vu plus tôt, présenter des dialogues ancrés dans un contexte peut être favorable pour l'apprenant souhaitant développer ses compétences communicatives. De plus, le fait d'avoir accès à la gestuelle et aux expressions faciales des participants peut s'avérer important pour comprendre les enjeux et le déroulement de l'interaction.

Utiliser une vidéo authentique permet dans un premier temps d'améliorer ses compétences linguistiques en entendant la langue dans une situation cadrée, où l'image vient se rajouter au son pour comprendre la situation de communication et ses enjeux. Les vidéos authentiques permettent aussi d'être exposé à des informations non verbales qui peuvent être liées à la façon de communiquer dans le/les pays où la langue apprise est parlée. En effet, savoir parler une langue peut s'avérer insuffisant si les codes sociolinguistiques n'ont pas été assimilés. Or, dans les vidéos authentiques, l'apprenant fait face à des situations de communication ancrées dans une réalité culturelle et a accès à un ensemble d'éléments non verbaux produits par des natifs. Tous ces éléments font de la vidéo authentique un support idéal pour répondre aux besoins pédagogiques de l'enseignant, qui doit également enseigner à ses apprenants comment interagir [Vargas, 2006].

1.3.3 Étude des situations de communication et des interactions

L'un des objectifs au cœur de l'enseignement des langues étrangères est de mettre les apprenants face à des situations de communication où ils doivent interagir dans la langue apprise. Le processus de communication est classiquement décrit par les étapes suivantes [Watzlawick et al., 1972] :

— le locuteur, par un encodage lexical, syntaxique, morphologique et phonologique produit une phrase, porteuse d'un message, d'une intention avec comme attente que son interlocuteur saisisse le message/l'intention en fonction de la situation de communication,
— l'interlocuteur décode la phrase et reconstruit l'intention du locuteur en se basant sur des inférences liées au contexte, à la situation de communication.

Le concept des « interactions langagières » permet de comprendre plus en détail les différents éléments qui entrent en jeu dans le processus de communication [Camus, 1999].

Que ce soit pour les interactions entre des participants parlant une même langue ou les interactions entre un natif et un non-natif d'une langue cible, il est nécessaire que les participants parviennent à s'entendre pour que l'interaction soit un succès. Pour comprendre et se faire comprendre, il faut d'abord avoir acquis des compétences linguistiques suffisantes pour être capable de communiquer avec l'autre. Mais, cela ne suffit pas, il faut également avoir acquis un certain nombre de codes sociolinguistiques indispensables, car il existe très souvent un lien fort entre langage et culture : une bonne connaissance linguistique ne suffit pas s'il manque la compréhension de la culture. Un exemple simple est l'utilisation du tutoiement et du vouvoiement en français : un apprenant en FLE doit maîtriser le tutoiement et le vouvoiement, mais il doit aussi savoir dans quels cas utiliser l'un ou l'autre. Un autre exemple est celui du Japon, où la hiérarchie tient une place centrale, et cela s'en ressent dans la langue : la façon de s'exprimer va fortement dépendre de la position sociale/hiérarchique de la personne à qui nous nous adressons. Par conséquent, même si l'encodage et le décodage de la langue sont un succès, l'interaction peut être vouée à l'échec, car les codes culturels n'ont pas été compris ou respectés. En effet, comme ils ne sont pas connus, le message et l'intention sont mal interprétés. Ainsi, l'un des deux participants peut être choqué, vexé ou tout simplement incapable de rattacher l'intention supposée au contexte dans lequel l'interaction se déroule.

En résumé, pour que nous puissions interagir correctement dans une langue étrangère, il faut :
— avoir les compétences linguistiques,
— connaître les codes sociolinguistiques.

Les situations liées à l'interaction sont idéales pour l'apprentissage de la langue, car l'apprenant est plongé dans un cadre où se combinent les nécessités de parler et de se comporter de façon appropriée. Il peut à la fois intégrer des connaissances linguistiques et socioculturelles.

1.4 Bilan

Dans ce chapitre, nous avons vu que l'évolution des méthodologies dans le domaine de la didactique des langues a amené l'interaction au cœur de l'enseignement, favorisant l'usage de documents authentiques pendant les classes de langue. La vidéo authentique constitue un support pertinent pour enseigner aux apprenants de quelle façon interagir, ce qui leur permet d'assimilier des compétences langagières mais aussi socioculturelles.

Cependant, il demeure un obstacle pour les enseignants qui souhaiteraient

utiliser des vidéos authentiques. En effet, il est nécessaire de déterminer le niveau de compréhensibilité d'une vidéo qu'ils voudraient utiliser en cours de langue, pour voir si celle-ci sera accessible à des étudiants d'un niveau spécifique. Si les enseignants sont capables d'estimer ce niveau de compréhensibilité eux-mêmes, c'est une tâche ardue et assez chronophage, ce qui peut limiter le nombre de vidéos avec lesquelles un enseignant sera amené à travailler. Actuellement, avec la diversification des méthodes de diffusion, mais aussi la popularisation des plateformes de visionnage vidéo, un très vaste choix de vidéos authentiques pourrait être exploité, mais ne l'est pas à cause de la contrainte que la didactisation amène. Il existe un besoin d'avoir accès à des vidéos authentiques préalablement annotées en termes de niveau de compréhensibilité, pour que les enseignants aient un plus large choix de vidéos authentiques à présenter pendant leurs cours. C'est ici que notre étude entre en jeu. L'objectif industriel, dans lequel s'inscrit cette thèse, est de mettre en place un service destiné aux enseignants qui leur permette d'exploiter des interactions issues de documents de fiction, en leur présentant des extraits pour lesquels le niveau de compréhensibilité a été estimé automatiquement. Cet indicateur leur permettra d'une part, d'affiner leur choix et et d'autre part d'alléger la tâche de didactisation de ces vidéos authentiques, sélectionnant ainsi des documents audiovisuels adaptés au niveau de leurs apprenants.

Pour cela, il faut mettre en place une méthode d'analyse de contenu qui permette de produire des indicateurs du niveau de compréhensibilité d'un document audiovisuel authentique. L'un des premiers chapitres de cette thèse porte donc sur l'impoact de ces facteurs sur la compppréhensibilité. Nous avons réalisé l'identification de ces facteurs grâce à une étude de la littérature dans les domaines de la didactique des langues et du traitement automatique. Cet état de l'art sert de point de départ pour réaliser une évaluation subjective de la compréhensibilité indispensable pour valider ces différentes facteurs et en quantifier l'influence. La connaissance de ces phénomènes permet alors de cibler un ensemble de paramètres qui sont corrélés et peuvent être extraits de façon automatique, ce qui constitue un premier pas vers la construction de notre mesure objective du niveau de compréhensibilité.

Chapitre 2

Facteurs affectant le niveau de compréhensibilité

Sommaire

2.1 Introduction

Les documents audiovisuels comprennent plusieurs modalités : le texte, l'audio et la vidéo (dans le sens d'une séquence d'images). Cette multimodalité implique que la compréhensibilité peut être influencée par des éléments venant soit des modalités prises individuellement soit de leurs combinaisons. Chaque personne confrontée à un film qui n'est pas visionné dans sa langue native a pu faire face à des difficultés à appréhender une situation de communication à cause de la vitesse d'élocution, de la présence de mots qu'ils ne connaissaient pas ou d'un surplus d'informations à assimiler simultanément. Aux vues de la pluralité des sources potentielles de difficulté ou de facilitation de la compréhensibilité, une première étape indispensable est de s'intéresser à ce que nous trouvons dans la littérature dans le domaine de la didactique des langues et du traitement automatique pour établir une liste suffisamment exhaustive des facteurs qui y ont été retenus. Dans ce chapitre, nous étudions donc les facteurs qui, dans la littérature, sont désignés comme influençant la compréhensibilité en considérant :

— ce qui est dit (que ce soit à l'oral ou à l'écrit),
— comment cela est dit,
— dans quel contexte cela est dit.

2.2 Compréhensibilité de l'écrit : mesures de lisibilité

L'évaluation de la difficulté globale de textes est un sujet qui intéresse les professeurs et les chercheurs depuis de nombreuses années [Sherman, 1893, Thorndike, 1921]. Ces études ont principalement été menées pour s'assurer que les personnes soient confrontées à des textes qui soient en accord avec leur niveau de langue. Ceci est un aspect important, notamment dans l'apprentissage des langues. Évaluer à quel point un texte est difficile à comprendre est crucial pour déterminer le niveau de langue requis pour pouvoir utiliser et comprendre un texte spécifique en classe. Un enfant n'aura pas la même maîtrise de la langue maternelle qu'un adolescent et un adulte. De même, un débutant aura moins de compétences linguistiques qu'un apprenant de niveau avancé lors de l'apprentissage de langue étrangère. L'étude de compréhensibilité de documents écrits peut servir pour la constitution de méthodes, ou pour rassembler un ensemble de documents de niveau de compréhensibilité équivalent. Une autre application possible est la rédaction de consignes dans les manuels techniques pour qu'ils soient accessibles au plus large panel de lecteurs possible.

2.2.1 Lisibilité : définitions

Dans le domaine de l'étude des textes, étudier la compréhensibilité revient à étudier la **lisibilité**, définie par Bourque comme l'« aptitude du texte à se faire comprendre » [Bourque, 1989]. Dale et Chall étayent cette définition en ajoutant

qu'il s'agit de la somme des éléments dans du contenu textuel qui vont affecter la compréhension du lecteur, sa vitesse de lecture et son niveau d'intérêt pour le contenu [Dale and Chall, 1949]. Dans ce contexte, nous dissocions la forme et le fond d'un texte avec :

— l'étude des aspects linguistiques et conceptuels d'un texte pour déterminer sa difficulté,
— l'étude des effets de l'aspect « esthétique » d'un texte (sa mise en page, sa typographie...) sur la compréhension.

Ici, nous nous concentrons sur la lisibilité dans le sens de l'étude des **aspects linguistiques**. La lisibilité d'un texte peut être soit mesurée, soit prédite. Mesurer la lisibilité consiste à demander à des experts (par exemple des enseignants) ou à un groupe de lecteurs représentatifs de juger la difficulté globale d'un texte donné. Un problème possible avec cette approche est que les résultats obtenus ne soient pas toujours reproductibles, notamment dans le cas où il n'est pas demandé systématiquement aux sujets de justifier leurs choix. Prédire la lisibilité d'un texte consiste à trouver des formules ayant pour but de « donner un index de difficulté globale probable pour les lecteurs » [Klare, 1974].

2.2.2 Historique des mesures de lisibilité

2.2.2.1 Mesures de lisibilité pour l'anglais

Les premiers travaux de lisibilité ont d'abord été conduits pour la langue anglaise. Sherman a été le premier chercheur à étudier la littérature d'un point de vue statistique [Sherman, 1893]. Ses recherches ont mis en avant une relation entre la **longueur des phrases, le vocabulaire et la lisibilité** : d'après ses observations, plus les phrases sont courtes et contiennent des termes concrets, plus la lisibilité augmente. Plus tard, des études de l'**influence du vocabulaire** sur la lisibilité ont été menées. Thorndike a créé la première liste de mots en anglais classés selon leur fréquence d'utilisation [Thorndike, 1921]. Elle comptait initialement 10 000 mots, mais Thorndike a ensuite plusieurs fois mis à jour cette liste.

Ses études ont pour une origine une constatation faite par des professeurs de langues russes et allemandes qui ont remarqué que plus un mot est utilisé fréquemment, plus il est simple à comprendre. Lively et Pressey ont construit la première formule de lisibilité adaptée aux enfants [Lively and Pressey, 1923], en se servant de cette liste de Thorndike. Ils ont tenté de mettre en place une mesure de difficulté du vocabulaire pour des manuels scolaires. Par groupe de 1000 mots, ils mesuraient le nombre de mots différents et le nombre de mots qui ne faisaient pas partie de la liste de 10 000 mots de Thorndike. Bien qu'elle soit compliquée à appliquer, cette formule a été l'une des premières utilisées pour mesurer la lisibilité. Parmi les plus connues, nous trouvons la formule de Flesch [Flesch, 1948] et la formule de Dale et Chall [Dale and Chall, 1948].

La formule de Flesch permet d'attribuer un score entre 0 et 100 à des textes.

Plus le score augmente, plus le texte est facile à comprendre :

$$206,835 - (1,815 * ASL) - (84,6 * ASW) \tag{2.1}$$

avec :
- ASL (Average Sentence Length) qui correspond au nombre moyen de mots par phrase du texte,
- ASW (Average of Syllables per Word) qui correspond au nombre moyen de syllabes par mot.

La formule de Dale et Chall réutilise la longueur moyenne des phrases, et s'exprime par la relation :

$$0,1579 * PDW + 0,496 * ASL + 3,6365 \tag{2.2}$$

avec PDW (Percentage of Difficult Words) qui correspond au pourcentage de mots considérés comme difficiles car absents de la liste de 3000 mots faciles définie par Dale et Chall.

Cette formule renvoie le niveau de lecture (en termes de niveau scolaire) nécessaire pour pouvoir comprendre un texte. Si la formule de Flesch se concentre plus sur un **aspect syntaxique**, en utilisant des mesures liées à la longueur des phrases et à la longueur des mots, la formule de Dale et Chall prend en compte à la fois la dimension syntaxique en incluant la longueur des phrases et la complexité lexicale.

L'intérêt pour la lisibilité s'est par la suite essoufflé, jusqu'au développement de méthodes de traitement automatique des langues et de méthodes d'apprentissage qui ont contribué à raviver l'intérêt autour du sujet. L'apparition de techniques d'apprentissage plus novatrices pour l'époque a permis d'exploiter une plus grande diversité de variables et de trouver des alternatives à des solutions déjà existantes. En 2005, Collins-Thompson et Callan ont mené une étude visant à évaluer la lisibilité de documents Web à partir de modèles de langage statistiques [Collins-Thompson and Callan, 2005]. Un modèle de langage statistique sert à modéliser la distribution de séquences de mots ou de symboles dans une langue, et est construit à partir d'exemples (des phrases ou des documents complets par exemple) d'où sont extraits des informations comme la fréquence relative des mots. Cette approche permet de mettre en avant des schémas d'utilisation du langage, le modèle et le type de schéma explicités dépendront de l'ensemble de documents utilisés pour l'apprentissage. Dans cette étude, par exemple, c'est le schéma d'utilisation des mots en fonction du niveau scolaire qui est mis en avant. Dans cette même étude, Collins-Thompson et Callan ont utilisé un type de modèle statistique simple : dans ce contexte le 1-gram est une liste de mots associés à leur probabilité $P(w)$ d'être utilisé dans un document avec un niveau scolaire spécifique (voir figure 2.1).

En connaissant la probabilité que chaque mot apparaisse en fonction du niveau scolaire ou grade d'un document, il est possible de prédire la lisibilité

Type *w*	Grade 1 *P*(*w*)	Grade 5 *P*(*w*)	Grade 12 *P*(*w*)
the	0.06000	0.07000	0.08000
a	0.06000	0.05000	0.06000
red	0.00080	0.00040	0.00020
ball	0.00010	0.00005	0.00001
was	0.01000	0.01000	0.02000
perimeter	0.00005	0.00030	0.00005
optimal	0.000001	0.00001	0.00010

Note. *P*(*w*) denotes the probability of type *w* in the model.

FIGURE 2.1 – Modèle de langage 1-gram très simplifié pour les niveaux scolaires 1, 5 et 12 en anglais (issu de [Collins-Thompson and Callan, 2005])

d'un texte, la lisibilité étant indiquée à l'aide du niveau scolaire attribuée au document Web. L'étude conclut à l'efficacité d'une telle approche en confrontant les prédictions obtenues à celles issues des formules de lisibilité traditionnelles, ce qui marque un tournant dans les méthodes de calcul de la lisibilité.

2.2.2.2 Mesures de lisibilité pour le français

Dans son article « Lisibilité et compréhension », publié en 1980, Georges Henry présente les variables entrant en jeu dans sa formule de lisibilité pour le français [Henry, 1980], destinée au traitement automatique de textes :

- le nombre de mots par phrase,
- le Type Token Ratio : le rapport entre le nombre de mots différents et le nombre total de mots du texte,
- le pourcentage de mots qui ne sont pas présents dans la liste de Gougenheim [Gougenheim et al., 1964],
- le pourcentage de pronoms personnels,
- le nombre d'indicateurs du dialogue, un indicateur du dialogue pouvait être un point d'exclamation, des guillemets ouvrant un dialogue ou encore des prénoms utilisés seuls,
- le pourcentage de noms présents dans une liste de noms concrets.

Dans la mouvance d'utilisation d'approches plus « modernes » pour estimer la lisibilité, François a par exemple testé des régressions linéaires multiples, des méthodes ensemblistes (le *bagging* et le *boosting*) pour réaliser une estimation automatique de la difficulté des textes [François, 2009]. Les régressions linéaires permettent ici d'expliciter la relation mathématique existant entre des variables

considérées et la lisibilité. Le bagging et le boosting permettent de combiner les prédictions issues de plusieurs modèles pour réaliser une nouvelle prédiction plus précise. Vingt variables ont été testées par François pour construire ses modèles de prédiction :

- le nombre moyen de lettres par mot,
- la longueur moyenne des phrases,
- les indicateurs du dialogue de Henry (qui correspondent aux points d'exclamation, aux guillements ouvrant un dialogue et au nombre de prénoms employés seuls [Henry, 1980]),
- la fréquence lexicale (mesurée avec un modèle de langage),
- onze variables liées à la complexité verbale.

En 2012, François a testé plus d'approches statistiques (en incluant notamment les machines à vecteurs supports et des arbres de classification) et en utilisant de nouveaux prédicteurs [François and Fairon, 2012]. Il a ajouté des paramètres syntaxiques, mais aussi des paramètres en lien avec les informations sémantiques. Ceci a montré que les **acquis** d'une personne dans une langue donnée, particulièrement le **vocabulaire** et la **grammaire**, sont très importants pour la compréhension écrite.

Ces études autour de la lisibilité ont permis de mettre en avant **l'influence de la syntaxe et la complexité du vocabulaire** sur la difficulté de compréhension de textes écrits. Les diverses formules de lisibilité, mais aussi les modèles statistiques de langage utilisés pour prédire la lisibilité, que ce soit pour l'anglais et le français, font entrer en jeu des variables liées à la **complexité syntaxique** (longueur des phrases, longueur des mots, indicateurs du dialogue) et/ou à la **complexité morphologique** (complexité verbale) et/ou des variables liées à la **complexité lexicale** (fréquence lexicale).

2.3 Compréhensibilité de l'oral

2.3.1 Influence des aspects linguistiques

Les facteurs qui influent sur la compréhension orale ont été étudiés pour les locuteurs natifs (L1) et les locuteurs non-natifs (L2). Comme pour la compréhension écrite, que ce soit pour la compréhension orale en L1 ou en L2, les **compétences linguistiques** jouent un rôle majeur [Buck, 2001, Anderson, 2005]. En effet, l'acquisition de compétences lexicales et grammaticales suffisantes est obligatoire pour pouvoir décoder les mots d'une phrase et leur inférer un sens [Carrow-Woolfolk, 1999]. Il est possible d'établir un parallèle entre l'influence du vocabulaire et de la grammaire sur la lisibilité et leur influence sur la compréhension orale. La capacité à décoder les phrases parlées va dépendre des connaissances acquises par la personne qui écoute en termes de vocabulaire et de grammaire.

En ce qui concerne l'impact de la complexité du vocabulaire sur la compréhension orale, Nation a établi une relation entre les compétences lexicales et la compréhension orale [Nation, 2006]. Il a notamment permis de montrer qu'il existe une forte corrélation entre la difficulté d'éléments textuels et des paramètres en lien avec le vocabulaire comme la fréquence lexicale. Nissan et al. ont confirmé cette observation, en menant une étude qui a abouti à la conclusion que la **fréquence lexicale** était liée à la difficulté générale d'un élément textuel (texte, phrase...), en particulier parce que l'augmentation du nombre de mots peu fréquents augmente la difficulté à comprendre un élément [Nissan et al., 1995]. Ces études démontrent l'importance du vocabulaire en tant que facteur de difficulté pour la compréhension orale. Il est d'ailleurs intéressant de noter qu'il a été identifié comme tel par des apprenants L2 [Goh, 1999].

Pour ce qui est de l'influence de la complexité grammaticale sur la compréhension orale, Chaudron avait avancé que la **complexité syntaxique** pouvait influencer la compréhension orale en L2 [Chaudron, 1983]. Rupp et al. ont montré que la **longueur moyenne des phrases** d'un passage textuel peut augmenter la difficulté de compréhension, car les phrases longues sont plus complexes d'un point de vue syntaxique [Rupp et al., 2001]. S'il est possible de conclure que la syntaxe joue un rôle dans la compréhension orale, aucune conclusion ne peut être faite quant à l'influence de la morphologie sur la compréhension orale.

2.3.2 Influence des aspects cognitifs

Les **aspects cognitifs** ont un rôle prépondérant dans la compréhension orale pour les natifs et les non-natifs. Un auditeur doit développer un ensemble de **stratégies cognitives** pour pouvoir mener à bien une tâche de compréhension orale. Vandergrift a insisté sur l'importance des stratégies compensatoires (comme l'utilisation d'informations contextuelles, visuelles ou du sens commun) pour une compréhension efficace de messages oraux [Vandergrift, 2007]. Comme les stratégies compensatoires dans la langue native seront utiles pour la compréhension orale de langues non natives, il est important de les développer dans sa langue native, pour pouvoir les exploiter avec les langues étrangères. La mémoire de travail, qui est la capacité de stocker et de manipuler des informations, est également cruciale que ce soit pour la compréhension de texte [Daneman and Merikle, 1996], les connaissances lexicales et grammaticales [Stokes and Klee, 2009, Robinson et al., 2003] ou la compréhension orale [Florit et al., 2013].

La mémoire de travail permet de stocker, de concaténer et d'articuler des données en entrée simultanément durant tout le processus d'écoute. Une mémoire de travail développée est utile lors de tâche de compréhension orale, puisque, contrairement à la tâche de compréhension écrite, il n'est pas possible de revenir en arrière quand nous le souhaitons, pour être sûrs de capter toutes les informations importantes de l'énoncé. Comme un texte isolé ne se suffit parfois pas à lui-même pour que l'auditeur puisse en saisir tout le sens, et parce

des informations peuvent manquer si l'auditeur manque de connaissances sur le fond, des stratégies supplémentaires doivent être développées par l'auditeur pour comprendre une situation. Il doit pouvoir :

- identifier les incohérences (suivi de compréhension),
- donner une signification à tout ce qui n'est pas dit de façon explicite dans le texte (inférence),
- interpréter la manière de raisonner et d'agir de ses pairs (théorie de l'esprit).

En résumé, avoir un certain niveau de raisonnement est nécessaire pour comprendre l'intention et l'attitude des locuteurs [Perfetti and Stafura, 2014, Kim and Phillips, 2014], cela permet à l'auditeur d'avoir plus d'indices sur l'interaction en cours.

La combinaison de toutes ces capacités cognitives est indispensable pour réussir une tâche de compréhension orale, du décodage à l'interprétation de ce qui a été entendu.

2.3.3 Influence de la dimension affective

En plus de la dimension linguistique et cognitive, un autre aspect peut affecter la compréhension orale : la **dimension affective**. Dans le cadre éducatif, les tâches de compréhension orale peuvent être stressantes pour les étudiants, parce qu'ils sont confrontés à la peur d'échouer. Cela peut avoir des conséquences négatives sur la réussite de la compréhension. Elkhafaifi et Aneiro ont mis en relief l'influence de l'**anxiété** des étudiants et de l'environnement sur la réussite de la compréhension orale [Elkhafaifi, 2005, Aneiro, 1990]. Noro a expliqué que la difficulté globale de la compréhension orale peut amener des émotions négatives comme le manque d'attention ou le manque de confiance en soi [Noro, 2006]. Si ces émotions négatives ne sont pas gérées correctement par les enseignants et les apprenants, la compréhension orale sera plus difficile. Diminuer l'anxiété des étudiants, en développant par exemple des **méthodes affectives** (comme l'entraide entre étudiants ou la mise en place d'outils pour faciliter la compréhension), peut augmenter la capacité des apprenants à comprendre la parole [Kurita, 2012].

2.3.4 Influence de la prosodie

Des études sur la compréhensibilité de contenus audio ont été menées dans plusieurs domaines. Une partie des recherches est dédiée à l'enseignement et se concentre souvent sur la recherche de ce qui rend la compréhension orale difficile pour les apprenants en langue. D'autres études sont consacrées à la **qualité du signal audio** ou à la **qualité de la production de la parole** des locuteurs : par exemple, pour s'assurer que des messages d'urgence soient émis de façon claire dans les lieux publics, ou améliorer la qualité du signal audio dans les milieux où l'échange de messages à l'aide d'appareils audio est obligatoire (cas des

pilotes qui communiquent avec les personnes basées dans les tours de contrôle).

La façon dont un message est délivré est importante pour la compréhension orale. Les enseignants interrogés dans l'étude de Boyle considèrent que la façon dont un message est produit (sa clarté, sa qualité, l'accent...) joue un rôle dans la compréhension orale [Boyle, 1984]. Cela nous amène à nous intéresser à la **prosodie**. La prosodie concerne tout ce qui touche **au rythme, à l'accentuation, à l'intonation** de la parole.

Pour l'anglais, **l'accentuation et l'intonation** peuvent jouer un rôle dans la compréhension d'un message [Wong and Waring, 2010]. Dans certaines langues, l'accentuation peut être utilisée pour mettre en avant les mots importants d'un énoncé.

Le **débit de parole** joue un rôle important sur la compréhensibilité de la parole. Le débit de parole est ce qui est perçu par l'humain quand nous considérons qu'un locuteur parle vite ou lentement. Goh et Hayati ont observé qu'un débit de parole rapide pouvait être une source de difficulté globale pour les étudiants, d'autant plus s'ils ne sont pas habitués à entendre la langue [Goh, 1999, Hayati, 2010]. Les préférences de débit de parole peuvent être liées à d'autres facteurs, comme le niveau de l'auditeur dans la langue apprise, la prononciation du locuteur [Zhao, 1997]...

Chang et Read ont étudié dans l'impact des **accents** peu familiers sur la compréhension [Chang and Read, 2008]. Par exemple, une personne étant habituée à entendre de l'anglais américain pourrait avoir plus de difficultés à comprendre l'anglais prononcé par un indien. Leurs études ont conclu que la compréhension est seulement affectée si l'auditeur n'a pas été souvent confronté à un certain accent. Ce problème impacte tout autant les natifs que les non-natifs d'une langue cible : Ikeno et Hansen, mais aussi Weil, ont démontré que les locuteurs natifs comprenaient mieux lorsqu'ils étaient habitués à l'accent d'une personne [Ikeno and Hansen, 2006, Weil, 2001].

Les **pauses et les hésitations** (aussi appelées **disfluences**) sont des phénomènes qui se rencontrent souvent dans des situations de communication, notamment quand il s'agit de parole spontanée (c'est-à-dire pas préparée), même si dans les conversations suivant un script (films, pièces de théâtre), il est possible d'imposer les pauses et hésitations pour moduler l'intention du message. Les pauses peuvent être silencieuses ou remplies et constituent une interruption dans la parole, que ce soit au sein même d'une phrase ou entre deux interventions successives. L'hésitation peut se présenter sous plusieurs formes, il peut par exemple s'agir de la répétition complète d'un mot ou de la répétition d'une partie de ce mot, mais cela peut également être représenté par l'insertion de marqueurs d'hésitation comme « euh » en français ou justement par des pauses. Les pauses et hésitations contribuent à modifier le rythme de la production de la parole et peuvent avoir des effets différents sur la compréhension pour les natifs

et les non-natifs. Pour les auditeurs natifs et non-natifs de niveau avancé, les hésitations et les pauses peuvent faire office d'informations supplémentaires sur le message entendu, en mettant l'emphase sur certains mots ou les intentions des locuteurs [Corley and Hartsuiker, 2003]. Un interlocuteur peut choisir de mettre l'emphase sur une information importante en insérant volontairement une pause avant cette information. Le fait de répéter plusieurs fois un même mot lorsque nous hésitons peut constituer un indice pour dire qu'il s'agit d'une information importante. Cependant, les effets bénéfiques des pauses et des hésitations sont nuancés par d'autres études disant qu'elles peuvent également être source de gêne chez l'auditeur. Si Blau admet que les pauses peuvent amener de l'aide à des auditeurs non-natifs [Blau, 1991], elle a également noté que cela pouvait être une source de confusion pour les auditeurs non-natifs ayant un niveau peu élevé dans la langue cible, ce qui peut s'expliquer par le fait que la concentration est plus axée sur le décodage de la parole quand nous débutons dans l'apprentissage d'une langue et que l'ajout de pauses peut nuire à ce processus de décodage. Quand nous sommes d'un niveau plus avancé, l'attention peut être portée sur des informations non verbales pouvant amener du contexte supplémentaire. Voss a un avis plus tranché, concluant dans une première étude [Voss, 1979] que la présence de pauses et d'hésitations dans la parole spontanée amenait des barrières perceptuelles aux apprenants de langue seconde. Dans une seconde étude, où il a étudié les erreurs de perceptions faites par des personnes qui avaient pour langue native l'allemand sur la compréhension orale en allemand puis en anglais, qui contenaient un grand nombre d'hésitations, il constate que percevoir correctement la parole dépend des stratégies cognitives adoptées pour décoder et reconstruire le message [Voss, 1984]. La maîtrise de ces stratégies évolue au fur et à mesure de la maîtrise de la langue, ce qui veut dire que plus une personne aura un niveau avancé dans un langage moins elle sera perturbée par les pauses et les hésitations.

Ainsi, en fonction des cas, les pauses et les hésitations peuvent être bénéfiques pour la compréhension orale ou distraire les auditeurs, selon leur niveau dans la langue.

Pour résumer, cette section nous a permis de voir que la réussite de la compréhension orale en L1 et en L2 dépend autant d'aspects linguistiques que des mécanismes cognitifs développés par l'auditeur. La gestion de l'anxiété d'un auditeur par rapport la tâche de compréhension orale peut également influencer le succès de cette tâche.

De même, la prosodie et les différents facteurs qui la composent (le débit de parole, l'intonation, l'accentuation, les pauses, les disfluences...) ont une influence sur la compréhension de contenu audio.

2.4 Documents audiovisuels et multimodalité : quels impacts sur la compréhensibilité ?

Les études réalisées sur les facteurs influençant la compréhension orale de documents audiovisuels sont moins nombreuses que celles réalisées autour de la lisibilité ou la compréhension orale de documents audio. Dans le domaine de l'apprentissage des langues, il s'agit soit de déterminer les sources de difficulté ou de facilitation dans la compréhension des vidéos, soit de comparer les supports audio et les supports audiovisuels pour savoir quel type de support est le plus adapté. Nous pouvons également trouver quelques études concernant la charge cognitive amenée par les sous-titres, elles portent notamment sur l'influence de la quantité d'informations reçues sur la compréhension.

2.4.1 Modalité audio et conditions sonores « réalistes » : influence de l'intelligibilité de la parole

L'environnement sonore peut être perçu comme une source de difficulté globale. Cela a été mentionné dans l'étude de Boyle par les professeurs interrogés [Boyle, 1984]. L'objectif de cette étude était de lister l'ensemble des facteurs qui avaient une influence sur la compréhension orale en L2, du point de vue des enseignants et des étudiants. Adank et Larsby expliquent que les locuteurs natifs peuvent aussi avoir des problèmes pour comprendre la parole si les conditions d'écoute ne sont pas idéales [Adank et al., 2009, Larsby et al., 2005]. Les effets de mauvaises conditions sonores sur la compréhension orale sont encore plus marqués chez les auditeurs non-natifs, parce qu'ils doivent faire face aux conditions d'audition difficiles et à leurs propres connaissances de la langue écoutée : ils vont être plus sensibles à l'augmentation du bruit ambiant, le brouhaha par-dessus la parole ou la réverbération. Les auditeurs bilingues, en revanche, peuvent être moins affectés par les conditions d'écoute [Lecumberri et al., 2010]. Dans les contenus audiovisuels pédagogiques, les conditions sonores sont contrôlées pour que la parole produite soit intelligible au maximum pour l'apprenant. Ce n'est pas forcément le cas pour le contenu audiovisuel authentique. Si nous prenons le cas de films d'action par exemple, les bruits liés à la scène d'action peuvent recouvrir la parole. La modalité vidéo peut amener des informations qui permettent de compenser la diminution d'intelligibilité que peuvent amener les conditions sonores liées au contexte de la situation de communication.

2.4.2 Modalité vidéo : influence des indices visuels

Lorsque nous communiquons, notre langage verbal s'accompagne d'un langage non verbal : les expressions faciales, mais aussi les gestes ou encore la posture... Des études orientées sur l'apprentissage des langues ont montré que la gestuelle, mais aussi les expressions faciales étaient un apport dans la compréhension d'une situation de communication. Cette constatation va de pair avec l'idée avancée plus tôt selon laquelle l'apprentissage de la langue repose sur des

connaissances linguistiques, mais aussi extralinguistiques.

Il a été démontré que la gestuelle (dans le sens du mouvement du corps y compris les expressions faciales) joue un rôle crucial dans la communication humaine et est bénéfique pour parler et apprendre les langues [Kellerman, 1992, Goldin-Meadow and Alibali, 2013]. Harmer a émis l'opinion que les supports vidéos étaient bénéfiques pour les apprenants en langue étrangère, grâce à l'accès aux expressions faciales, aux gestes et aux autres indices visuels [Harmer, 2007]. Dahl et Ludvigsen ont fait une étude qui a prouvé que les gestes facilitent la compréhension orale des langues natives et étrangères : même si les non-natifs n'utiliseront pas les informations gestuelles de la même manière que des natifs, cela les aidera à mieux comprendre ce qui est dit [Dahl and Ludvigsen, 2014]. Par exemple, la gestuelle permet de fournir des stimuli concernant l'état émotionnel d'une personne (un exemple serait une personne qui croiserait les bras et froncerait les sourcils quand elle est mécontente), mais la gestuelle peut aussi permettre d'appuyer ce qui est dit (en le mimant par exemple). L'étude de Sueyoshi et Hardison a montré que, lorsque le document présenté à des apprenants est un tutoriel, l'accès à la vidéo avec des gestes et des indices visuels aidait les apprenants en langue seconde à mieux comprendre les tâches à réaliser [Sueyoshi and Hardison, 2005]. Les gestes pouvant aider à accompagner la parole par le mouvement pour illustrer la tâche à accomplir. Pour les apprenants d'un niveau plus avancé, ce sont les expressions faciales qui les aident le plus, les apprenants avec un niveau débutant ou débutant/intermédiaire ont besoin simultanément des expressions faciales et des gestes.

2.4.3 Modalité texte : apport des sous-titres

L'étude de l'influence des sous-titres sur la compréhension de vidéos est un sujet qui a son intérêt dans l'apprentissage des langues. Plusieurs études ont été faites pour déterminer si les sous-titres étaient bénéfiques ou gênants pour la compréhension des vidéos chez les apprenants de L2. Le protocole expérimental consistait à comparer les performances des étudiants L2 exposés aux vidéos avec les sous-titres aux performances d'étudiants de L2 ayant vu les mêmes vidéos, mais sans sous-titres [Hayati and Mohmedi, 2011].

2.4.3.1 Apport pour la compréhension

Il existe deux types de sous-titres utilisés pour l'apprentissage des langues : les sous-titres de L1, où les sous-titres sont dans la langue native des apprenants, et les sous-titres de L2, où les sous-titres sont dans la langue apprise. La plupart des études ont prouvé que les sous-titres (L1 ou L2) ont un effet positif sur la compréhension pour les apprenants en langue seconde : ils ont de meilleurs résultats aux tests de compréhension [Perez et al., 2013]. Les sous-titres de L1 sont plus adaptés pour les apprenants de niveau débutant, car ils ont moins de connaissances de la langue cible, cependant les sous-titres de L2 ont des effets plus marquants, car ils apportent une redondance entre ce qui est dit et ce

qui est lu [Hayati and Mohmedi, 2011]. Une étude de Mitterer et McQueen a montré que les sous-titres dans la langue native étaient nuisibles au processus de compréhension pour les apprenants non-natifs, alors que les sous-titres en langue étrangère sont une source d'aide [Mitterer and McQueen, 2009]. D'autres études ont porté sur les effets des sous-titres dans d'autres domaines que celui de l'apprentissage des langues. Ces études ont également conclu que les sous-titres aidaient à mieux comprendre le contenu de la vidéo, qu'ils soient dans la langue native ou non-native [Markham et al., 2001].

2.4.3.2 Source de surcharge cognitive ?

En général, les études ont montré que les sous-titres étaient bénéfiques à la compréhension. Mais, les sous-titres étant du texte venant s'ajouter à la vidéo, il est possible que l'addition de ces informations aux autres informations inhérentes aux images de la vidéo amène une charge cognitive supplémentaire aux spectateurs et dégrade ainsi la compréhension du contenu. Les études sur le sujet ont donné lieu à des conclusions mitigées. Kalyuga *et al.* ont fait une étude semblant appuyer le fait que le sous-titre était une source de surcharge cognitive [Kalyuga et al., 1999]. Mais des études plus récentes sur l'impact des sous-titres sur la charge cognitive ont montré que les sous-titres n'amenaient pas de surcharge [Kruger et al., 2013]. Il serait compliqué de conclure sur cette question s'il n'avait pas été constaté que la vidéo constituait un apport bénéfique pour la compréhension orale.

2.4.4 Document audio *vs* document audiovisuel

L'accès simultané aux images, au son et à du contenu textuel contribue à améliorer la compréhension : la multimodalité (c'est-à-dire l'exposition simultanée à plusieurs modalités) tend à améliorer la compréhensibilité d'un contenu. Si l'avantage apporté par les sous-titres a déjà été mis en avant, cela peut également être conforté par le fait que toutes les études qui avaient pour objectif de comparer les résultats d'un enseignement ayant recours uniquement à des documents audio et d'un enseignement ayant recours à des documents audiovisuels ont démontré que les documents audiovisuels ont des effets plus probants sur la compréhension des étudiants en langues étrangères [Jones and Plass, 2002, Ali Batel, 2014, Yasin et al., 2017]. Les élèves ayant été exposés à du contenu audiovisuel auront de meilleurs résultats aux tests de compréhension que des élèves qui ne se sont vus présenter que des documents audio.

Pour résumer, dans cette section, de nouveaux facteurs influençant la compréhension de contenu audiovisuels ont été identifiés : les indices visuels, la présence de sous-titres, mais aussi le type de sous-titres (en langue native ou en langue apprise). Elle a aussi permis de montrer que l'enseignement à l'aide de contenu audiovisuel permettait aux étudiants d'obtenir de meilleurs résultats en terme de compréhension, ce qui veut dire que la compréhensibilité est également influencée par les modalités disponibles.

2.5 Bilan

Dans ce chapitre, nous avons exploré l'ensemble des phénomènes répertoriés dans la littérature qui peuvent influencer la compréhensibilité des contenus présentés à des apprenants. Nous avons étudié la question en considérant aussi bien les documents écrits, les documents audio, et les documents audiovisuels en nous intéressant plus particulièrement, dans ce dernier cas, à l'apport des différentes modalités qui les constituent. Plusieurs familles de facteurs ont ainsi pu être identifiées, notamment la complexité lexicale, la complexité grammaticale et l'intelligibilité de la parole comme rappelé dans les tableaux 2.1 et 2.2.

TABLE 2.1 – Principaux facteurs de complexité du vocabulaire et de complexité grammaticale mentionnés dans la littérature

Complexité	Facteurs notables	Références
Lexicale	fréquence lexicale	[Lively and Pressey, 1923] [Dale and Chall, 1948] [Nation, 2006]
	diversité lexicale	[Lively and Pressey, 1923] [Henry, 1975]
Grammaticale	longueur des phrases	[Sherman, 1893] [Flesch, 1948] [Rupp et al., 2001]
	indicateurs du dialogue	[Henry, 1975]
	complexité morphologique	[François, 2009]

TABLE 2.2 – Principaux facteurs affectant l'intelligibilité de parole identifiés dans la littérature

Dimension	Facteurs notables	Références
Prosodie	accentuation et intonation	[Wong and Waring, 2010]
	disfluences	[Corley and Hartsuiker, 2003] [Blau, 1991] [Voss, 1979]
	débit de parole	[Goh, 1999] [Hayati, 2010]
Accents	accents « non standards »	[Chang and Read, 2008]
Mauvaises conditions sonores	bruits, brouhaha, réverbération	[Boyle, 1984] [Adank et al., 2009] [Larsby et al., 2005]

Si tous ces facteurs ont été identifiés dans la littérature comme ayant une influence sur le niveau de compréhensibilité, il n'existe cependant pas de collections de données annotées qui mettraient en jeu l'ensemble de ces phénomènes et permettant de démontrer et de quantifier cette influence. Pour mettre en place

la mesure objective du niveau de compréhensibilité au cœur de cette thèse, il s'est avéré indispensable de constituer un corpus de contenus audiovisuels et de collecter des annotations auprès d'experts ayant une bonne connaissance des problématiques autour de la compréhensibilité. Même si plusieurs facteurs ont été identifiés grâce à l'état de l'art présenté dans ce chapitre, tous ne seront pas considérés pour l'élaboration du corpus. L'étude se concentrera sur les facteurs inhérents aux documents audiovisuels qui ont une influence sur la compréhensibilité. Les facteurs directement liés à la dimension affective et la dimension cognitive seront ignorés, car nous nous intéressons à des facteurs quantifiables et indépendants de l'individu qui est exposé à la vidéo.

La constitution de ce corpus et les annotations subjectives collectées font l'objet du chapitre suivant.

Chapitre 3

Création et analyse d'un corpus de référence (ESCAL) : étude subjective de la compréhensibilité pour l'apprentissage des langues

Sommaire

3.1 Introduction

Notre objectif est de constituer un corpus qui associe contenus audiovisuels authentiques et évaluation de leur niveau de compréhensibilité. Un tel corpus nous permettrait de disposer d'un corpus de référence fondé sur des évaluations subjectives, qui permettraient ensuite d'évaluer les performances d'un système de prédiction de niveau de compréhensibilité de documents audiovisuels authentiques dans un second temps. Une contribution majeure de ce travail de thèse a été de constituer le corpus qui répond à ce besoin. Nous avons rassemblé une collection d'extraits de films, et avons demandé à des experts d'évaluer un certains nombre de facteurs qui influent sur la compréhensibilité. Ces facteurs ont été choisis en nous basant sur les résultats de l'étude présentée au chapitre précédent et notamment sur les familles de facteurs liés au vocabulaire, à la grammaire, à l'intelligibilité de la parole en considérant l'apport de chaque modalité ou combinaison de modalités.

Ce chapitre présente donc la mise en place du corpus que nous avons nommé ESCAL (Évaluation Subjective de la Compréhensibilité pour l'Apprentissage des Langues), ainsi que la procédure de collecte d'annotations. Nous complétons ce chapitre en présentant également une analyse des annotations collectées, à la fois d'un point de vue quantitatif en comparant et en analysant les évaluations des annotateurs et qualitatif en analysant les commentaires qui accompagnent les annotations.

3.2 Création du corpus ESCAL

3.2.1 Choix des documents : extraits de films

Dans le premier chapitre, nous avons montré que les doucments audiovisuels étaient un support de choix pour permettre aux enseignants de présenter différentes situations de communication à leurs apprenants et d'appliquer l'approche communicative en leur inculquant à la fois des connaissances langagières mais les codes sociolinguistiques de la langue enseignée.

Le fait de pouvoir entendre et voir simultanément ce qui se passe lors d'une situation de communication favorise l'apprentissage en parallèle de ces deux aspects. La dimension authentique constitue un avantage supplémentaire, car non seulement les conditions de la situation de communication se rapprochent le plus possible de ce que les apprenants seront amenés à rencontrer dans des conditions réelles, mais en plus, il y a un côté motivant lié au fait qu'il s'agit de documents qui ont été conçus originellement pour les natifs de la langue. Il s'avère que les films regroupent les deux éléments essentiels : ils contiennent des situations de communication et ils sont attrayants, mais surtout motivants pour les apprenants. Ensuite, les films ont des contenus très diversifiés : il y aura beaucoup de variations dans le vocabulaire, la grammaire, l'intelligibilité de la parole en fonction du type de film, de son sujet, mais aussi des acteurs, ce qui rendra le contenu plus ou moins difficile à appréhender selon le niveau des

apprenants dans la langue cible. De plus, les films et plus précisément les extraits de films font partie de supports fréquemment exploités, mais aussi appréciés par les enseignants pour la construction leurs cours [Kotula, 2014].

Pour la constitution du corpus ESCAL, nous nous sommes orientée vers une sélection d'extraits de documents de fiction. L'objectif industriel de ce travail de thèse était de proposer en premier lieu des outils dédiés à des enseignants de français langue étrangère. Le choix s'est porté sur des films français.

Quinze films ont été retenus, de manière à ce qu'il y ait une diversité de genres, d'époques, de qualités audiovisuelles restituées, mais aussi de registres de langue.

3.2.1.1 Segmentation en extraits : approche basée sur l'interaction

Les films sont intéressants pour l'apprentissage d'une langue : en effet, ils contiennent des interactions qui s'inscrivent dans un contexte proche de ce qui peut être rencontré par les apprenants dans la réalité. Bien souvent, ils ne sont pas exploités en entier : seules des situations de communication correspondant à des interactions sont utilisées par la suite.

Pour se faire, il a été nécessaire d'abord de segmenter le contenu en *Zones d'Interaction Potentielles* avant d'identifier parmi ces zones celles qui correspondaient réellement à des interactions. Plusieurs règles ont été définies pour automatiser et simplifier ce processus de segmentation.

Nous avons observé que dans les documents de fiction, une situation de communication ne contenait pas de trop longs silences. Une première règle de segmentation élémentaire a été appliquée lorsque le silence excédait une certaine durée.

Pour cela, nous nous sommes basée sur les sous-titres de films : nous avons utilisé le corpus OPUS[1], qui est une collection de sous-titres multilingues issus de Open Subtitles composée de sous-titres réalisés par un public de fans. Si en termes de qualité du texte ces sous-titres ne sont pas toujours fiables (fautes d'orthographe, erreurs de typographies...), ils ont l'avantage d'être synchronisés correctement suivant la *timeline* de la vidéo correspondante. Les proportions de temps de parole et temps de silence restent donc exploitables.

En se référant au temps de pause entre deux sous-titres successifs et une valeur seuil, une nouvelle *Zones d'Interaction Potentielles* est créée. Des valeurs seuils allant de quatre à dix secondes ont été testées, et nous avons trouvé qu'une valeur de sept secondes permettait d'obtenir des zones intéressantes : avec ce seuil, les interactions potentielles contiennent moins de ruptures qui amènent des incohérences (présence de sous-titres qui ne semblent pas en lien avec les sous-titres précédents, zones constituées d'un seul sous-titre ne pouvant donc pas correspondre à une interaction...).

1. http://www.opensubtitles.org/

3.2.1.2 Critères de sélection des extraits

Une fois ce premier niveau de segmentation en *Zones d'Interaction Potentielles* obtenu, il a été nécessaire de ne garder que les zones correspondant effectivement à des zones d'interaction. Cette étape a été réalisée manuellement. Les extraits ont été choisis de telle façon qu'ils contiennent des interactions suffisamment longues entre les personnages, pour que ce soit exploitable. Ainsi, les segments d'une durée inférieure à 10 secondes ont été éliminés d'office car des interactions aussi courtes ne sont pas (ou peu) exploitées par les professeurs en cours de langue étrangère. Les scènes d'action ont été exclues de notre sélection, car notre objectif est d'étudier des situations de communication et que nous avons considéré qu'une scène d'action serait trop parasitée (par le bruit, beaucoup de mouvements...) pour être pertinente.

Après cette première phase de sélection, le choix et la délimitation des interactions ont été faits en reprenant la définition d'interaction que donne Traverso [Traverso, 1996] faisant état d'une cohérence en termes :

— d'objectif,
— de cadre spatial,
— de cadre temporel,
— de participants.

Dans le cas où une zone manquait de cohérence dans l'un de ces aspects, à cause par exemple, d'une phrase manquante ou ajoutée lors de la phase de découpage automatique, alors les frontières ont été modifiées manuellement. Si une zone était constituée de quelques répliques sans contexte, elle n'était pas retenue. Si les zones étaient constituées de plusieurs interactions sans pause longue pour les séparer, un redécoupage était fait à la main. À l'issue de cette étape, 300 extraits correspondant à 300 situations d'interaction ont été identifiés parmi l'ensemble des *Zones d'Interaction Potentielles* restantes.

3.2.1.3 Collection à annoter

Pour chacun des films retenus, trois à cinq extraits, répondant à la définition d'une interaction, ont été sélectionnés. Les flux audio et vidéo de chacun d'eux ont été récupérés. La collection à annoter est décrite dans le tableau 3.1.

Les *Zones d'Interaction Potentielles*, construites à partir des sous-titres de Open Subtitles, sont issues d'un travail d'annotateurs amateur (c'est ce que nous appelons le « *fansub* »), donc leur qualité peut parfois être remise en question. Pour se placer dans un cas idéal dans la suite de l'étude, ils ont été retravaillés pour aboutir à la transcription exacte de chaque interaction. En effet, les sous-titres sont parfois simplifiées pour seulement conserver le sens.

Notre corpus est ainsi composé de 55 extraits (2541 secondes). Pour chacun des 55 extraits, les trois flux suivants sont disponibles :

— le texte : il s'agit de la transcription exacte des 55 extraits (7225 mots au total),

— l'audio : la bande sonore seule,
— les images composant le flux vidéo.

Table 3.1 – Extraits du corpus

Titre du film	Année	Nombre d'extraits
Le fabuleux destin d'Amélie Poulain	2001	3
Cyrano de Bergerac	1990	4
Delicatessen	1991	3
Embrassez qui vous voudrez	2002	4
Intouchables	2011	3
La chèvre	1981	4
La folie des grandeurs	1971	3
La gloire de mon père	1990	5
La grande vadrouille	1966	4
Le petit Nicolas	2009	4
Les choristes	2004	4
Les plages d'Agnès	2008	3
Qu'est-ce qu'on a fait au bon Dieu	2014	3
Séraphine	2008	4
Un long dimanche de fiançailles	2004	4

3.2.2 Modes de présentation des extraits

Afin d'étudier l'influence de la complexité du vocabulaire, de la complexité grammaticale, et de l'intelligibilité de la parole sur le niveau de compréhensibilité des documents ainsi rassemblés, mais également pour analyser l'influence des modalités sur la compréhensibilité, les 55 extraits ont été proposés aux participants selon 5 modes de présentations différents impliquant soit une modalité seule, soit une combinaison de modalités.

— texte seul (T) : seule la transcription de l'extrait est présentée, dans ce cas nous étudions uniquement l'influence du vocabulaire et de la grammaire sur la compréhensibilité puisqu'il n'y a pas de parole produite ni d'image disponible,
— audio seul (A) : dans ce cas nous étudions l'influence de la grammaire, du vocabulaire et de l'intelligibilité sur la compréhensibilité de l'oral.
— combinaison de deux modalités : texte et audio (AT),
— combinaison de deux modalités : audio et images (composants le flux vidéo) (AV),
— combinaison des trois modalités : audio, image et texte (AVT).

Chaque combinaison (extrait, mode de présentation) constituera un *document* présenté en vue d'être annoté. Nous disposons donc de 275 documents.

3.3 Annotation du corpus

3.3.1 Protocole d'annotation des extraits de films

3.3.1.1 Dimensions à considérer

Suite à l'état de l'art qui a été réalisé et présenté dans le chapitre précédent, le choix a été de se concentrer sur l'impact de la complexité lexicale, de la complexité grammaticale et de l'intelligibilité de la parole sur la compréhensibilité. Quatre dimensions ont ainsi été étudiées :

- le vocabulaire,
- la grammaire,
- l'intelligibilité,
- la difficulté globale (que nous associons au niveau de compréhensibilité du document dans sa globalité).

À travers la constitution de ce corpus, le premier objectif est de quantifier l'influence des trois premières dimensions sur la difficulté globale perçue. Le second objectif est d'étudier l'influence des différentes modalités audio, vidéo et texte et de leurs combinaisons sur la perception de la difficulté globale par les experts sollicités pour enrichir le corpus ESCAL.

3.3.1.2 Sélection d'un panel de participants

Afin de pouvoir évaluer les matériaux collectés, il fallait l'expertise de personnes famliarisées avec l'étude de la compréhensibilité. C'est pour cela que nous avons fait appel à des enseignants de Français Langue Étrangères, car ils sont souvent confrontés à la problématique de l'évaluation de documents (texte, audio ou audiovisuel) appropriés pour leurs étudiants en terme de difficulté globale.

Pour que le corpus permette de produire des résultats qui sont statistiquement significatifs (et donc pertinents et exploitables), quinze professeurs de FLE ont été recrutés en se basant sur les critères suivants :

- être français natifs pour que l'expérience ne soit pas biaisée par leur propre niveau de maîtrise de la langue française,
- avoir une expérience d'enseignement d'au moins trois ans avec des apprenants de niveaux de français différents. Nous sommes ainsi assurée qu'ils étaient suffisamment sensibles aux variations de compréhensibilité,
- être familiers avec l'utilisation d'extraits de films en cours,
- être (auto-diagnostiqués) normo-entendants pour que leur perception de l'intelligibilité de la parole ne soit pas influencée par des problèmes d'audition.

Les 15 professeurs (13 femmes, 2 hommes) ont entre 27 et 63 ans (moyenne d'âge : 37 ans) et une expérience d'enseignement entre 3 et 40 ans (moyenne : 11 ans, écart-type : 9 ans).

3.3.2 Collecte des annotations

3.3.2.1 Description de l'interface d'annotation

Pour simplifier la consultation des 275 documents et leur annotation, une interface graphique (GUI) en ligne a été développée pour présenter chacun d'eux aux différents experts participants. Cette interface est présentée dans la figure 3.1.

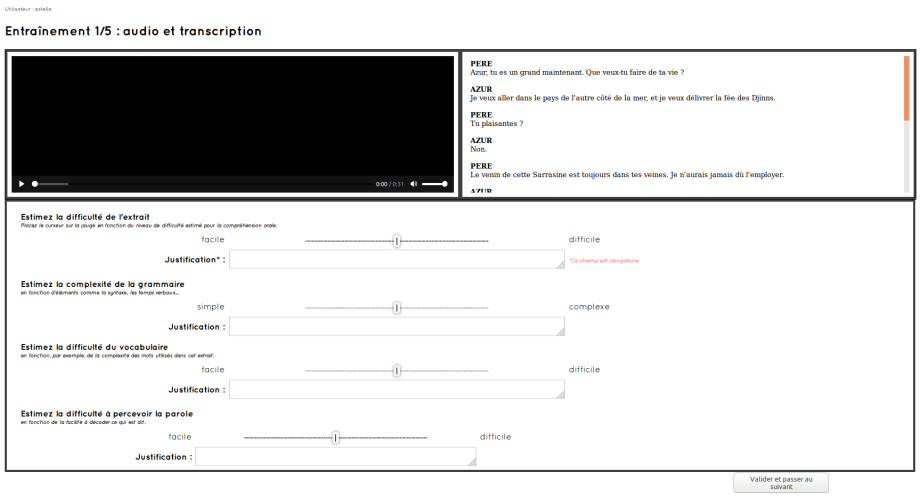

FIGURE 3.1 – Interface graphique pour l'annotation d'extraits de films

En haut à gauche nous avons le lecteur qui permet soit de lire l'audio seul (A) soit la vidéo et l'audio (AV, AVT). À droite il y a le texte qui correspond à la transcription exacte et qui est affiché selon le mode de présentation (T, AT, AVT). Chaque intervention est précédée par le nom du personnage ou locuteur courant. En bas, il y a l'interface d'évaluation où les experts peuvent attribuer un score à l'aide de trois à quatre curseurs selon le mode de présentation (l'intelligibilité ne pouvant pas être évaluée en l'absence de la modalité audio). La position initiale de chaque curseur est au centre. Le choix a été fait de faire annoter les extraits à l'aide de curseurs non gradués de 0 (très facile) à 100 (très difficile). Ce choix a d'abord été fait pour s'assurer que les enseignants ne se basent pas sur une échelle commune dans le domaine de l'enseignement des langues étrangères pour réaliser l'annotation (par exemple l'échelle CECRL du référentiel commun européen [Conseil de l'Europe, 2003] pour l'apprentissage des langues étrangères), mais aussi pour que nous travaillons sur des données continues et pas sur des données discrètes.

Des zones de saisie de texte permettaient aux participants de mettre un commentaire pour justifier l'ajustement de chaque curseur ou nuancer leurs choix. La justification est obligatoire pour l'évaluation de la difficulté globale et facultative pour les autres évaluations.

3.3.2.2 Conditions de collecte

Chaque participant devait analyser 55 documents de façon à traiter chacun des 55 extraits uniquement sous l'un des 5 modes de présentation, soit 11 documents par mode. En les exposant ainsi à chaque mode, ils réalisaient la tâche d'évaluation un nombre équivalent de fois, ce qui assurait qu'ils n'étaient pas influencés par une condition de présentation spécifique. Ainsi, l'étude de l'influence des modalités pouvait être réalisée.

Les enseignants devaient utiliser les curseurs pour noter le document présenté en termes de :

- **difficulté globale :** de 0 - « très facile » à 100 - « très difficile » ;
- **difficulté du vocabulaire :** de 0 - « très facile » à 100 - « très difficile » ;
- **complexité grammaticale :** de 0 - « très facile » à 100 - « très complexe » ;
- **intelligibilité de la parole :** lorsque que la modalité audio est disponible, de « totalement intelligible » à « totalement inintelligible ».

Chaque document a été annoté par exactement trois participants. Cette contrainte a été mise en place pour pouvoir ensuite calculer des accords inter-annotateurs et éviter des cas où deux annotateurs présenteraient un fort désaccord.

Pour s'assurer que l'évaluation n'était pas affectée par un biais lié à la liste des documents présentée aux participants, chaque participant s'est vu attribuer une liste de documents uniques, présentée par la suite dans un ordre aléatoire. Les participants ont réalisé l'expérience en ligne avec leur propre matériel. Ils avaient pour consigne de réaliser l'expérience dans un endroit calme, en utilisant un ordinateur et des écouteurs.

3.3.2.3 Description des annotations du corpus ESCAL

Chaque annotateur a ainsi attribué 19 scores à chaque document traité, un score pour chaque dimension (soit 4 scores pour les modalités comportant de l'audio (A, AT, AV, AVT) et 3 pour la modalité sans audio (T)). Chaque document a été évalué par 3 experts, ce qui représente 57 scores par document soit :

- 15 scores de difficulté globale,
- 15 scores de complexité du vocabulaire,
- 15 scores de complexité grammaticale,
- 12 scores d'intelligibilité.

Le corpus ESCAL comprend en totalité 3135 scores ainsi que 825 commentaires visant à justifier les scores de difficulté globale attribués (275 documents x 3 annotateurs).Les justifications n'étant pas obligatoires pour les scores de complexité du vocabulaire, de grammaire et d'intelligibilité, le nombre de com-

mentaires comptabilisé au total n'est donc pas équivalent selon la dimension considérée. Nous avons ainsi :

— 478 justifications pour le vocabulaire,
— 360 justifications pour la grammaire,
— 418 justifications pour l'intelligibilité.

Un tableau récapitulatif décrivant le corpus ESCAL est donné dans la section Bilan de ce chapitre(Table 3.5).

3.4 Analyse quantitative et qualitative

Pour avoir une connaissance plus fine du corpus ESCAL, nous avons analysé les mesures subjectives collectées. Une analyse quantitative nous a permis de comparer les scores donnés par les différents annotateurs pour voir s'il se dégage une perception commune des quatre dimensions étudiées (cf. section 3.3.1). Ainsi, nous pouvons décrire le corpus à l'aide de statistiques et répondre aux deux objectifs qui ont motivé la création de ce corpus :

1. démontrer (et quantifier) l'influence de la complexité du vocabulaire, de la complexité grammaticale et de l'intelligibilité sur le niveau de compréhensibilité,
2. démontrer l'influence du mode de présentation (et donc des modalités) sur le niveau de compréhensibilité.

L'étude du lien entre les différentes dimensions, se fait à l'aide d'un calcul de corrélation ; des régressions linéaires multiples permettent de prendre en compte en même temps le vocabulaire, la grammaire et l'intelligibilité de la parole pour quantifier leur influence sur le niveau de compréhensibilité (représenté par la difficulté globale). Une comparaison est alors établie entre les scores attribués en fonction du mode de présentation.

Nous avons également procédé à une analyse qualitative. Celle-ci consiste à étudier les commentaires des annotateurs réalisés pendant l'expérience, pour justifier leurs choix quant au placement des curseurs pour chacun des documents qui leur ont été présentés.

3.4.1 Étude des accords inter-annotateurs

Afin de déterminer s'il existe une manière commune de percevoir chacune des dimensions (vocabulaire, grammaire, intelligibilité et difficulté globale), et ce indépendamment du mode de présentation, nous nous sommes intéressée à l'accord inter-annotateurs pour chaque dimension évaluée.

3.4.1.1 Principe de base et méthodes

L'accord inter-annotateurs permet d'évaluer à quel point les annotateurs sont d'accord sur le même élément, ici le score à attribuer à un document. Il

s'agit donc d'étudier la corrélation entre leurs annotations pour savoir s'il existe un lien dans leur manière de les réaliser.

Plusieurs méthodes existent pour calculer cet accord, la méthode à adopter dépend tout d'abord du type de tâche qui a été réalisée. Il peut s'agir :

- d'attribuer un label parmi plusieurs à un document dans ce cas l'accord inter-annotateurs vise à comparer si les annotateurs ont souvent donné le même label au même document ;
- d'une tâche où il faut classer les documents et dans ce cas l'accord inter-annotateurs repose sur la comparaison des classements de tous les annotateurs pour voir à quelle fréquence les documents ont été classés de la même façon ;
- d'une tâche d'attribution d'un score, et dans ce cas pour calculer l'accord inter-annotateurs il faut s'intéresser à la proximité des scores attribués à chaque document.

Dans le cadre de la constitution de ce corpus, nous nous situons dans ce dernier cas.

3.4.1.2 Calcul de corrélation

L'évaluation de l'accord inter-annotateurs dans le cas où nous travaillons sur des données continues peut être réalisée en étudiant la corrélation entre les scores de deux annotateurs. La corrélation sert à mesurer le lien entre deux variables et à voir si elles évoluent « ensemble ». Dans notre cas, la corrélation permet de déterminer si la façon d'annoter des annotateurs est proche et à quel point. Il existe plusieurs coefficients de corrélation : les deux plus connus sont le coefficient de Pearson [Pearson, 1895] et le coefficient de Spearman [Artusi et al., 2002].

Le coefficient de Pearson permet d'évaluer la relation **linéaire** entre deux variables **continues**, cette relation linéaire se traduit par le fait que lorsqu'une variable subit une modification alors cela entraîne une modification proportionnelle sur une autre variable ;

Le coefficient de Spearman étudie la **monotonie** de la relation entre deux variables. La nuance est que nous nous attendons à observer que les données changent ensemble, mais nous n'attendons aucune relation proportionnelle dans les modifications des deux variables [Chok, 2010].

Dans notre problématique, il ne s'agit pas de chercher une relation linéaire entre les scores attribués par les annotateurs, mais de repérer s'ils ont tendance à noter de la même manière. La corrélation de Spearman semble donc plus appropriée.

Approche choisie

La corrélation de Spearman est fondée sur l'étude des rangs, donc pour évaluer l'accord inter-annotateurs nous avons ramené le problème à une comparaison de classements où chaque score correspond à la position du document dans le classement et où nous considérons qu'il peut y avoir des cas d'égalité. Poser cette hypothèse revient à comparer les classements réalisés par chaque annotateur et la corrélation de Spearman peut s'appliquer.

Dans l'étude des corrélations, il faut prendre en compte à la fois :
- **la résistance (valeur absolue du coefficient) :** elle donnera une indication sur la force de la corrélation (une corrélation peut être nulle, faible, moyenne, forte et très forte),
- **le sens (signe du coefficient) :** il déterminera la direction de la relation qui existe entre deux échantillons comparés (si elle existe),
- **la significativité :** étudiée à l'aide d'une valeur seuil appelée *p-value* qui accompagne le coefficient de corrélation, elle indiquera la probabilité que cette valeur de coefficient ait été obtenue par hasard. Nous dirons que les corrélations sont significatives si les p-values des corrélations sont inférieures au seuil $p = 0,05$, qui est une valeur usuelle de seuil [Kennedy-Shaffer, 2019].

Le but du calcul de l'accord inter-annotateurs est de déterminer, ici, si pour chaque dimension considérée, les annotateurs tendent à donner les scores de la même manière, ce qui reflétera alors une vision commune de la manière de percevoir la complexité des documents (vocabulaire, grammaticale, intelligibilité ou compréhensibilité). Nous comparons les scores des annotateurs qui ont annoté des documents en commun.

Dans la suite, nous nous intéressons uniquement aux corrélations significatives pour que l'étude des accords inter-annotateurs soit pertinente.

3.4.1.3 Significativité des corrélations

Pour chaque dimension, nous calculons 105 corrélations de Spearman (ce qui correspond au nombre de paires distinctes d'annotateurs) et observons ainsi un nombre de corrélations significatives représentant un certain pourcentage de corrélations :
- **difficulté globale** : 22 corrélations significatives (soit 21%),
- **complexité du vocabulaire** : 32 corrélations significatives (soit 30%),
- **complexité grammaticale** : 14 corrélations significatives(soit 13%),
- **intelligibilité** : 17 corrélations significatives (soit 16%).

Le faible taux de corrélations significatives pour toutes les dimensions peut s'expliquer par le fait que les annotateurs ne sont pas toujours sensibles aux mêmes éléments pour attribuer les scores.

3.4.1.4 Résistance des corrélations

La figure 3.2 permet de représenter la distribution des coefficients de corrélations significatifs pour chaque dimension évaluée.

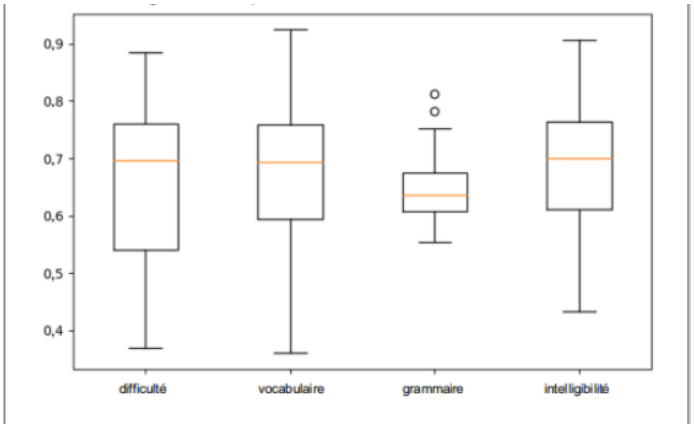

FIGURE 3.2 – Diagramme représentant les coefficients de corrélation de Spearman significatifs pour l'accord inter-annotateurs

Nous constatons un fort accord entre les annotateurs que ce soit pour attribuer les scores de complexité lexicale ou de complexité grammaticale. La dispersion des scores d'intelligibilité reflète que même dans les cas extrêmes, les annotateurs ont un accord moyen sur la perception de l'intelligibilité.

Nous notons en observant la distribution des scores de complexité grammaticale, que les annotateurs s'éloignaient peu de la moyenne pour attribuer leurs scores. Ce phénomène peut refléter une incertitude quant à la façon d'évaluer la complexité de la grammaire. ´ l'inverse, la distribution des scores d'intelligibilité montre que l'ensemble de la plage de notation a été utilisée pour évaluer l'intelligibilité, ce qui signifie qu'elle a été évaluée de manière plus fine.

Dans le cas de la difficulté globale et de la complexité du vocabulaire, les annotateurs peuvent être faiblement d'accord, ce qui veut dire qu'il y a un consensus moins fort quant à la perception de ces dimensions.

Nous avons constaté dans la section 3.4.1.3 une proportion peu élevée de corrélations significatives et donc une part trop importante de sensibilité individuelle de chacun des annotateurs pour généraliser la manière dont chacune des dimensions a été annotée. Cependant, le fait de se pencher sur les cas de corrélations significatives (voir figure 3.2) et de voir que pour chacune des dimensions, il existe en moyenne un accord fort entre les annotateurs met en avant des traits communs dans la façon de percevoir et d'évaluer la complexité du vocabulaire, la complexité grammaticale, l'intelligibilité et la compréhensibilité à travers le niveau de difficulté globale.

Pour la suite, nous allons utiliser les moyennes de scores par documents (couple extrait-mode de présentation) pour essayer d'établir l'existence d'une relation entre vocabulaire, grammaire, intelligibilité et difficulté globale. L'idée est de les quantifier tout en s'affranchissant de cette sensibilité individuelle des annotateurs et en se centrant sur ces aspects communs dans la façon de réaliser les annotations.

3.4.2 Étude des scores attribués par les annotateurs

Pour chaque dimension qui a été évaluée, nous avons calculé la médiane et avons illustré la distribution de l'ensemble des scores (non moyennés) à l'aide d'histogrammes. Ces informations permettent d'observer de quelle façon les annotateurs ont eu tendance à annoter chacune des dimensions.

3.4.2.1 Distribution des scores

En fonction de la plage de valeurs exploitée pour évaluer une dimension, cela nous donne une information quant à la variabilité du corpus, qui sera ensuite utilisé comme corpus de référence. Cela permet aussi de voir si les annotateurs ont fait leurs annotations de façon équilibrée ou s'ils ont plus favorisé un intervalle d'annotation plutôt que d'utiliser l'ensemble de la plage des scores.

Complexité du vocabulaire : la figure 3.3 permet de voir que les annotateurs ont utilisé l'intégralité de la plage de valeurs pour évaluer la complexité du vocabulaire ce qui indique qu'il y avait de la diversité dans la complexité lexicale des documents. La médiane située à 55 semble indiquer que le corpus est équilibré en terme de complexité lexicale. Nous notons une distribution dissymétrique à gauche des scores de vocabulaire, ce qui montre que la plus grande partie des scores attribués est supérieure à 50, ce qui veut dire qu'il y a une légère tendance à noter au-dessus de 50, il y avait donc un peu plus d'extraits considérés comme moyens à difficiles en termes de vocabulaire par les annotateurs ;

Complexité grammaticale : la figure 3.4 montre que des notes de 0 à 100 ont été attribuées aux documents par les annotateurs bien qu'ils se soient globalement peu éloignés de la moyenne pour noter la complexité grammaticale. Cependant, même si toute la plage de valeurs possible a été exploitée, nous constatons que la médiane est basse : elle se situe à 36, ce qui veut dire que les extraits ont été considérés comme n'étant pas forcément compliqués par les annotateurs. Cette tendance de notation se confirme sur l'histogramme : il est dissymétrique à droite et la plupart des scores se situent entre 0 et 50 ;

Intelligibilité de la parole : les scores d'intelligibilité ont été attribués de manière très équilibrée (voir figure 3.5) : la médiane des scores se

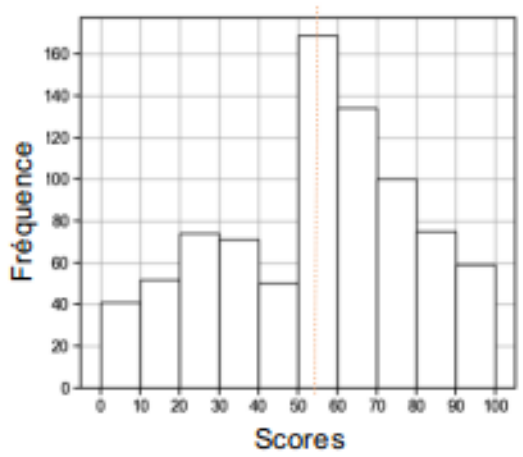

FIGURE 3.3 – Distribution des scores de vocabulaire

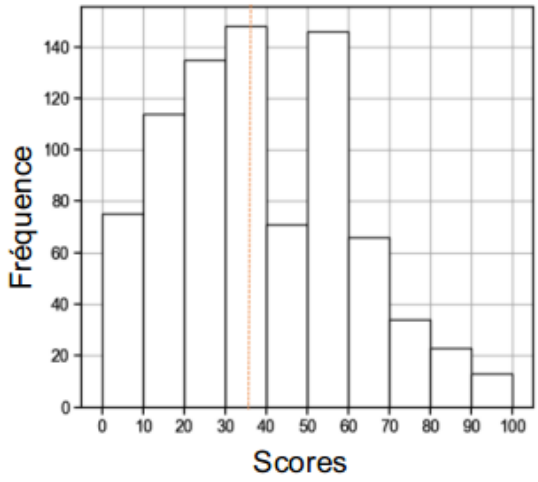

FIGURE 3.4 – Distribution des scores de grammaire

situe exactement à 50. Bien que le plus grand nombre de scores attribués soit entre 50 et 60 comme l'indique l'asymétrie de la distribution, il n'y a pas eu de tendance à noter plus en dessous ou au-dessus de 50 ;

Niveau de compréhensibilité : pour l'évaluation de la difficulté globale, nous notons des tendances proches de l'histogramme présenté dans la

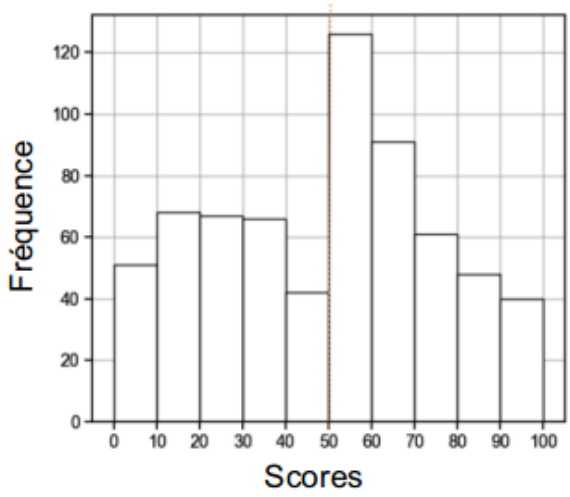

FIGURE 3.5 – Distribution des scores d'intelligibilité

figure 3.6 que dans l'histogramme présenté pour le vocabulaire. Les médianes sont proches, puisque celle des scores de difficulté globale est égale à 58 (contre 55 pour le vocabulaire) et nous voyons qu'il y a une tendance à attribuer des scores de difficulté globale au-dessus de 50. Ces constatations peuvent faire penser qu'il y a un lien entre la manière dont les annotateurs ont évalué la difficulté globale et le vocabulaire.

3.4.2.2 Corrélation entre la difficulté globale et les autres dimensions

Il s'agit de mettre en avant une possible relation entre le niveau de compréhensibilité représenté par les scores de difficultés globale et les scores liés respectivement à la complexité du vocabulaire, la complexité grammaticale et l'intelligibilité, nous nous sommes intéressée à la relation qui existait entre les différentes dimensions évaluées.

Méthode : la première étape a été d'analyser la corrélation existant entre les scores de compréhensibilité et les scores pour les trois autres dimensions. Nous avons vu plus tôt (cf. section 3.4.1), lorsque nous parlions des accords inter-annotateurs, qu'il existait deux façons d'étudier la relation entre deux variables continues : la première était de réaliser une corrélation de Pearson, la seconde était de réaliser une corrélation de Spearman. Pour rappel, la corrélation de Pearson permet d'identifier une relation linéaire entre les variables, la corrélation de Spearman permet

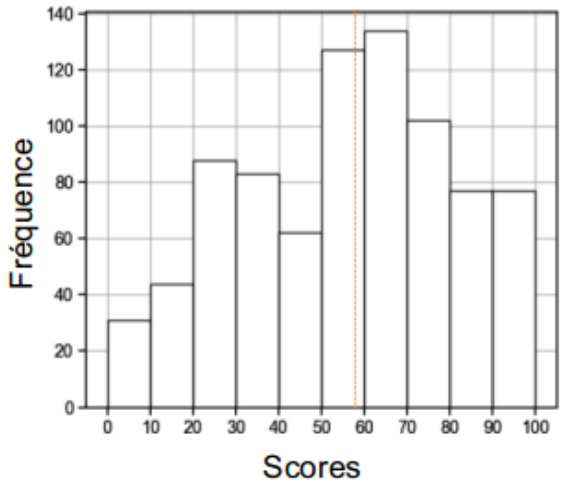

FIGURE 3.6 – Distribution des scores de difficulté globale

d'identifier une relation monotone, mais pas forcément proportionnelle entre les variables. Ici, l'intérêt est simplement de déterminer l'existence d'une relation entre les variables et nous n'avions pas d'hypothèses sur l'existence d'une relation linéaire ou non entre les scores des différentes dimensions, de ce fait, peu importait le choix de la corrélation calculée. Comme celle-ci a été utilisée plus tôt, nous reprenons la corrélation de Spearman.

Application et résultats : nous calculons la corrélation de Spearman bivariée (*i.e.* nous ne faisons pas d'hypothèse sur le signe que doit prendre le coefficient de corrélation) entre les scores de difficulté globale et :

- les scores de complexité du vocabulaire,
- les scores de complexité grammaticale,
- les scores d'intelligibilité de la parole. Dans le cas de l'intelligibilité de la parole, nous n'avons pris en compte que les documents annotés avec la modalité audio comprise dans le mode de présentation.

Le tableau 3.2 présente les résultats de corrélations de Spearman entre la difficulté globale et les trois autres dimensions.

Nous observons que les corrélations sont toutes positives et significatives (ce qui veut dire que les résultats obtenus ne sont probablement pas liés au hasard). La complexité grammaticale est moyennement corrélée à la difficulté globale, tandis que la complexité du vocabulaire et l'intelligibilité de la parole sont fortement corrélées. Cela illustre que :

TABLE 3.2 – Corrélations de Spearman bivariées entre la difficulté globale et la complexité du vocabulaire, la complexité grammaticale et l'intelligibilité de la parole (*** $p \leq 0,001$)

	Complexité vocabulaire	Complexité grammaticale	Intelligibilité de la parole
Difficulté globale	0,74***	0,56***	0,63***

— plus les extraits sont perçus comme étant compliqués du point de vue du vocabulaire et/ou de la grammaire, plus le score de difficulté globale attribué au document est élevé,
— plus la parole est intelligible, plus le document est considéré comme facile donc plus le score de difficulté globale attribué diminue.

Cette première étape a permis de démontrer les relations existant entre le niveau de compréhensibilité et les trois autres dimensions. Nous avons approfondi cette étude en réalisant des régressions linéaires multiples, pour quantifier l'influence du vocabulaire, de la grammaire et l'intelligibilité sur la difficulté quand elles sont considérées simultanément.

3.4.2.3 Corrélation des autres dimensions avec la difficulté globale

Si nous avons pu grâce aux corrélations bivariées statuer sur l'existence d'une corrélation positive significative entre les quatre dimensions étudiées, il faut maintenant regarder de quelle façon le vocabulaire, la grammaire et l'intelligibilité vont influer sur le niveau de compréhensibilité. Le but est de savoir dans quelle mesure une modification de la complexité du vocabulaire, de la grammaire ou de l'intelligibilité va modifier la difficulté globale.

Méthode : pour rendre explicite la relation qui existe entre plusieurs variables et une variable cible (appelée variable dépendante) la méthode la plus commune est la **régression linéaire multiple**. Il s'agit de la méthode la plus simple pour établir, si elle existe, une relation linéaire entre n variables et une variable dépendante. La relation sera du type :

$$y = x_1 * \alpha_1 + x_2 * \alpha_2 + ... + x_n * \alpha_n + \beta \tag{3.1}$$

où :
— y est la variable dépendante,
— x_i est la i^{ème} variable indépendante,
— α_i est le coefficient de régression de la i^{ème} variable indépendante,
— β est une constante qui correspond à l'intercept. Si toutes les variables indépendantes sont nulles, la variable dépendante prend la valeur de β.

Le modèle obtenu avec la régression linéaire multiple permet de prédire des scores de difficulté globale. La relation est établie en appliquant la **méthode des moindres carrés** [Chabert, 1989].

Pour évaluer la qualité du modèle, nous utilisons comme métrique d'évaluation le coefficient de détermination R^2 **ajusté**. Ce coefficient est une mesure située entre 0 et 1 qui permet de mesurer à quel point un modèle de régression linéaire multiple va être adéquat pour prédire la variable dépendante. Plus la valeur du coefficient de détermination se rapproche de 1, plus le modèle de régression linéaire est adéquat. Le R^2 donne également une information sur le pourcentage de la variance de la variable dépendante que le modèle permet d'expliquer. À noter qu'un grand nombre de variables explicatives fait augmenter le coefficient de détermination, mais cela tend à rendre le modèle peu robuste. Il vaut mieux en prendre en compte le nombre de variables explicatives pour s'assurer que le modèle est robuste, le R^2 ajusté prend en compte le nombre de variables explicatives qui ont été utilisées pour construire la relation linéaire, et nous nous assurons, en l'utilisant, de la fiabilité du modèle [Miles, 2014].

Application et résultats : ce qui sera exploité dans la suite pour chacune des dimensions sera la **moyenne de scores** par document.

Dans un premier temps, nous avons cherché à voir ce qui se passait si l'intelligibilité n'entrait pas en jeu dans la régression linéaire multiple, certains documents n'étant pas associé à la modalité audio, donc ne disposant pas de score d'intelligibilité. Une première régression linéaire multiple a été calculée en considérant comme variable dépendante la difficulté globale et comme variables indépendantes le vocabulaire et la grammaire, tous les modes de présentations ont été considérés pour calculer ce modèle. Cette régression permet d'obtenir un score de corrélation élevé entre les scores de difficulté globale réels et les scores de difficulté globale prédits à l'aide du modèle de régression obtenu. Le R^2 ajusté atteint une valeur de 0,76. Les coefficients non standardisés (NsCoef), qui permettent d'interpréter l'impact de chaque variable sur la sortie de la régression, montrent que la complexité du vocabulaire (NsCoef = 0,69) a plus de poids que la complexité grammaticale (NsCoef = 034) sur la variation de la difficulté globale. La figure 3.7 représente un nuage de points mettant en relation les scores prédits de difficulté globale avec les moyennes de scores humains de difficulté globale.

Ces premiers résultats sont intéressants, mais l'intelligibilité n'a pas été prise en compte. L'objectif est maintenant de voir si l'intelligibilité a une influence. Une deuxième régression linéaire multiple est réalisée en prenant en compte l'intelligibilité parmi les variables indépendantes. Pour cette régression, les scores de tous les modes de présentation ont été considérés à l'exception du mode de présentation T où l'intelligibilité n'entre pas en compte.

Cela permet d'obtenir un R^2 ajusté qui atteint une valeur de 0,82. La

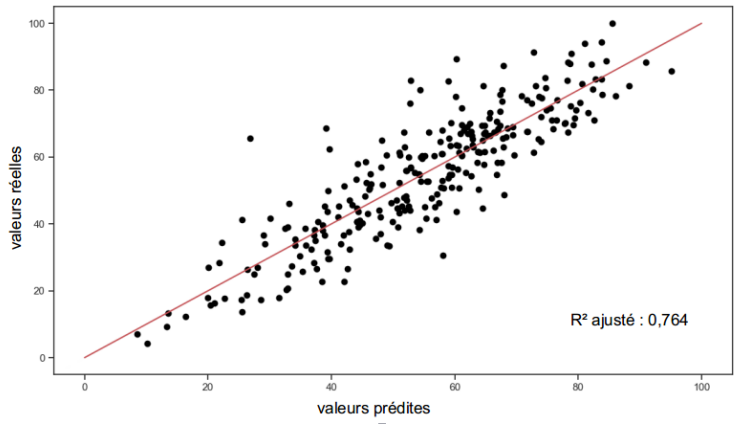

FIGURE 3.7 – Nuage de points mettant en relation les scores humains de difficulté globale et les scores prédits de difficulté globale, calculés à partir d'une régression linéaire multiple avec la complexité du vocabulaire et la complexité grammaticale comme variables indépendantes

complexité du vocabulaire a toujours le poids le plus élevé (NsCoef = 0,55), suivi par la complexité grammaticale (NsCoef = 0,31) et l'intelligibilité (NsCoef = 0,28). La figure 3.8 met en avant la relation entre les scores prédits et les scores réels.

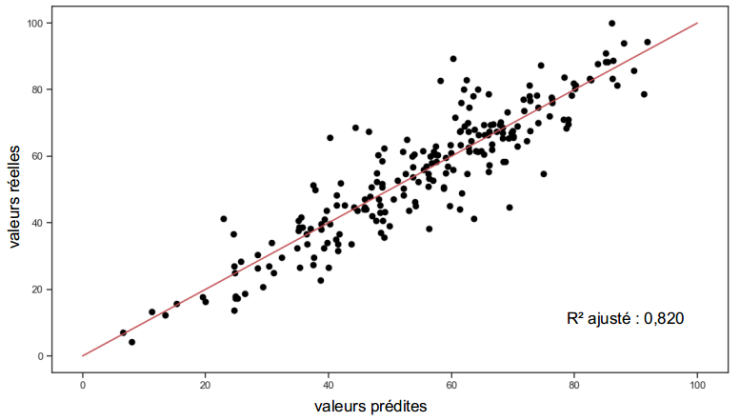

FIGURE 3.8 – Nuage de points mettant en relation les scores humains de difficulté globale et les scores prédits de difficulté globale, calculés à partir d'une régression linéaire multiple avec la complexité du vocabulaire, la complexité grammaticale et l'intelligibilité comme variables indépendantes.

Bilan : l'utilisation de régressions linéaires multiples a permis de quanti-

fier la relation qui existe entre vocabulaire, grammaire, intelligibilité et difficulté globale. La combinaison de ces trois dimensions pour construire le modèle de prédiction de difficulté globale permet d'obtenir un modèle qui permet d'expliquer 82% de la variance de la difficulté globale. Le modèle obtenu nous informe que les aspects linguistiques jouent un rôle prépondérant sur la compréhensibilité, mais que le vocabulaire est celui qui aura le plus d'influence.

3.4.3 Influence des modes de présentation sur les scores

Le second objectif de cette étude qualitative est d'étudier l'influence du mode de présentation (et donc des modalités disponibles) sur les scores donnés. Les hypothèses suivantes peuvent être faites :

— au vu de l'état de l'art, nous pouvons supposer que la combinaison de modalités simplifie la compréhension d'un document. Ainsi, la difficulté globale perçue est plus élevée pour le mode de présentation audio seul (A) que pour le mode de présentation audio+texte (AT) et le mode de présentation audio+vidéo (AV),
— comme il s'agit du mode de présentation où toutes les modalités sont utilisées, la difficulté globale est moins élevée pour le mode de présentation audio+vidéo+texte (AVT),
— nous avons vu dans l'état de l'art que lorsqu'elles étaient visibles la gestuelle et les expressions faciales aidaient à la compréhension orale, de même que la présence de sous-titres : la parole est plus intelligible si nous combinons la modalité audio avec les modalités vidéo et/ou texte.

Rien dans la littérature ne permet de déduire qu'il existe un quelconque lien direct entre modalité et complexité linguistique. Aucune relation n'est attendue entre le mode de présentation et la complexité du vocabulaire et la complexité grammaticale. Pour chaque mode de présentation et pour chaque dimension étudiée, nous avons calculé : la moyenne des scores et l'écart-type des scores pour pouvoir étudier l'évolution des scores en fonction des modes de présentation.

3.4.3.1 Modes de présentation et difficulté globale

Il faut mesurer si les différences entre la moyenne des scores de difficulté globale inter-mode de présentations sont significatives pour s'assurer qu'il existe une réelle différence dans les méthodes de notation de la difficulté en fonction des modalités. Dans le cas où nous observons une absence de différence significative dans les moyennes, alors cela signifie que les annotations sont réalisées de la même façon, quels que soient les modes de présentation.

La figure 3.9 permet d'observer que même si les écarts sont légers entre les scores, les scores de difficulté globale sont les plus élevés pour le mode de présentation audio seul (A).

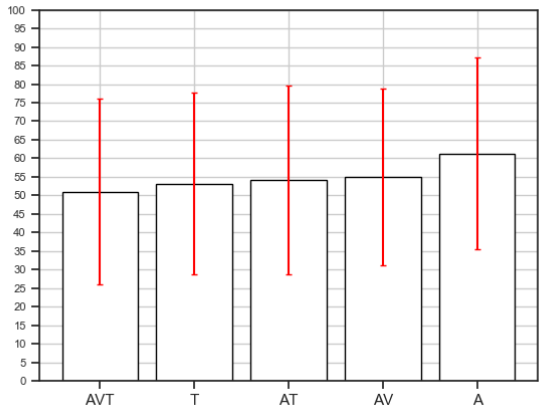

FIGURE 3.9 – Scores moyens de difficulté globale en fonction du mode de présentation (A : audio seul, T : texte seul, AV : audio+vidéo, AT : audio+texte, AVT : audio+vidéo+texte). Les barres d'erreur représentent ± 1 écart-type

Dans le corpus, les extraits présentés en mode audio seul (A) ont été considérés comme étant les plus difficiles. La moyenne des scores de difficulté globale est la plus basse pour le mode de présentation le plus complet (AVT), le mode de présentation texte seul (T) est le moins compliqué, suivi du mode présentation associant audio et texte(AT), puis du mode associant audio et vidéo (AV).

Ces observations semblent étayer l'idée selon laquelle le niveau de compréhensibilité est affecté par les modalités disponibles et que l'ajout de modalités constitue une source de facilitation. Pour le vérifier, un *t-test* [Xu et al., 2017] a été appliqué sur des échantillons indépendants (cf. tableau 3.3).

Les moyennes sont significativement différentes pour les scores de difficulté globale des modes de présentation A et T et pour les modes de présentation A et AVT.

TABLE 3.3 – Comparaison des moyennes de scores de difficulté inter-modes de présentation à l'aide du t-test pour variables indépendantes

Modes de présentation	T	AV	AT	AVT
A	**0,02**	0,08	0,05	**0,005**
T	x	0,62	0,8	0,53
AV	x	x	0,82	0,28
AT	x	x	x	0,4

Nous pouvons conclure que :

— la façon dont les extraits sont notés en terme de difficulté globale est semblable pour les modes de présentation A, AV et AT,

— les notations des modes T et AVT se sont faites d'une façon différente par rapport au mode A,
— le fait de combiner les autres modalités avec la modalité audio (modes AV, AT, AVT) permet de diminuer la difficulté globale perçue, minimisée en combinant toutes les modalités (mode AVT).

3.4.3.2 Modes de présentation et intelligibilité

L'état de l'art a mis en avant le fait que regarder la vidéo, tout en ayant accès à la transcription (texte), devrait apporter un avantage pour décoder la parole : si la vidéo seule ne permet pas de désambiguïser ce qui est dit, mettre à disposition le texte est susceptible d'éliminer ce problème. Pour le vérifier, nous comparons l'évolution des scores d'intelligibilité en fonction des modes de présentation, illustrée par la figure 3.10.

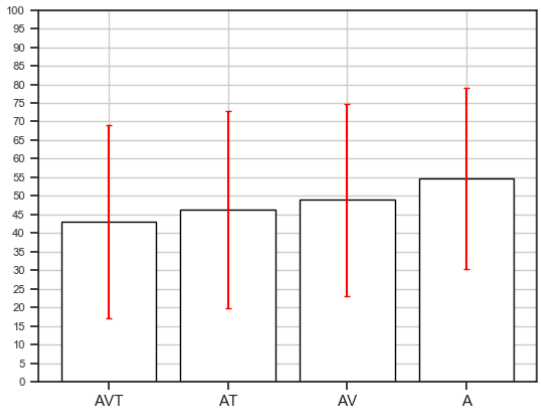

FIGURE 3.10 – Scores moyens d'intelligibilité en fonction du mode de présentation (AVT, AT, AV, A). Les barres d'erreur représentent ± 1 écart-type

Les extraits présentés sous le mode de présentation audio seul (A) sont ceux qui ont été considérés comme étant les plus inintelligibles par les annotateurs. Ajouter les modalités vidéo et/ou texte contribue à améliorer les scores d'intelligibilité. En se basant sur la figure 3.10, les extraits les plus intelligibles sont ceux présentés avec toutes les modalités.

Pour savoir s'il existe une réelle différence dans la façon d'annoter l'intelligibilité en fonction des modes de présentation, nous réalisons une comparaison des moyennes des scores à l'aide du *t-test* pour voir si elles sont significativement différentes (cf. tableau 3.4). Il y a une différence significative des moyennes entre les scores des modes de présentation A et AT et ceux des modes de présentation

A et AVT.

TABLE 3.4 – Comparaison des moyennes de scores d'intelligibilité inter-modes à l'aide du t-test pour variables indépendantes

Modalité de présentation	AV	AT	AVT
A	0,12	**0,02**	**0,002**
AV	x	0,47	0,13
AT	x	x	0,38

Cela nous permet de conclure que :
— la stratégie de notation de l'intelligibilité est différente si l'audio est la seule modalité disponible et si nous y ajoutons le texte et/ou la vidéo,
— le fait que les modes de présentations AV et A et les modes de présentation AT et AVT n'amènent pas de moyennes de scores significativement différentes peut signifier que c'est le texte qui constituera un véritable apport pour maximiser l'intelligibilité.

3.4.4 Modes de présentation et complexité lexicale et grammaticale

Les figures 3.11 et 3.12 ne permettent pas de dégager une tendance dans l'évolution des scores moyens par modes qui permettrait de mettre en avant une quelconque relation entre la complexité du vocabulaire, la complexité grammaticale et les modes de présentation.

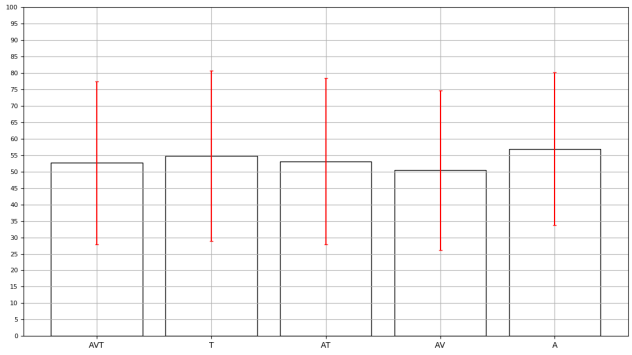

FIGURE 3.11 – Scores moyens de complexité du vocabulaire en fonction du mode de présentation (A, AV, AT, AVT). Les barres d'erreur représentent ± 1 écart-type

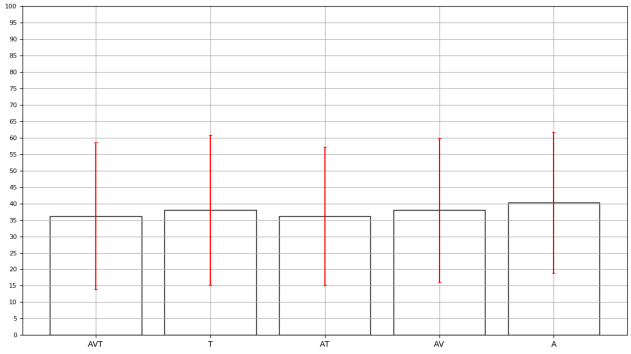

FIGURE 3.12 – Scores moyens de complexité grammaticale en fonction du mode de présentation (A, AV, AT, AVT). Les barres d'erreur représentent ± 1 écart-type

Réaliser une comparaison des moyennes des scores pour la grammaire et pour le vocabulaire à l'aide de t-tests, permet de confirmer qu'il n'existe pas de différences de moyennes significatives inter-mode de présentation. Ces observations permettent de conclure que pour la grammaire et le vocabulaire, la façon d'évaluer est indépendante des modalités disponibles.

Bilan de l'analyse quantitative menée sur les scores obtenus

Les analyses quantitatives des résultats ont permis de confirmer et de quantifier l'influence de la complexité du vocabulaire, de la complexité grammaticale et de l'intelligibilité sur le niveau de compréhensibilité. L'ensemble de ces dimensions combinées pour prédire la difficulté globale dans un modèle de régression linéaire permet d'expliquer 82% de la variance de la difficulté. L'analyse quantitative a également permis de confirmer le rôle prépondérant que jouent les modalités sur la difficulté globale et l'intelligibilité. Il a également été démontré dans cette partie que le fait de combiner les modalité audio, vidéo et texte permet de minimiser la difficulté globale perçue et de maximiser l'intelligibilité. En plus de l'analyse quantitative, il faut réaliser l'étude qualitative de cette expérience en analysant les commentaires qui ont été laissés par les participants.

3.4.5 Analyse qualitative des commentaires des experts

Les commentaires ont été prévus pour permettre de justifier les choix d'annotations des experts et notamment de détailler quels aspects en lien avec le vocabulaire, la grammaire et/ou l'intelligibilité ont joué un rôle dans les scores attribués. De plus, les commentaires peuvent mettre en avant des facteurs supplémentaires susceptibles d'affecter les différentes dimensions considérées. Nous

avons donc procédé à une étude détaillée de l'ensemble des commentaires des annotateurs.

3.4.5.1 Commentaires sur la difficulté globale

Il était obligatoire pour les annotateurs de justifier la note attribuée pour la difficulté globale, ce qui représente 825 commentaires (275 documents x 3 annotateurs) Les arguments avancés le plus souvent pour donner la note de difficulté globale étaient en lien avec une des trois autres dimensions (vocabulaire, grammaire, intelligibilité) mais également avec le manque de contexte ce qui peut être plus ou moins problématique suivant l'extrait de film considéré.

La figure 3.13 représente la proportion de commentaires appartenant à chacune des catégories citées ci-dessus. .

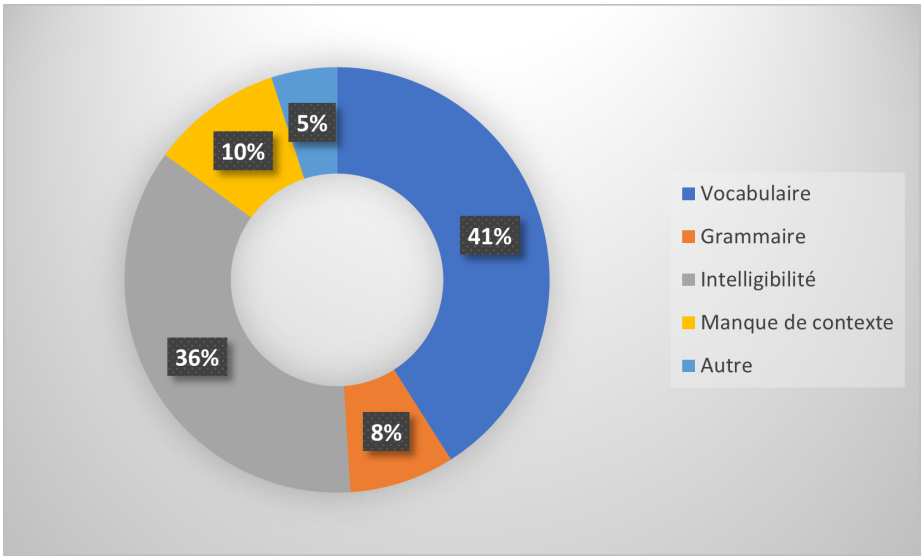

FIGURE 3.13 – Diagramme représentant la catégorie des commentaires des annotateurs pour l'évaluation de la difficulté globale

Dans la catégorie « Autre », nous comptons des commentaires qui indiquent que la difficulté est due à des effets de longueur (longueur de l'extrait, longueur des répliques) mais aussi à des particularités dans l'écriture comme l'usage de ton humoristique (quiproquos, jeux de mots), de ton ironique ou de métaphores qui peuvent être difficiles à saisir lorsqu'on a une mauvaise connaissance de la langue apprise.

Si nous nous intéressons de plus près à la proportion de chaque catégorie, nous notons que le vocabulaire et l'intelligibilité ont eu un rôle très important dans l'estimation de la difficulté globale, tandis que la grammaire a eu un rôle marginal même en comparaison du manque de contexte. Nous remarquons également un parallèle entre cette figure et les constats faits après avoir étudié

les corrélations entre les différentes dimensions : le vocabulaire est la dimension ayant le plus d'influence sur la difficulté globale, suivie de l'intelligibilité et enfin de la grammaire.

Si nous entrons plus dans le détail, nous pouvons noter que, quand la difficulté avait pour origine une mauvaise intelligibilité, ce sont des éléments en lien avec la prosodie qui ont été cités pour expliquer la difficulté perçue du document :

— le débit de parole trop rapide,
— l'accent des locuteurs,
— des problèmes d'articulation.

Nous retrouvons ici ce qui avait été constaté dans l'état de l'art : la prosodie joue un rôle important dans la compréhension orale. Le fait que les extraits soient issus de films professionnels implique un certain contrôle dans les conditions de captation, mais aussi dans la gestion de l'environnement sonore, cela peut expliquer le fait que des éléments liés à l'environnement sonore n'aient pas été évoqués comme des facteurs principaux d'une mauvaise compréhensibilité des extraits.

Quand les sources de difficulté sont liées au vocabulaire, les annotateurs soulignent notamment :

— la présence de vocabulaire compliqué soit par la rareté des mots ou de leur utilisation, soit parce que le vocabulaire appartenait à un champ lexical spécifique (ou les deux),
— l'utilisation de registre de langue qui sort du registre courant.

Dans la littérature, nous avions déjà pu voir que le vocabulaire jouait un rôle dans la compréhensibilité et ce quelque soit le type de support. Néanmoins, si nous avions déjà pu identifier que les mots rares pouvaient être source de complication, le champ lexical spécifique n'était pas ressorti comme une source de difficulté. Mais, utiliser du jargon médical revient à se servir d'un vocabulaire qui n'est pas connu de tous et donc qui est rare, alors les deux aspects rareté/spécificité sont peut-être difficilement dissociables. Le registre de langue apparaît pour la première fois dans l'ensemble cette étude comme une source de complexité lexicale. Une possibilité est que le registre de langue courant est plus « facile » pour des apprenants, car il s'agit du registre qui est le plus étudié pour l'apprentissage des langues. L'utilisation des registres familier, vulgaire, mais aussi soutenu oblige à avoir une connaissance plus fine de la langue : en effet, apprendre et maîtriser l'usage de jargon et/ou de mots vulgaires ou apprendre à utiliser correctement du langage soutenu requiert un niveau plus élevé dans la langue apprise. Cela veut dire que tout mot utilisé avec un sens qui n'est pas du registre courant contribue à augmenter la complexité du vocabulaire et affecte le niveau de compréhensibilité d'un document.

Les annotateurs ont cité la présence de références implicites et la présence de références culturelles comme source de difficulté globale liée au contexte : le fait de ne pas avoir vu le reste du film, mais aussi le manque de connaissances culturelles est susceptible d'amener des problèmes dans la compréhension globale

dans le cas où des références sont faites. Ces références se manifestent régulièrement sous la forme de noms propres comme des noms de personnages (comme « Bastoche » cité souvent dans les extraits d'*Un long dimanche de fiançailles*), des noms de lieux (par exemple « Montmartre » cité dans *Amélie Poulain*) ou de personnalités ou figures historiques (dans un extrait de *Séraphine*, les personnages parlent de Picasso, Braque ou encore du Douanier Rousseau). Ne pas connaître ces noms propres et à quoi et à qui ils réfèrent entraîne un manque de contexte qui nuit à la compréhension. La lacune culturelle se reflète alors par une lacune en terme de vocabulaire.

Pour ce qui est de la grammaire, bien que les commentaires l'évoquent peu comme source de difficulté globale, les éléments cités demeurent intéressants. Nous retrouvons des facteurs qui avaient déjà été identifiés comme source de complexité dans la littérature, notamment :

— les structures syntaxiques complexes,
— les temps verbaux difficiles.

Après avoir vu les éléments cités comme source de difficulté globale par les annotateurs, la même analyse peut être faite sur les autres dimensions évaluées.

3.4.5.2 Commentaires sur la complexité du vocabulaire

Les commentaires sur la notation du vocabulaire sont au nombre de 478. Ils ont permis de mettre en avant six types de phénomènes pertinents qui, selon les annotateurs, expliquaient la complexité du vocabulaire (voir figure 3.14) :

— la présence de mots ou d'expressions considérées comme compliqués, car ils sont peu utilisés,
— l'utilisation de champs lexicaux spécifiques (comme le champ lexical de l'astrologie dans Delicatessen ou le champ lexical de la cuisine dans *Cyrano de Bergerac*),
— le registre de langue (familier, vulgaire, argotique ou soutenu),
— les mots nécessitant une référence culturelle pour être compris (il peut s'agir de noms de lieux comme Montmartre, dans *Le fabuleux destin d'Amélie Poulain* ou de noms de peintres célèbres dans *Séraphine*),
— la diversité lexicale,
— les mots inconnus qui sont enseignés aux apprenants de niveau avancé (ils correspondent aux mots qui sont au-dessus du niveau B1-B2 du CECRL, qui correspond à un niveau intermédiaire dans le domaine de l'apprentissage des langues étrangères).

Dans une moindre mesure, les annotateurs ont également parlé de la polysémie, de la présence de nombres cardinaux complexes, des marques d'oralité ou de l'utilisation de régionalismes comme sources de difficulté du vocabulaire dans un document (dans *La gloire de mon père*, Joseph Pagnol dit par exemple « Bonjour » pour dire « Au revoir »).

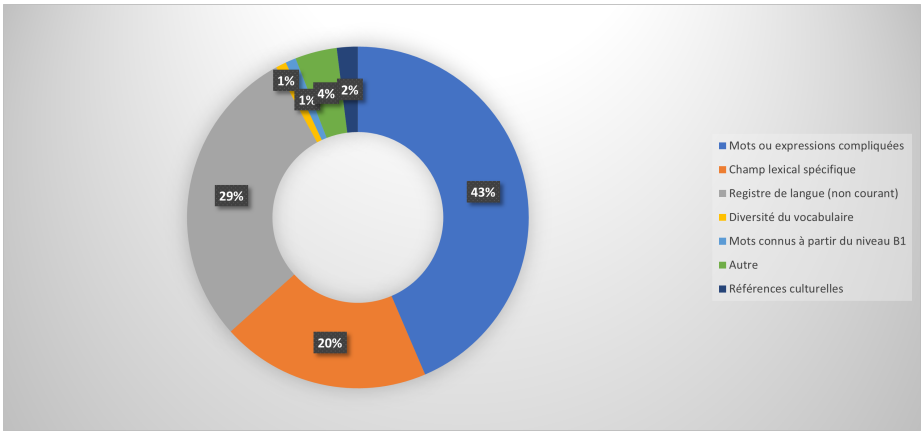

FIGURE 3.14 – Diagramme représentant la catégorie des commentaires des annotateurs pour l'évaluation de la complexité du vocabulaire.

3.4.5.3 Commentaires sur la complexité grammaticale

Nous comptons au total 360 commentaires pour justifier les scores de grammaire. C'est l'aspect qui a été le moins justifié par les annotateurs. Nous restons dans une certaine tendance qui est de penser que la grammaire est la dimension qui a été évaluée de la façon la moins poussée par les annotateurs.

Cependant, bien que les commentaires soient moins nombreux, ils permettent de relever des phénomènes influençant la complexité grammaticale qui seront intéressants à utiliser pour tenter de prédire la complexité grammaticale par la suite. Nous relevons notamment des éléments liés directement à la forme des mots, notamment la flexion des verbes, ces éléments sont en lien avec ce qu'on appelle la complexité morphologique :

— la diversité des temps verbaux utilisés,
— les temps verbaux et les modes « compliqués » (passé simple, plus-que-parfait, conditionnel, impératif, subjonctif).

Nous notons aussi des éléments en lien avec la complexité de la structure et des relations dans les phrases (c'est ce que nous appelons la complexité syntaxique) :

— les tournures syntaxiques complexes,
— la longueur des phrases,
— la voix passive,
— les propositions coordonnées et les propositions subordonnées.

La figure 3.15 représente la proportion de commentaires correspondant aux éléments cités ci-dessus.

Ponctuellement, des commentaires soulignent certaines structures grammaticales très spécifiques pour expliquer que cela puisse amener de la complexité,

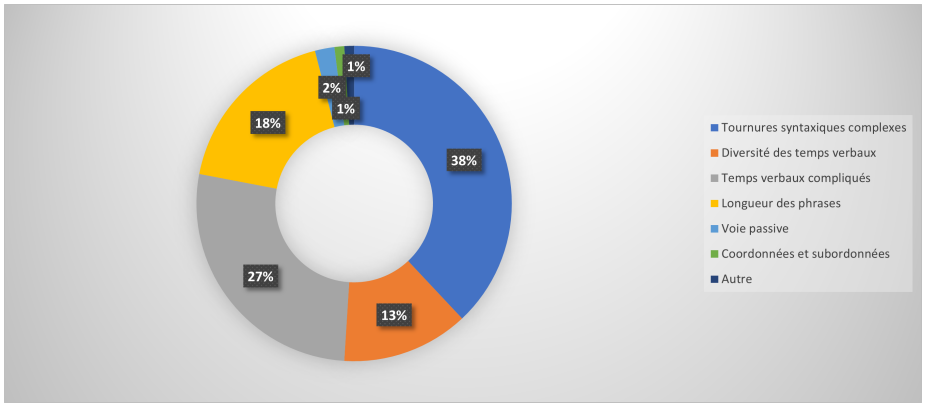

FIGURE 3.15 – Diagramme représentant la catégorie des commentaires des annotateurs pour l'évaluation de la complexité grammaticale

comme l'utilisation du « y » et du « en », l'expression de l'hypothèse, la structure « ne... que »... Nous constatons que la présence de voix passive, de coordonnées et de subordonnées ainsi que la longueur des phrases sont très peu citées par les annotateurs comme des facteurs de complexité grammaticale, alors qu'ils sont au contraire très évoqués dans la littérature.

3.4.5.4 Commentaires sur l'intelligibilité

En termes de proportion, avec 418 commentaires, l'intelligibilité est l'aspect qui a été le plus commenté par les annotateurs. Même si le nombre de commentaires est inférieur à celui pour le vocabulaire, il faut se souvenir qu'il y avait moins de justification à apporter pour l'intelligibilité puisqu'il faut prendre en compte le mode de présentation T qui ne contenait pas d'audio.

Si nous regardons les commentaires des annotateurs pour expliquer leur score d'intelligibilité, nous pouvons dégager les catégories d'explication suivantes :

- le débit de parole (38 %) : soit le débit est trop rapide soit il est saccadé,
- l'environnement sonore (22 %) : les annotateurs évoqueront un environnement sonore trop bruité (brouhaha) ou pollué par la présence d'un fond musical,
- l'accent des locuteurs (12%) : présence d'accents non standards dans certains extraits (provençal (*la Gloire de mon père*) ou africain (*Intouchables*),
- la mauvaise élocution des locuteurs (8%) : des mots mâchés quand ils parlent trop vite (c'est le cas de Gérard Depardieu dans le film *La Chèvre* par exemple),
- le volume sonore des locuteurs (voix forte, voix chuchotée) (10%) : certains parlent trop fort comparé à d'autres par exemple tandis que certains chuchotent (dans *Séraphine* par exemple),
- la parole superposée (5%) : il arrive quelques fois que les locuteurs parlent

en même temps.

Nous notons que les éléments de cette liste sont ceux qui avaient déjà été identifiés comme des facteurs influant sur le niveau de compréhensibilité dans l'état de l'art réalisé dans le chapitre 2.

La figure 3.16 donne des détails sur les proportions des commentaires des annotateurs.

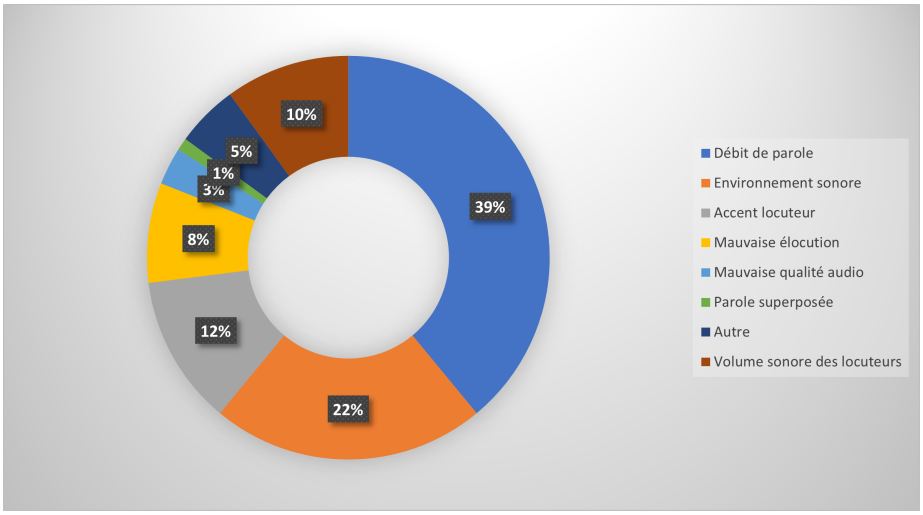

FIGURE 3.16 – Diagramme représentant la catégorie des commentaires des annotateurs pour l'évaluation de l'intelligibilité

3.5 Bilan

Pour résumer, le **corpus ESCAL** créé dans le cadre de ce travail de thèse rassemble des documents auxquels ont été associés un ensemble de scores attribués par un panel d'experts. Ces derniers ont eu accès à une interface spécialement développée pour leur permettre de noter des documents qui leur ont été affectés et ce, suivant quatre dimensions différentes : en évaluant la complexité du vocabulaire et de la grammaire employés, ainsi que l'intelligibilité et le niveau de compréhensibilité ou difficulté globale. Le tableau 3.5 résume les principales carcatéristiques de ce corpus, valorisé par les annotations d'experts qui apportent des éléments subjectifs pour mesurer le niveau de compréhensibilité.

Une analyse fine des scores obtenus ainsi que l'étude qualitative des commentaires additionnels laissés par les experts ont permis dans un premier temps de démontrer que le vocabulaire, la grammaire et l'intelligibilité de la parole sont des facteurs importants pour expliquer la variation du niveau de compréhensibilité d'un contenu.

TABLE 3.5 – Description synthétique du corpus ESCAL

Caractéristique	Détails
Documents	Extraits = 55 (issus de 15 films cf Table 3.1) Mode de présentation = 5 (A, T, AV, AT, AVT) Nombre total de documents = 275
Annnotateurs	Experts FLE = 15 (13 femmes/2 hommes ; de 27 à 63 ans) Nombre d'annotateurs par document = 3 Nombre de documents par annotateur = 55 (11 par mode de présentation)
Scores	Fournis par un annotateur = 19 (3 pour T et 4 pour A, AV, AT, AVT) Nombre total de scores par extrait = 57 Nombre total de scores = 3135
Commentaires fournis par les annotateurs	difficulté globale = 825 vocabulaire = 478 grammaire = 360 intelligibilité = 418
Durée totale du corpus	1 heure 20

L'**étude quantitative** a notamment permis d'étudier de manière croisée l'influence de la difficulté globale sur chacune des trois autres dimensions et inversement. Le fait que le vocabulaire a le plus d'influence sur la difficulté globale, suivi par l'intelligibilité et enfin la grammaire a pu être mis en avant. La prise en compte de ces trois dimensions (vocabulaire, grammaire et intelligibilité) pour construire un modèle, basé sur la régression linéraire multiple, a permis d'aboutir à un modèle expliquant 82% de la variance de la difficulté globale.

L'**étude qualitative** des commentaires des annotateurs a permis d'identifier plus en profondeur les phénomènes qui entraient en jeu dans la perception de chacune des dimensions qui ont été étudiées. Or, ces phénomènes sont importants pour atteindre l'objectif central de cette thèse qui vise à construire une mesure objective du niveau de compréhensibilité. Cette connaissance permettra par la suite de cibler des paramètres à calculer et à inclure dans la construction de systèmes automatiques de prédiction de niveau de compréhensibilité. Le travail réalisé autour du corpus ESCAL à fait l'objet d'une publication nationale [Randria et al., 2020a] et d'une publication internationale [Randria et al., 2020b].

Le chapitre suivant, qui clôt cette première partie permet de lister le plus exhaustivement possible les phénomènes et facteurs pertinents pour déterminer les paramètres qui pourront intervenir dans le calcul de la mesure objective du niveau de compréhensibilité.

Chapitre 4

Phénomènes et facteurs pertinents pour le calcul d'une mesure objective du niveau de compréhensibilité

Sommaire

4.1 Introduction

Le domaine de la didactique des langues étrangères et ses pratiques ont été explorés dans les chapitres précédents pour connaître les facteurs qui contribuent de manière générale à mesurer le niveau de compréhensibilité, que ce soit par rapport à l'écrit, à l'oral ou aux contenus audiovisuels et aux modalités qui les composent. Nous avons également créé un corpus centré sur l'évaluation subjective du niveau de compréhensibilité d'un ensemble de documents. Ce corpus a permis d'étudier plus en détail les éléments qui entrent en compte dans les scores attribués par les experts annotateurs. Pour compléter et clore cette première partie, il nous reste une dernière étude à mener.

L'objectif de cette thèse, nous le rappelons, est de proposer une mesure objective du niveau de compréhensibilité. Cela passe par le développement de systèmes permettant de prédire cette mesure pour chaque nouveau document considéré. Comme nous le verrons dans la partie suivante, plusieurs approches sont possibles, basées sur des paramètres ou des représentations extraits des contenus audiovisuels et modalités traités. Pour choisir ces paramètres, il est nécessaire de comprendre plus finement les phénomènes et les facteurs qui sont liés à la mesure du niveau de compréhensibilité en tenant compte de l'éclairage donné par notre analyse du corpus ESCAL.

Dans ce chapitre, nous examinons donc chacune des quatre dimensions présentées dans les chapitres précédents comme contribuant à la mesure de la compréhensibilité. Notre objectif est de faire le lien avec les ressources disponibles et les méthodes de calcul associées aux différents facteurs retenus. Cela nous permettra de faire le lien, dans la seconde partie du manuscrit, avec les outils de traitement automatique des textes, des images et du son qui permettront d'extraire les paramètres correspondants.

4.2 Facteurs liés à la complexité du vocabulaire

4.2.1 Fréquence lexicale

Que ce soit dans notre étude ou dans la littérature, la fréquence des mots (ou fréquence lexicale) a un impact sur la complexité du vocabulaire telle qu'elle est perçue par un non-natif d'une langue. Dans diverses études touchant à l'apprentissage des langues, la fréquence lexicale est régulièrement citée comme affectant la compréhensibilité des documents (peu importe le type de support utilisé). Les mots peu fréquents d'un texte auront plus de chance de ne pas être connus de l'apprenant, rendant le texte plus difficile à comprendre. Le rôle de la fréquence lexicale dans la compréhensibilité d'un document s'illustre notamment dans la littérature par le biais des diverses formules de lisibilité qui incluent régulièrement le pourcentage de mots peu fréquents [Lively and Pressey, 1923, Dale and Chall, 1948, Henry, 1975]. La fréquence lexicale comme source de complexité du vocabulaire revient également dans les commentaires des annotateurs du corpus lorsque ceux-ci évoquent la présence de mots peu communs soit, car

ils appartiennent à un champ lexical spécifique, soit parce qu'ils sont utilisés avec un sens n'appartenant pas au registre de langue courant. La connaissance de la fréquence des mots est donc un pré-requis important pour la mesure objective de la complexité du vocabulaire.

Il existe plusieurs manières d'évaluer la fréquence lexicale. Dans les formules de lisibilité, les mots peu fréquents étaient identifiés comme étant ceux absents de la liste de mots fréquents établis par des linguistes comme Thordnike [Thorndike, 1921] pour l'anglais, ou Gougheneim [Gougenheim et al., 1964] pour le français. Actuellement, il existe des bases de données donnant directement la fréquence des mots dans des bases de données. Comme bases de données lexicales connues, nous pouvons citer FlElex [François et al., 2014], qui est un lexique pour l'enseignement du FLE qui répertorie la fréquence normalisée des mots en fonction du niveau européen commun de référence et qui se base sur un corpus de manuels de FLE ; ou Lexique [New et al., 2004], qui est une base de données lexicale construite autour de mots appartenant à des films ou livres français. Il est possible de faire varier les approches pour étudier la fréquence lexicale en utilisant ce type de base de données. La première manière d'exploiter ces bases de données est simplement d'utiliser les fréquences lexicales pour déterminer la proportion de mots rares dans un document. Mais le caractère plus exhaustif qu'apporte une base de données peut aussi permettre de repérer des mots qui sont absents de la base de données. Il peut s'agir des noms propres ou des mots qui ne sont pas dans la langue cible, mais il est également possible que ces mots soient absents de la base de données, car il s'agit de néologismes ou encore d'argot ou de mots vulgaires. Par exemple, dans la base de données de Lexique, nous constatons que certains mots du registre vulgaire sont absents de la base (c'est le cas de « putain »), ces mots sont peut-être usuels dans le quotidien (il peut s'agir de régionalisme) mais il ne s'agit pas de mots qui sont systématiquement enseignés aux apprenants. En calculant la proportion des mots « inconnus » de la base de données, nous pouvons obtenir une information complémentaire sur la complexité lexicale.

En résumé, la fréquence lexicale ou encore l'absence de fréquence dans une base de données d'informations lexicales sont deux phénomènes mesurables qui peuvent jouer un rôle prépondérant dans la complexité du vocabulaire.

4.2.2 Richesse lexicale

Lorsque nous parlons de richesse lexicale, il faut comprendre le terme « richesse » d'un point de vue quantitatif. Comme l'expliquent Thoiron et Arnaud, la richesse lexicale ne fait pas intervenir de notion de « rareté » ou encore de « sophistication du vocabulaire » [Thoiron and Arndaud, 1992]. Un texte est riche s'il contient un grand nombre de mots différents. Dans la littérature, il est dit que la richesse lexicale augmente au fur et à mesure qu'une personne parlant une langue a des connaissances approfondies de celle-ci. Dans le sens inverse, plus la richesse lexicale d'un texte augmente plus il faut que les connaissances

lexicales des personnes exposées au texte soient élevées. Ceci met en avant la relation qui existe entre la richesse lexicale et la complexité lexicale.

L'étude de la richesse lexicale est liée à la diversité lexicale et à la densité lexicale [Johansson, 2008]. La diversité lexicale s'intéresse à la variabilité du vocabulaire : plus le vocabulaire est varié plus la diversité lexicale augmente. Cette variabilité du vocabulaire dans un document se mesure habituellement en faisant le rapport entre le nombre de mots distincts et le nombre de mots total. Cela correspond à la mesure la plus connue de diversité lexicale : le Type-Token Ratio, mais il existe des déclinaisons comme l'index de Guiraud (ou Root Type Token Ratio) [Guiraud, 1954], où le nombre de mots total est ramené à sa racine pour essayer de s'affranchir de la longueur du document en termes de nombre de mots. La densité lexicale se différencie de la diversité lexicale parce qu'elle permet d'obtenir une information sur la quantité d'informations apportée par la diversité des mots lexicaux. Elle se calcule généralement en faisant le rapport entre le nombre de mots porteurs de sens (exemple : nom, adverbe, verbe, adjectif) d'un document et le nombre total de mots du-dit document.

4.2.3 Niveau CECRL

Deux participants de l'expérience ont noté la présence de mots qui ne sont connus qu'à partir du niveau CECRL B1 ou B2 et ont considéré que cela était une source de complexité du vocabulaire. Le CECRL a défini à l'échelle européenne des niveaux seuils pour catégoriser le niveau des apprenants européens de langue étrangère, nous reprenons les termes utilisés dans le CECRL pour présenter chacun des niveaux [Conseil de l'Europe, 2003] :

— A1 et A2 correspondent au niveau « utilisateur élémentaire »,
— B1 et B2 correspondent au niveau « utilisateur indépendant »,
— C1 et C2 correspondent au niveau « utilisateur expérimenté ».

Les grands débutants (A1) et les débutants (A2) doivent acquérir du vocabulaire élémentaire, ce qui correspond à du vocabulaire fréquemment utilisé. La présence des mots de niveau B1 et/ou B2 augmente naturellement la complexité du vocabulaire. En se basant sur des ouvrages listant les mots à connaître pour un apprenant de niveau A1-A2, il est possible de calculer un pourcentage de mots de ce niveau dans un document. Nous faisons l'hypothèse que plus celui-ci est élevé, plus nous pouvons considérer que le document est simple d'un point de vue lexical.

4.2.4 Registre de langue

Parmi les commentaires des annotateurs de l'expérience, le registre de langue est évoqué à plusieurs reprises (il est évoqué dans 17% des commentaires pour expliquer les scores de vocabulaire) comme un facteur de complexité lexicale. Le registre de langue correspond à un mode d'expression qui varie en fonction du contexte, du locuteur... Nous pouvons citer par exemple le registre courant, le registre familier, le registre vulgaire et le registre soutenu. Les mots qui ne sont

pas utilisés avec une définition du registre courant ne sont pas forcément étudiés dans le cadre des cours de langue. Soit parce qu'il s'agit de mots littéraires qui ne seront vus que par des apprenants d'un niveau avancé, soit parce que ces mots sont trop familiers ou trop vulgaires pour être appropriés ou pertinents à enseigner dans le cadre de l'apprentissage des langues étrangères. Dans les deux cas, nous supposons que la présence de ces mots tend à augmenter la difficulté du vocabulaire.

Pour étudier le registre de langue, il est possible de se baser sur un dictionnaire de mots de la langue française où, pour chaque définition d'un mot, le registre de langue associé à cette définition est précisé. De par la tendance polysémique de certains mots, en fonction du contexte un même mot peut être du registre courant, mais aussi du registre vulgaire. Par exemple le mot « saleté » peut être utilisé dans le sens courant ou dans un sens plus vulgaire en guise d'insulte. En prenant en compte l'aspect polysémique d'un mot et donc en incluant une certaine incertitude quant au sens à utiliser il est possible de mesurer la proportion de mots susceptibles d'être utilisés dans un registre de langue vulgaire, familier...

4.2.5 Longueur des mots

La longueur des mots ne semble pas avoir joué un rôle dans la perception de la complexité du vocabulaire chez les participants de l'expérience. Cependant, nous trouvons dans la littérature des réflexions intéressantes sur l'impact de la longueur des mots sur la complexité lexicale. D'après la « loi de Zipf », les mots les plus courts tendent à être les mots les plus fréquents [Zipf, 1935]. Si nous nous basons sur cette observation, il est raisonnable de supposer que la longueur des mots va varier avec le niveau de l'apprenant : plus le niveau d'une personne dans une langue cible va évoluer, plus la longueur des mots qu'il va utiliser va augmenter [Granfeldt, 2006].

Pour calculer la longueur des mots, il est possible d'utiliser des mesures traditionnelles de longueur de mots comme :
— le nombre de caractères alphabétiques par mot,
— le nombre de syllabes par mot.

Calculer la longueur des mots en se basant sur le nombre de caractères alphabétiques est très fréquent : nous trouvons le calcul de la longueur via le nombre de syllabes dans la formule de lisibilité de Flesch [Flesch, 1948] par exemple. Les deux manières de calculer la longueur des mots peuvent aussi être incluses simultanément dans certaines études sur la complexité lexicale [Gala et al., 2014].

4.2.6 Polysémie

Dans les commentaires, certains annotateurs ont cité l'utilisation de mots polysémiques comme une source de complexité lexicale. La polysémie peut consti-

tuer un frein dans la compréhension dans le cadre de l'apprentissage des langues : un apprenant qui a un niveau faible dans la langue apprise, en l'absence d'indices contextuels explicites, ne peut pas se reposer sur son expérience personnelle pour discriminer quel est le sens à allouer à un mot polysémique. Il est même possible qu'un seul des sens de ce mot soit connu de l'apprenant. *A contrario*, un apprenant de niveau avancé, qui a plus de connaissances et d'expériences personnelles peut éventuellement se reposer sur ses acquis pour induire le sens d'un mot polysémique dans un contexte spécifique [Hollard, 2010].

L'accès à un dictionnaire permet d'avoir accès aux multiples définitions d'un même mot. Cela peut être utilisé par exemple pour calculer le nombre moyen de définitions par mot. Les mots avec le plus de sens sont ceux qui sont les plus fréquents (par exemple le verbe « faire ») et aussi les plus simples vus par les utilisateurs élémentaires (niveaux A1 et A2 [Beacco and Porquier, 2007, Beacco et al., 2004]).

4.3 Facteurs liés à la complexité grammaticale

S'intéresser à la complexité grammaticale c'est s'intéresser à la complexité morphosyntaxique et donc à l'aspect syntaxique (structure des phrases) et à l'aspect morphologique (structure des mots) du texte.

4.3.1 Complexité morphologique

La morphologie recouvre tout ce qui touche à la forme des mots, leurs flexions et leurs variantes, ainsi la complexité morphologique désigne la complexité des mots. Pour évaluer la complexité morphologique, nous pouvons nous intéresser à deux aspects :

- les temps verbaux, évoqués de manière récurrente par les annotateurs dans leurs commentaires,
- la composition des mots, avec l'étude des morphèmes. Bien que les morphèmes n'aient pas été cités dans les commentaires, des études en linguistique induisent que la composition morphémique des mots (et des phrases) est liée à la complexité morphologique [Colé et al., 2004].

4.3.1.1 Complexité verbale

La fréquence des temps verbaux (présent de l'indicatif, passé simple...), le nombre de différentes formes verbales et la variété de temps du passé sont identifiés dans la littérature comme des phénomènes permettant d'expliquer la complexité morphologique [Bulté and Housen, 2012]. Les annotateurs de l'expérience ont eux-mêmes régulièrement évoqué les temps verbaux comme source de complexité grammaticale. Les enseignants sont sûrement plus sensibles à cet aspect de la grammaire parce que le système verbal français est considéré comme complexe pour des apprenants en langue étrangère, c'est le cas pour les

japonais étudiant le français en L2 par exemple [Kashioka, 1990]. La maîtrise des temps verbaux des apprenants en langue native ou seconde va évoluer avec leur niveau dans la langue. Plus l'apprenant devient expérimenté, plus il est capable d'utiliser des temps verbaux complexes et plus il est en mesure d'utiliser plusieurs temps verbaux différents. Cette évolution simultanée entre les temps verbaux maîtrisés et le niveau des apprenants se reflète dans les référentiels de programmes de l'Alliance Française [Chauvet, 2008] : dans ces référentiels, les auteurs précisent quels temps verbaux doivent être acquis en fonction du niveau (en terme de CECRL) des apprenants (par exemple l'apprentissage du présent correspond au niveau A1).

Un texte contenant des temps verbaux complexes et diversifiés est plus complexe du point de vue grammatical. Deux premiers paramètres à calculer pour évaluer la complexité grammaticale d'un texte sont :

— le nombre de temps verbaux différents,
— la fréquence des temps verbaux complexes.

Le calcul de la **proportion de verbes conjugués aux temps du passé** paraît aussi être intéressante. Bien que les annotateurs aient fait peu de commentaires pour la grammaire, en terme de proportion, l'évocation des temps du passé comme source de complexité grammaticale fait partie des plus fréquemment citées. Il semble que, dans notre étude, les enseignants en langue étrangère considèrent que les temps du passé sont les temps verbaux qui amènent le plus de difficulté. Comme mesure complémentaire, et pour faire le lien avec les référentiels de programme, il est possible de calculer aussi **la proportion de chacun des temps verbaux** présent dans le document.

4.3.1.2 Complexité morphémique

Le morphème est le plus petit élément du langage porteur du sens. En d'autres termes, il s'agit de la plus petite unité susceptible d'amener un changement dans le sens du mot. Les morphèmes sont de plusieurs types, il y a :

— les désinences : les marqueurs du pluriel (-s), du féminin (-e) ou les terminaisons verbales (-ons, -ez...), ce sont des morphèmes grammaticaux,
— les morphèmes lexicaux qui sont les mots simples qui constituent le lexique (noms, verbes, adverbes, adjectifs),
— les suffixes et les préfixes, qui sont porteurs du sens, mais seulement s'ils sont liés à des radicaux (ou mots racines).

Par exemple le mot « mangeons » est constitué de deux morphèmes : le verbe « manger » (il s'agit du radical ou de la racine) et la terminaison « -ons » qui marque la deuxième personne du pluriel.

En se servant d'informations issues d'une base de données lexicales qui renseigne notamment sur le nombre de morphèmes de chaque entrée lexicale et en prenant en comptant la forme du mot, nous pouvons calculer le nombre moyen

de morphèmes par mot dans document. Selon le nombre de morphèmes qu'il contient, un mot est considéré comme simple ou complexe d'un point de vue morphologique [Diependaele et al., 2012]. Plus la grammaire et le vocabulaire d'une personne sont développés dans une langue donnée, plus cette personne est capable d'utiliser des mots composés de plusieurs morphèmes et donc plus complexes [Carlisle and Goodwin, 2014]. L'utilisation de mots composés de plusieurs morphèmes reflète de la complexité grammaticale du texte. Sachant cela, nous faisons l'hypothèse que plus le nombre moyen de morphèmes par mot est élevé, plus le document est complexe d'un point de vue grammatical.

4.3.2 Complexité syntaxique

La complexité syntaxique désigne la complexité des phrases en termes de structure. Étudier la syntaxe revient à étudier l'organisation des mots ou des groupes de mots dans une phrase ainsi que la relation qui existe entre eux. La structure des phrases étant notamment influencée par leur longueur, cela implique d'étudier ces deux dimensions.

4.3.2.1 Longueur des phrases

La longueur des phrases est considérée comme un bon indicateur de la complexité structurelle. Hunt a montré que la longueur des phrases était un indice probant pour mesurer la maturité syntaxique [Hunt, 1965]. Une personne progressant dans une langue (native ou non native) fait des phrases plus longues en fonction de son avancée dans l'apprentissage. De la même façon, plus la phrase est longue, plus elle est difficile à appréhender par l'apprenant s'il n'a pas le niveau adéquat. L'hypothèse faite sur la longueur est que plus une phrase est longue, plus elle est susceptible d'être complexe syntaxiquement [Rupp et al., 2001]. Cette hypothèse est soutenue par les commentaires des annotateurs, qui ont noté que la longueur des phrases amène de la complexité grammaticale. Si en moyenne, un texte contient des phrases longues, alors le texte est syntaxiquement plus compliqué.

Pour étudier la longueur des phrases, plusieurs approches sont possibles, nous pouvons calculer :

— le nombre de caractères de la phrase,
— le nombre de syllabes de la phrase,
— le nombre de morphèmes de la phrase.

4.3.2.2 Structure des phrases

La seconde dimension qui peut être étudiée est la structure des phrases, plus spécifiquement, nous pouvons étudier les relations qui existent entre les mots et entre les portions de phrases (ou propositions). L'analyse de la structure des phrases peut se faire en s'intéressant au nombre de propositions qui constituent la phrase ou encore aux arbres syntaxiques.

Profondeur de l'arbre syntaxique

Une phrase avec un nombre de propositions plus élevé a aussi un arbre syntaxique profond. Une phrase ayant un arbre profond est plus complexe d'un point de vue syntaxique [Ferreira, 1991]. Une possibilité pour estimer la profondeur d'un arbre est de calculer son nombre de gouverneurs. Le gouverneur d'un mot est celui qui justifie sa présence dans la phrase. Si nous regardons l'exemple d'arbre syntaxique repris de Kahane [Kahane, 2001] (figure 4.1), dans la phrase « le petit garçon parle à Marie », le verbe « parle » est la racine de la phrase, le mot « garçon » est le gouverneur des mots « le » et « petit », l'adverbe « à » est le gouverneur du nom propre « Marie ».

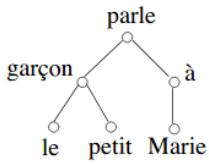

FIGURE 4.1 – Exemple d'arbre syntaxique (issu de [Kahane, 2001])

L'utilisation d'un analyseur morphosyntaxique permet d'extraire des informations morphologiques et syntaxiques telles que le temps verbal et la classe grammaticale, mais aussi les gouverneurs de chaque mot d'une phrase. À partir des informations de l'analyseur il est possible de calculer le nombre moyen de gouverneurs par phrase. Plus le nombre moyen de gouverneurs par phrase augmente, plus les arbres syntaxiques des phrases sont profonds et plus la complexité syntaxique augmente.

Propositions subordonnées ou coordonnées et voix passive

Le type de relations présentes dans la phrase va avoir une influence sur la complexité grammaticale perçue. L'utilisation de propositions coordonnées ou subordonnées et de voix passive donne lieu à des tournures syntaxiques complexes. Les annotateurs évoquent une tournure syntaxique complexe ou littéraire comme facteur de complexité grammaticale. La présence de propositions et de subordonnées est identifiée dans la littérature [Norris and Ortega, 2009, Bulté and Housen, 2012] et par les participants de l'expérience comme une source de complexité syntaxique. Une phrase qui compte un nombre de propositions élevé tend à être plus complexe [de Clercq, 2016]. La voix passive est identifiée dans l'état de l'art comme étant plus difficile à traiter que d'autres structures [Byrnes and Sinicrope, 2008] et les annotateurs ont aussi évoqué ce phénomène comme une source de complexité syntaxique.

La méthode la plus simple pour repérer les propositions coordonnées et subordonnées est de trouver les conjonctions de coordination ou de subordination (mais, ou, et, donc, qui, que...) qui permettent de les introduire. Une fois que nous avons retrouvé ces mots outils, il est possible de calculer le nombre moyen de propositions subordonnées et de propositions coordonnées par phrase en nous

basant sur le nombre de conjonctions [Zhang et al., 2013].

Notons que la fréquence d'utilisation des formes passives peut également être utilisée pour mesurer la complexité syntaxique [Bulté and Housen, 2012, Brindley and Wigglesworth, 1997]. En se servant d'informations issues d'un analyseur morphosyntaxique, il est possible d'identifier le nombre de mots qui sont utilisés à la voix passive et de calculer le nombre de fois où la voix passive est utilisée dans un texte. Nous supposons que plus la voix passive est utilisée, plus un document est grammaticalement complexe.

4.4 Facteurs liés à l'intelligibilité de la parole

Nous retrouvons des similitudes entre les phénomènes identifiés par les annotateurs comme nuisibles pour l'intelligibilité et les phénomènes qui ont été cités dans la littérature comme facteurs de difficulté pour la compréhension orale. Comme certains phénomènes ont déjà été expliqués en détail dans le chapitre 2, cette section est parfois moins détaillée pour éviter les redites.

4.4.1 Prosodie : débit de parole et qualité de l'élocution

Les annotateurs ont, pour la grande partie, ciblé des problèmes de prosodie que nous retrouvons dans la littérature comme des facteurs affectant l'intelligibilité, tels que :

- le débit de parole,
- l'accent des locuteurs,
- la mauvaise élocution.

Dans notre corpus, la proportion de documents contenant de la parole avec des accents semble trop faible pour prendre en compte les phénomènes liés à l'accent pour la mesure de l'intelligibilité. Le choix a été fait de se concentrer sur uniquement deux phénomènes parmi les trois cités ci-dessus :

- le débit de parole : mesurable en calculant le nombre moyen de phonèmes par seconde ou le nombre de syllabes par seconde,
- les problèmes d'élocution : en étant en possession de la transcription exacte, il est par exemple possible d'estimer un écart entre ce qui est vraiment dit et ce qui est compris par un système de reconnaissance de la parole par exemple. Nous pouvons mesurer la distance entre la prédiction du système et la transcription. Dans cette approche, il faut cependant considérer le fait que le système de reconnaissance n'est pas un système parfait.

4.4.2 Environnement sonore

Dans leurs commentaires, les annotateurs ont beaucoup cité le bruit ou la musique de fond comme pouvant être dérangeants pour la compréhension orale. Dans le domaine du traitement automatique, mais aussi de la didactique, la

réverbération est reconnue comme un facteur pouvant affecter la compréhension orale, c'est pour cela qu'il peut être intéressant de l'ajouter aux phénomènes à considérer.

4.4.2.1 Bruit

Nous avons vu dans l'état de l'art, mais aussi dans les commentaires des annotateurs du corpus, que le bruit pouvait être une nuisance pour la compréhension orale. Il peut être estimé en calculant ce que nous appelons le rapport signal sur bruit (ou SNR pour « Signal to Noise Ratio » en anglais) [Johnson, 2006]. Celui-ci peut être calculé de diverses façons : soit en calculant le rapport entre la puissance du signal et la puissance du bruit, mais ça induit de connaître au préalable le bruit présent dans le signal. Comme cela n'est pas toujours possible, d'autres méthodes ont été mises en place dans lesquelles une estimation du bruit est réalisée, c'est le cas par exemple de l'outil STNR (Speech Signal to Noise Ratio) développé par le NIST et dédié à la mesure du SNR dans du signal contenant de la parole [Kim and Stern, 2008]. L'outil WADA-SNR propose une méthode alternative fondée sur les hypothèses que la parole « propre » suit une distribution Gamma et que le bruit suit une distribution gaussienne [Seong et al., 2019].

4.4.2.2 Réverbération

En acoustique, la réverbération correspond à la persistance du son après interruption de la source sonore. Bien qu'elle n'ait pas été évoquée par les annotateurs et qu'elle n'apparaisse pas comme facteur d'intelligibilité dans des études dédiées à l'apprentissage des langues, nous suspectons que la réverbération est un élément qui peut affecter la compréhension de l'oral. Une forte réverbération a pour effet de recouvrir le signal émis initialement avec ce que nous pourrions qualifier de « bruit » réverbérant, ce qui amène des complications pour comprendre les mots qui sont les plus « noyés » dans ce bruit. Plus il y a de réverbération, plus la quantité de mots difficiles à distinguer augmente, et moins la parole est intelligible. Cependant, une absence totale de réverbération a également un impact négatif sur l'intelligibilité. S'il n'y a pas de persistance du son, alors, le son décroît avant d'avoir pu être propagé dans l'environnement. En conséquence bien qu'initialement le signal de parole émis soit intelligible à presque 100%, comme le message n'arrive pas au destinataire, la communication est quand même détériorée, et donc l'intelligibilité aussi [Long, 2014].

La réverbération peut se mesurer en se servant de prédicteurs déjà existants tels que l'outil SRMR (speech-to-reverberation modulation energy ratio) [Falk et al., 2010].

4.4.3 Prédictibilité phonétique des mots

De façon identique à une tâche de lecture, lors des tâches d'écoute, il existe également une étape de sélection lexicale [Dufour and Frauenfelder, 2007]. En

partant du principe que le traitement du signal audio en entrée se fait de façon séquentielle chez l'humain, la reconnaissance d'un mot, en supposant qu'il est basé sur le modèle de COHORT [Marslen-Wilson and Welsh, 1978], se déroule en deux phases :

— l'activation du lexique,
— la sélection lexicale.

Lorsque le mot commence à être entendu, l'auditeur active son lexique et plus particulièrement il active un ensemble de mots qui partagent les mêmes premiers phonèmes. Au fur et à mesure que le mot se déroule et que des phonèmes supplémentaires sont entendus, les informations partielles que l'auditeur a sur le mot augmentent ce qui diminue le nombre de mots candidats possibles.

Point d'unicité phonologique

Comme pour la lecture, il est possible de reconnaître un mot dès qu'un point d'unicité phonologique est atteint, il s'agira alors du phonème du mot qui permet de discriminer ce mot par rapport à d'autres mots candidats. Si nous reprenons l'exemple dans [Ghio et al., 2016], pour le mot « vérité » le point d'unicité se situe à la troisième syllabe. Une fois ce point atteint, la sélection lexicale a lieu et nous sommes sûrs que le mot prononcé n'était pas « véritable » par exemple. Un mot dont le point d'unicité est éloigné peut mettre plus de temps à être reconnu, si le décodage du mot se fait suivant le modèle de COHORT.

Les bases de données lexicales fournissent parfois des informations de l'ordre phonétique et peuvent par exemple renseigner le point d'unicité phonologique des entrées lexicales en donnant la position du phonème dans le mot qui va permettre de le discriminer d'autres mots potentiellement activés dans le lexique d'une personne. En connaissant la position du point d'unicité de chaque mot, il est possible de calculer sa position moyenne dans un document. Une fois le point d'unicité atteint, l'auditeur peut prédire la fin du mot en cours de traitement et anticiper le mot suivant [Marslen-Wilson and Welsh, 1978]. Si nous supposons que l'auditeur fonctionne selon le modèle de COHORT, nous pouvons faire l'hypothèse qu'un point d'unicité phonologique situé tôt dans le mot favorise le décodage et donc a une influence positive sur l'intelligibilité de la parole.

Voisinage phonologique

Il a été observé qu'un nombre de voisins orthographiques importants peut nuire à l'identification d'un mot lors de la tâche de lecture. Un constat similaire peut peut-être être fait pour la tâche de compréhension orale, quant au voisinage phonologique des mots. Nous pouvons supposer qu'un mot ayant de nombreux voisins phonologiques active plus de vocabulaire qu'un mot en possédant très peu. Le nombre de candidats potentiels croît si le mot possède de nombreux voisins, ralentissant ainsi la phase d'identification d'un mot, cela peut donc également nuire à la phase de décodage du message entendu et à l'intelligibilité.

Les bases de données lexicales peuvent mettre à disposition le nombre de voisins orthographiques, mais aussi le nombre de voisins phonologiques des mots

qu'elles contiennent. Le nombre moyen de voisins phonologiques par mot est ainsi calculable. Nous posons l'hypothèse que plus le nombre de voisins phonologiques augmente, plus cela nuit à l'intelligibilité de la parole.

4.4.4 Pureté de la parole

La parole peut être qualifiée de « pure » si elle est enregistrée dans des conditions sonores idéales, et de bruitée ou de très bruitée, quand la parole est enregistrée dans des conditions sonores très dégradées [Pinquier, 2004]. La pureté de la parole est un aspect pertinent à étudier du point de vue de l'intelligibilité, car de la parole pure n'est pas intelligible de la même manière que de la parole qui est superposée à un fond musical ou à des bruits environnementaux. Nous supposons que plus la parole est pure plus elle est intelligible. Pour estimer la pureté de la parole dans un signal audio, il est intéressant de se pencher sur des caractéristiques de la parole qui permettent de la discriminer d'autres types d'événements sonores tels que la musique, la parole bruitée, etc. Par exemple, la parole se distingue de la musique parce que, à court terme, l'énergie du signal temporel varie beaucoup plus [Saunders, 1996].

Pour évaluer la pureté de la parole, il existe des outils tels que la mesure de l'entropie du signal audio [Pinquier, 2004]. L'entropie mesure le désordre dans un signal, ce qui permet de donner une information quant à la pureté de la parole, car la parole seule n'a pas la même entropie que de la parole bruitée.

4.4.5 Redondance audio et image

L'analyse quantitative du corpus a démontré que les modalités disponibles ont joué un rôle important pour les participants de l'expérience. Nous avons vu que le mode de présentation AVT était celui qui permettait d'obtenir la meilleure intelligibilité et qui permettait de minimiser la difficulté globale. Lorsque nous nous intéressons, non plus aux sources de complexification, mais à la facilitation de la compréhensibilité, en nous penchant sur certains commentaires des annotateurs pour expliquer ce qui a permis de les aider à palier à certains problèmes de compréhension, nous retrouvons comme phénomènes intéressants :

- la redondance audio/texte : lorsque la parole est parfois difficilement intelligible, le fait d'avoir accès à la modalité texte « annule » la difficulté des annotateurs à comprendre ce qui est dit,
- la redondance audio/image : soit l'image permet de mieux cerner ce qui est dit parce que les participants voient les locuteurs, soit l'image apporte un contexte grâce à la visualisation du cadre spatial et temporel.

En nous référant aux études [Dahl and Ludvigsen, 2014] qui ont prouvé que l'accès au visage permettait d'améliorer la compréhension orale des natifs et des non-natifs, nous pouvons supposer que la difficulté d'un document est minimisée s'il y a des visages en gros plans des personnes en train de parler. Le fait de voir le visage d'une personne en train de parler peut aider à cerner ce qu'il dit

d'une meilleure façon grâce aux mouvements des lèvres, mais aussi à cerner ses intentions par le biais de ses expressions faciales.

Il n'est pas possible d'affirmer que ce phénomène a un lien avec l'intelligibilité. Cependant, il est intéressant d'analyser l'influence éventuelle de cette redondance sur l'intelligibilité. Cela peut se faire par exemple en exploitant simultanément la modalité audio et la modalité vidéo, nous pouvons analyser la quantité de gros plans superposés à une zone de parole et voir si cette quantité est corrélée à l'intelligibilité de la parole. Il faut néanmoins prendre en compte une limite : une personne apparaissant à l'écran lorsqu'il y a de la parole n'est pas forcément le locuteur courant, la personne qui parle peut se trouver hors champ (c'est le cas de la voix « off » par exemple).

4.5 Autres facteurs liés à la compréhensibilité globale

En plus des phénomènes décrits précédemment et relatifs aux trois dimensions reliées à la complexité du vocabulaire, la complexité grammaticale et l'intelligibilité, certains phénomènes complémentaires peuvent avoir un impact direct sur la difficulté globale, et ce sans être liés aux autres dimensions étudiées, mais en ayant un impact au niveau cognitif.

4.5.1 Longueur des documents

De façon intuitive, nous pouvons supposer que plus un document (audio, vidéo ou texte) est long (en termes de durée ou en terme de nombre de mots ou de phrases), plus cela est susceptible d'affecter la tâche de compréhension. Thompson et Rubin avaient noté que les étudiants en langue seconde considéraient les segments de plus de 2,5 minutes comme trop longs pour qu'ils puissent rester concentrés [Thompson and Rubin, 1996]. Dans [Bloomfield et al., 2011], nous trouvons plusieurs études sur l'influence de la longueur d'un passage sur la compréhension orale suggérant, qu'en réalité, celle-ci n'a en fait qu'un rôle mineur sur la compréhension. Aux vues de ces études contradictoires, inclure des paramètres en lien avec la longueur des extraits utilisés dans l'étude reste intéressant. Nous pouvons supposer que ces paramètres jouent un rôle plus ou moins important dans le calcul du niveau de compréhensibilité.

La longueur des documents peut se calculer de diverses façons, en fonction des modalités à disposition :

— en nombre absolu de caractères,
— en nombre absolu de phonèmes,
— en nombre absolu de phrases,
— ou simplement en terme en durée (secondes).

4.5.2 Cohérence du discours

Dans les sections qui traitent des facteurs affectant la complexité lexicale et la complexité grammaticale, l'unité de référence la plus grande est la phrase. Mais, il s'avère qu'il y a des mécanismes au niveau supérieur du discours qui peuvent avoir une influence sur la compréhension orale. Notamment, des études ont montré qu'un discours structuré avec des marqueurs explicites (comme « en premier lieu », « ensuite »...) améliore la compréhension orale des interlocuteurs [Camiciottoli, 2004, Chaudron and Richard, 1986]. Les marqueurs du discours permettent d'établir des relations entre des propositions, des phrases, et ils permettent également d'établir des relations entre plusieurs passages d'un énoncé. Les marqueurs du discours qui relient des propositions sont appelés des micro-marqueurs, ceux qui permettent de relier plusieurs passages sont appelés les macro-marqueurs. Ces marqueurs amènent de la cohérence au discours, et sont reconnus comme des potentiels indicateurs intéressants pour estimer le niveau de compréhensibilité [Bloomfield et al., 2010].

Il est possible de trouver des listes des marqueurs du discours pour le français. Nous pouvons cité par exemple LEXCONN [Roze et al., 2012]. Une telle liste peut servir à repérer les marqueurs du discours d'un document dont la transcription est accessible et estimer un nombre moyen de marqueurs de discours par phrase par exemple, ou le nombre absolu. Aux vues des constats tirés de la littérature, l'hypothèse à poser est que plus le nombre de marqueurs du discours augmente, plus le document est simple en terme de compréhensibilité.

4.5.3 Quantité d'informations visuelles et charge cognitive

Des études ont introduit l'idée que se servir de textes oraux et d'indices visuels, qui aident à relier un élément visuel à ce qui est entendu, permettrait d'améliorer l'apprentissage [Mayer, 2001, Sweller et al., 1988]. Une étude a tenté de généraliser ce constat, mais n'a pas abouti à des résultats probants [Tabbers et al., 2004]. Si nous ne remettons pas en question le fait que l'accès à certaines informations visuelles puisse être un appui pour la compréhension orale, vient quand même la question de ce qui se passe lorsqu'il y a trop d'informations disponibles à l'écran. Ceci peut nuire au traitement de l'information : le spectateur devant faire le tri de beaucoup plus d'informations, il est possible que cela amène à une surcharge cognitive et nuise à la qualité du traitement des données [Miller, 1956].

En nous servant d'outils de détection d'objets et de personnes (comme par exemple YOLO [Redmon et al., 2016]) nous pouvons inclure des paramètres permettant d'estimer la quantité d'éléments différents présents à l'écran (en termes d'objets et de personnes) ainsi que l'espace qu'ils occupent à l'écran (de petits éléments seront logiquement plus difficiles à exploiter). Comme les informations visuelles ne s'arrêtent pas à la quantité d'éléments présents à l'écran, nous avons aussi étudié le flux optique des vidéos en utilisant des outils disponibles spécia-

lisés dans le traitement de l'image (OpenCV [Bradski and Kaehler, 2000]). Le flux optique donne une information sur les mouvements des objets dans une vidéo : beaucoup de mouvements dans une vidéo peut être une source de complication pour le traitement de l'information et peut avoir un impact sur la compréhensibilité.

4.6 Bilan

Dans ce quatrième chapitre nous avons fait le tour de l'ensemble des facteurs qui, selon la littérature et les pratiques en didactique des langues, peuvent contribuer directement ou indirectement, à l'évaluation du niveau de compréhensibilité des contenus audiovisuels. Nous avons passé en revue chacune des différentes dimensions liées au vocabulaire, à la grammaire, à l'intelligibilité, sans oublier de considérer le niveau global. Nous avons fait le point sur les méthodes de calcul des différents facteurs identifiés, que celles-ci soient relativement facile à mettre en oeuvre, qu'elles s'appuient sur l'existence de ressources lexicales ou fassent appel aux possibilités offertes par des méthodes de traitement automatique ou outils du numériques permettant de traiter notamment le son et les images.

Conclusion

Cette thèse est motivée par le constat de l'inexistence d'indicateur de niveau de compréhensibilité pour les documents audiovisuels authentiques utilisés en apprentissage des langues étrangères. Notre proposition est de mettre en place une mesure objective calculée automatiquement et permettant de prédire un tel indicateur.

Pour atteindre cet objectif et proposer des méthodes de calcul de cette mesure, il a été indispensable de comprendre comment la compréhensibilité était abordée dans le domaine de la didactique des langues, à quoi correspondait un document authentique, quels étaient les objectifs des enseignants en utilisant ce type de document, quels étaient les aspects qu'ils souhaitaient faire travailler à leurs apprenants. Ce travail fait l'objet du premier chapitre de cette partie du manuscrit.

Un travail important a ensuite consisté à étudier les facteurs qui entraient en jeu dans la mesure de la compréhensibilité en fonction des types de documents authentiques utilisés dans l'enseignement des langues étrangères : documents écrits, enregistrements audio, ou support audiovisuels. Le chapitre 2 de ce manuscrit est consacré à un état de l'art centré sur le domaine de la didactique des langues.

Ce travail a notamment permis d'identifier plusieurs familles de facteurs. Nous avons donc considéré dans la suite de notre étude, quatre dimensions principales suivant lesquelles l'analyse de la compréhensibilité pouvaient être étudiée :

— la complexité du vocabulaire,
— la complexité grammaticale,
— l'intelligibilité de la parole,
— la difficulté globale.

Au delà de ces quatre dimensions, ce travail a également mis en avant le fait que la perception de la difficulté était influencée par les modalités disponibles : texte, audio et/ou vidéo. Les documents audiovisuels aidaient davantage dans la tâche de compréhension orale que les documents audio seuls. L'accès aux sous-titres avait un effet bénéfique pour la compréhension des documents audiovisuels. Deux facteurs, plus corrélés à l'aspect humain, notamment les dimensions affective et cognitive, très subjectives, ont été identifiés mais non pris en compte dans la suite de notre étude.

Même si un ensemble de facteurs ont été identifiés, car vastement étudiés dans la littérature, nous avons constaté qu'il n'existait cependant pas de corpus annoté permettant d'étudier et d'estimer le rôle de ces facteurs dans l'évaluation de la compréhensibilité telle que la pratique les enseignants de langues. Pour aller plus loin dans l'étude de la mesure du niveau de compréhensibilité, et répondre au manque identifié, nous avons créé un corpus dédié à l'Évaluation Subjective de la Compréhensibilité pour l'Apprentissage des Langues, nommé ESCAL.

Nous avons construit le corpus ESCAL à partir de 55 extraits de films, proposés sous cinq modes de présentation différents : audio seul (A), texte seul (T), audio et vidéo (AV), audio et texte (AT), audio, vidéo et texte (AVT). Un panel de 15 enseignants de FLE a contribué à l'évaluation des documents mis à disposition en tenant compte des quatre dimensions rappelées plus haut.

Le chapitre 3 de ce mémoire est donc dédié à la description du corpus ESCAL, au protocole d'évaluation proposé aux experts annotateurs et accessible au travers d'une interface développée spécifiquement. Un analyse fine du corpus et des éléments collectés auprès des enseignants a permis d'étudier l'influence des différentes dimensions et de prendre en compte les compléments apportés dans les commentaires laissés par les annotateurs. Ainsi, l'analyse quantitative du corpus a permis de confirmer que :

- il existe une corrélation entre le niveau de compréhensibilité et chacune des trois autres dimensions : la complexité du vocabulaire, la complexité grammaticale et l'intelligibilité,
- ces mêmes trois dimensions jouent également un rôle prépondérant pour expliquer la difficulté globale : l'inclusion de la complexité du vocabulaire, de la complexité grammaticale et de l'intelligibilité comme variables indépendantes pour la construction d'un modèle de régression linéaire multiple permet d'expliquer 82% de la variance de la difficulté globale,
- le niveau de compréhensibilité dépend des modalités disponibles : la combinaison de celles-ci permet de diminuer la difficulté globale et d'augmenter l'intelligibilité de la parole ; le fait de combiner les modalités audio, vidéo et texte maximise l'intelligibilité et minimise la difficulté globale.

L'analyse qualitative du corpus par le biais de l'étude des commentaires des annotateurs a permis de conforter ces constats, mais aussi d'identifier les éléments qui ont influencé l'évaluation des quatre dimensions. Si certains ont déjà été cités dans la littérature, les commentaires ont permis de dégager d'autres phénomènes complémentaires entrant en jeu dans la perception de la compréhensibilité : par exemple le registre de langue, mais aussi le niveau CECRL.

A la lumière de ces analyses et des éléments identifiés dans la partie état de l'art, il a été possible de faire un bilan de l'ensemble des phénomènes pertinents à prendre en compte pour construire un système de prédiction du niveau de compréhensibilité. Le chapitre 4 reprend ces différents phénomènes et revient sur le calcul des différents facteurs qui permettraient d'obtenir un indicateur sur le niveau de compréhensibilité de documents audiovisuels authentiques.

Les facteurs retenus serviront à la construction d'une première mesure objec-

tive du niveau de compréhensibilité. Deux approches seront utilisée et comparées, notamment à travers le prisme de l'explicabilité. Une approche d'apprentissage supervisée sera entièrement guidée par les constats et les conclusions tirées de l'état de l'art et de l'analyse du corpus ESCAL, mais aussi par l'exploitation des annotations des experts pour vérifier la pertinence des différentes mesures construites. Une seconde approche non supervisée sera étudiée, dans laquelle la mesure objective sera construite en se basant uniquement sur des représentations issues de réseaux de neurones. Les mesures obtenues seront ensuite confrontées à la perception des enseignants et/ou des apprenants de FLE.

Cette thèse étant une thèse CIFRE avec la société Archean Technologies, il s'agit également de montrer comment un indicateur prédisant le niveau de compréhensibilité pourrait être intégré ensuite dans une plateforme dédiée à l'apprentissage des langues.

Deuxième partie

Élaboration d'une mesure objective du niveau de compréhensibilité de documents audiovisuels

Introduction

La première partie de ce manuscrit était consacrée à l'étude de la compréhensibilité telle qu'elle est considérée dans le domaine de la didactique des langues. Une meilleure connaissance des pratiques actuelles d'enseignement de langues étrangères mais aussi du rôle des documents authentiques, notamment les supports audiovisuels, et de l'interaction, ont permis d'orienter nos travaux vers la définition d'une mesure du niveau de compréhensibilité des documents audiovisuels. Pour les besoins de cette étude nous avons créé un corpus et collecté des annotations auprès d'enseignants de langues étrangères familiers de la tâche d'évaluation de compréhensibilité (corpus ESCAL). L'analyse fine de ces données nous a permis dans un premier temps d'identifier les phénomènes et les facteurs qui influent sur la compréhensibilité. Ainsi, différents points de vue ont été considérés dans cette étude : celui des niveaux linguistiques à travers la complexité du vocabulaire et la complexité grammaticale, celui du niveau acoustique avec un *focus* sur l'intelligibilité de la parole et enfin, le niveau de compréhensibilité global appelé encore difficulté globale. (cf. figure 4.2).

Ainsi, à partir de cette étude fondée sur l'analyse :

— des commentaires collectés auprès des annotateurs du corpus ESCAL,

— de la littérature dans le domaine de l'apprentissage des langues,

— et de la littérature dans le domaine du traitement automatique,

nous avons répertorié un ensemble de facteurs ainsi que leur mode de calcul (cf. figure 4.2).

La seconde partie du manuscrit s'appuie sur un ensemble de ressources lexicales et d'outils de traitement automatique pour calculer un ensemble de paramètres associés à chacun des facteurs et phénomènes identifiés précédemment. Chaque paramètre est une valeur numérique extraite des contenus audiovisuels et sont issues d'une ou plusieurs des modalités considérées dans cette étude : le texte, l'audio ou la vidéo en tant que séquence d'images.

Ces paramètres contribueront à créer des modèles permettant de prédire la complexité du vocabulaire, la complexité grammaticale, l'intelligibilité et le niveau de compréhensibilité.

Le premier chapitre de cette seconde partie est dédié à la présentation plus formelle de ces paramètres ainsi qu'aux méthodes et outils qui permettent de les extraire du contenu traité. Nous nous intéressons ensuite à la construction de modèles pour la prédiction du niveau de compréhensibilité. Pour cela, nous

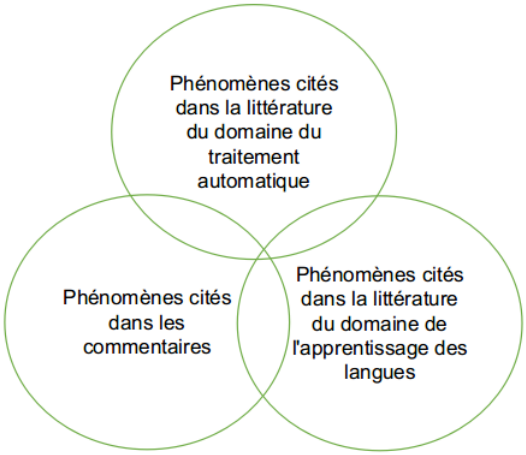

FIGURE 4.2 – Paramètres considérés

avons fait le choix de comparer deux approches :

— une approche dite « classique », fondée sur une phase d'extraction des paramètres issus des différentes modalités puis sur une phase de classification supervisée. La mise en oeuvre de cette approche repose sur la nécessité d'arriver à un modèle de prédiction pour lequel nous pourrons **décrire, expliquer et expliciter** les relations qui existent entre nos prédictions et les paramètres [Denis and Varenne, 2019]. L'un des intérêts d'une telle approche est de permettre aux enseignants de comprendre aisément ce qui impacte le niveau de compréhensibilité prédit.

— une approche dite *neuronale*, fondée sur des représentations extraites à l'aide de réseaux de neurones et également issues des différentes modalités. Ces représentations seront soit exploitées individuellement, soit fusionnées. La motivation ici est de voir si d'autres catégories de modèles issues de technologies plus actuelles et notamment de l'apprentissage profond, sont plus performants que les approches classiques basées sur des méthodes d'apprentissage que nous pourrions qualifier de classiques.

La première approche sera qualifiée dans la suite du document d'approche *interprétable* et la seconde d'approche *neuronale*.

Ces deux approches seront appliquées au corpus ESCAL décrit dans la première partie de ce mémoire, puis seront comparées par le biais des résultats de classification obtenus.

Chapitre 5

Mesure objective de la compréhensibilité : contribution des paramètres

5.1 Introduction

L'étude de la littérature en didactique des langues et la constitution du corpus ESCAL nous ont permis d'identifier un ensemble de phénomènes et de facteurs influençant la mesure de la compréhensibilité telle qu'elle peut être évaluée par les enseignants de langues. L'objectif de cette seconde partie, et notamment ce chapitre, est de se focaliser sur les méthodes et outils numériques qui vont permettre de représenter ces différents facteurs en les associant à des valeurs numériques. Il s'agit alors d'extraire un ensemble de paramètres des contenus audiovisuels traités en procédant par modalité tout en considérant différents niveaux de granularité. Après avoir défini les niveaux de granulatité auxquels nous nous référons, nous présenterons ensuite les paramètres extraits et leur lien avec les dimensions linguistiques relatives à la complexité lexicale et grammaticale, l'intelligibilité de la parole et le niveau de compréhensibilité global.

5.2 Modalités et niveaux de granularité

Les trois composantes, audio, texte et séquences d'images, constituant un document audiovisuel, peuvent être analysées séparément soit d'un point de vue global (en les considérant dans leur intégralité) soit plus finement, à différents niveaux de granularité.

La **composante texte** peut correspondre soit :
- à des sous-titres associés au document audiovisuel conçus par des professionnels et répondant à des normes (code couleur en fonction du type d'événement sonore retranscrit, longueur contrôlée des sous-titres...) ;
- à des sous-titres réalisés par des amateurs et que l'on appelle *fansub*, cependant, comme ils ne sont pas codifiés ils peuvent contenir différents types d'erreurs (typographie, erreurs d'orthographe, mauvaise accentuation...) ;
- à la transcription exacte de la parole au sein du document audiovisuel qui sera réalisée manuellement.

Dans les trois cas, la composante texte sera représentée au format de sous-titres avec : le numéro des sous-titres, les temps de début et de fin des sous-titres (ou *timecodes*) et le texte associé.

Le texte peut être analysé :
- dans son intégralité (après concaténation de chaque sous-titre) ;
- sous-titre par sous-titre, qu'ils soient professionnels, amateurs ou la transcription exacte ;
- mot par mot (ou lemme par lemme) ;
- lettre par lettre (ou caractère par caractère, un caractère étant une lettre ou un signe de ponctuation).

Le texte intégral correspond au niveau de granularité le plus élevé, l'étude au

niveau des lettres/caractères sera le grain le plus fin.

La **composante audio** correspond au signal issu de la bande sonore du document audiovisuel. Ce signal peut se décomposer en segments : il peut s'agir de segments de parole, de musique, de bruit et de silence mais il est aussi possible qu'un segment soit de plusieurs types (par exemple il peut contenir de la parole et de la musique superposées). Le choix peut être fait de traiter le signal audio tel quel ou de l'étudier après l'avoir débruité. Ensuite il peut être analysé à différents niveaux :

- dans sa totalité sur la durée du document ;
- sous-titre par sous-titre : dans ce cas, nous nous intéressons aux segments audio qui correspondent aux *timecodes* de début et de fin de chaque sous-titre ;
- par segment de parole (ou de non parole : musique, bruits) (un segment étant ici une zone temporelle continue), détectés et délimités à l'aide d'outil dédié comme un outil de détection automatique de l'activité vocale (noté *VAD* pour *Voice Activity Detection* en anglais) ;
- par phonème, les phonèmes sont obtenus à l'issue d'une conversion du texte vers les phonèmes soit à partir d'une transcription automatique de la parole, soit à partir d'une transcription manuelle du texte.

La **composante vidéo** correspond à la séquence d'images correspondant au document audiovisuel. Nous pouvons nous intéresser :

- par ensemble de la séquence d'images ;
- par séquence d'images alignées au début et à la fin des sous-titres à partir des *timecodes* ;
- par séquence d'images alignée aux segments de paroles/non parole ;
- par image (ou *frames*) constituant le signal vidéo.

Quelle que soit la composante considérée, si nous nous situons au niveau global, nous prendrons en compte la durée totale du document (D_{totale}) ; si nous nous positionnons au niveau du sous-titre la durée sera définie par ses *timecodes* de début et de fin (D_{ST}) ; au niveau d'un segment la durée dépendra du temps de début et de fin du segment ($D_{segment}$).

L'extraction d'un ensemble déterminé de paramètres pouvant contribuer à la prédiction du niveau de compréhensibilité et de ses différentes dimensions (lexicale, grammaticale, intelligibilité) présente un double intérêt. Elle permet d'interpréter les résultats en fonction de la contribution de chacun des paramètres, qu'il soit issu d'une modalité ou d'une combinaison de plusieurs modalités. Elle nous offre la possibilité de proposer une approche « interprétable ». D'autre part, cette approche repose sur l'exploitation d'outils de traitement automatique et de ressources lexicales existants, l'objectif de la thèse étant non pas de développer des outils spécifiques, mais de faire appel à des ressources à disposition. Afin d'obtenir les modèles les plus pertinents possibles, nous avons fait le choix de calculer certains des paramètres de plusieurs façons possibles, en faisant varier les outils et ressources numériques ainsi que les méthodes.

5.3 Paramètres liés à la complexité linguistique

Cette section est consacrée à l'ensemble des paramètres que nous avons extraits puis explorés pour prédire la complexité linguistique. Cela repose sur l'utilisation d'un ensemble des outils existants et accessibles pour réaliser cette extraction.

5.3.1 Outils et ressources pour l'analyse linguistique

Dans le chapitre précédent, nous avons vu que des éléments comme la longueur des phrases et la fréquence lexicale étaient des facteurs entrant en compte dans la compréhensibilité. Pour l'analyse de la complexité grammaticale, nous avons tout d'abord besoin d'un analyseur morphosyntaxique, c'est-à-dire un outil permettant d'analyser un texte au niveau de la phrase et de sa structure (syntaxe) et au niveau du mot et de sa forme (morphologie), ou de ressources lexicales apportant des informations complémentaires, comme la fréquence lexicale.

5.3.1.1 Ressources lexicales

Nous avons fait appel à plusieurs ressources :
— une base de données lexicales,
— un dictionnaire,
— une liste de connecteurs du discours,
— et une liste de mots n'ayant aucune valeur de sens, désignée généralement par liste de mots vides (ou *stoplist*).

Base de données lexicale : *Lexique 3*

Nous utilisons la ressource Lexique (version 3) [New et al., 2004], qui est une base de données lexicales, pour obtenir des informations sur 140.000 mots de la langue française. Celle-ci est disponible en ligne sous licence de type GNU[1]. Si d'autres bases de données lexicales citées dans la première partie ont été évoquées, le choix de Lexique 3 a été motivé par le fait que les mots étaient étudiés sur la base de sous-titres de films français, cela est intéressant, car le corpus conçu dans le cadre de cette thèse est basé sur un ensemble d'extraits issus de films.

Construite et maintenue depuis 1999, cette base de données propose, pour chaque mot :
— sa représentation phonémique,
— son nombre de lettres,
— son nombre de phonèmes,
— son nombre de syllabes,
— sa forme orthographique syllabée,
— sa fréquence d'occurrences ainsi que celle du lemme correspondant,

1. http://www.lexique.org/

— son nombre de morphèmes...

La fréquence d'occurrences est calculée d'après différents corpus et plus particulièrement d'après un corpus de 50 millions de mots issus de 9474 sous-titres de films. L'ensemble des informations disponibles est présenté en détail dans le manuel de Lexique 3 [New et al., 2005a].

Dictionnaire : *Wiktionnaire*

Lorsque nous étudions le vocabulaire d'un texte, le registre de langue des mots utilisés peut être utile à exploiter. Des mots sortant du registre courant peuvent être perçus comme plus difficiles, et donc augmenter la complexité du vocabulaire. Les dictionnaires sont des ressources idéales pour avoir une information sur les différents sens d'un mot, mais aussi sur ses différents registres, celui-ci varier selon la façon dont un mot est utilisé si celui-ci est polysémique. Par exemple le mot *chien* appartient au registre courant quand il est utilisé pour désigner l'animal, mais il appartient au registre familier quand il est utilisé pour dénigrer une personne. C'est pourquoi, dans notre étude, le dictionnaire Wiktionnaire[2] est utilisé. Chaque définition associée à un mot a une étiquette qui correspond à un registre :

— courant, si la définition n'a pas d'étiquette associée au registre,
— familier,
— argotique,
— vulgaire,
— désuet pour les étiquettes comme « vieux », « désuet » ou encore « obsolète » données par Wiktionnaire), ce registre pourrait être assimilé au registre soutenu.

Pour les besoins de l'étude, nous représentons un mot par son pourcentage de définitions appartenant à un registre spécifique. J'ai également créé un nouveau registre englobant les registres familier, vulgaire et argotique et noté par la suite *fva*, un mot sera classé dans ce super-registre s'il a au moins une étiquette de ce registre.

Si nous reprenons l'exemple du mot *chien*, celui-ci a 33% de définitions appartenant au registre familier et 66% de définitions étiquetées dans le registre courant, de ce fait elle a 33% de définitions appartenant au registre *fva*.

Liste de connecteurs du discours : *LEXCONN*

Dans la sous-section 4.5.2 de la première partie, nous avons cité la cohérence de discours comme un élément qui peut jouer sur la compréhensibilité d'un document. Celle-ci est reflétée par la manière dont sont utilisés les marqueurs du discours. Pour identifier ces marqueurs, j'ai choisi d'utiliser LEXCONN [Roze et al., 2012]. Il s'agit d'un lexique regroupant un ensemble de connecteurs du discours de la langue française.

2. https://fr.wiktionary.org/

Liste de mots vides (ou *stoplist*)

Pour calculer les paramètres lexicaux sur le texte, il est nécessaire de filtrer les mots trop fréquents usuellement désignés comme mots vides (*stop word* en anglais). Bien que nécessaire à l'articulation d'un texte et à sa cohérence, le mot vide n'a pas de signification s'il est utilisé individuellement. Il est ainsi opposé à la notion de mots pleins. Un mot sera considéré comme non significatif à partir du moment où il est utilisé à une très haute fréquence dans une collection de textes écrits dans une langue cible. En effet, il ne permet pas de discriminer un texte d'un autre dans cette collection.

En français, les principaux mots vides sont : les articles (le, la, les...), les prépositions (à, dans, par...), les auxiliaires *avoir* et *être*, et leurs formes conjuguées. Les mots vides d'une langue sont rassemblés dans ce qu'on appelle une *stoplist*. La liste utilisée est issue de la bibliothèque logicielle NLTK (Natural Language Toolkit) développée en Python et dédiée au traitement automatique des langues[3].

5.3.1.2 Analyseurs morphosyntaxiques

L'étude [Falk et al., 2014] compare les résultats de différents analyseurs syntaxiques. Ainsi, Talismane [Urieli and Tanguy, 2013] a une *accuracy* de 97,8% sur le corpus d'apprentissage French TreeBank (FTB) [Abeillé et al., 2003]. Sur le corpus de mots nouveaux, il étiquette correctement 81,45% de l'ensemble des mots, toutes catégories grammaticales confondues (contre 85,45% pour l'outil de Stanford [Manning et al., 2014] qui est l'analyseur le plus performant).

Talismane

Bien que n'étant pas considéré comme le meilleur outil d'étiquetage morphosyntaxique dans cette étude, nous avons choisi l'outil Talismane car il est facilement accessible et il est possible d'optimiser son temps de traitement en l'utilisant en mode serveur/client. Il permet de réaliser les opérations suivantes sur le texte fourni en entrée :

- segmentation en mots,
- segmentation en phrases,
- étiquetage morphosyntaxique pour attribuer des informations grammaticales à chaque mot du texte (genre, nombre, catégorie grammaticale...),
- repérage des dépendances syntaxiques pour préciser les relations entre les mots d'une même phrase.

Plus de détails sur le fonctionnement de cet analyseur morphosyntaxique peuvent être trouvés dans la documentation de l'outil[4].

L'utilisation de Talismane et des ressources lexicales citées précédemment vont nous permettre d'extraire un ensemble de paramètres candidats pour réaliser la tâche de prédiction de la complexité du vocabulaire et de la complexité grammaticale.

3. http://www.nltk.org/
4. http://joliciel-informatique.github.io/talismane/

5.3.2 Paramètres liés à la complexité du vocabulaire

5.3.2.1 Fréquence lexicale

Dans la première partie de ce manuscrit, la fréquence lexicale est citée régulièrement comme étant liée à la difficulté de compréhension, que ce soit pour la compréhension écrite ou la compréhension orale. Cette information joue donc un rôle important au niveau de la complexité du vocabulaire.

Listes de mots fréquents

Une approche traditionnelle pour analyser la fréquence lexicale est d'étudier la proportion de mots présents ou absents d'une liste de mots considérés comme fréquents dans une langue cible.

Pour le français, la première liste qui a été utilisée pour l'étude de la lisibilité est celle de Gougenheim [Gougenheim et al., 1964]. Pour s'inscrire dans un cadre plus en adéquation avec le contexte applicatif qui sera tourné vers l'apprentissage des langues étrangères, nous avons choisi d'utiliser une liste plus récente établie par le lexicologue Étienne Brunet pour aider les personnes chargées de la rédaction des programmes scolaires. Cette liste rassemble les 1500 mots les plus fréquents de la langue française[5]. Elle contient moins d'occurrences que la liste de Gougenheim (qui compte plus de 8000 mots), mais a l'avantage de rendre compte de la langue que lisent les élèves francophones, ce qui se rattache davantage au contexte de la thèse. Les mots sont extraits de plusieurs types de sources (littéraires ou non) et ont été ramenés à leur forme lexicale de base (ou lemme). Pour chaque document audiovisuel, à partir de l'intégralité de la transcription associée (d laquelle a été retirée la *stoplist*), nous avons calculé la proportion de mots ramenés à leurs lemmes présents dans la liste de Brunet.

Ce calcul est inspiré de la littérature relative à la lisibilité. Comme le sujet porte sur la difficulté de compréhension de vidéos authentiques issues de films, nous nous sommes également intéressée à des ressources dont les fréquences lexicales ont été calculées à partir de sous-titres de films.

Fréquence lexicale issue de Lexique 3

Dans la base de données Lexique 3, il est possible d'avoir des informations sur la fréquence lexicale dans les sous-titres de films [New et al., 2005b]. Celle-ci désigne le nombre total d'apparitions (par million d'occurrences) des mots (ou des lemmes) dans le corpus utilisé pour construire Lexique 3. En étudiant ces informations, un ensemble de paramètres ont été calculés :

Proportion des mots rares et fréquents : Dans Lexique 3, l'auteur donne des indications sur la fréquence ou la rareté des mots.

Les mots ayant une fréquence :

— inférieure à 5 sont considérés comme **très rares**,
— comprise entre 5 et 10 sont considérés comme **rares**,
— comprise entre 20 et 50 sont considérés comme **fréquents**,

5. https://eduscol.education.fr/186/liste-de-frequence-lexicale

— supérieure à 50 sont **très fréquents**.

Aucune indication n'est donnée sur les mots dont la fréquence est comprise entre 10 et 20. Ces informations nous permettent de classer les mots dans deux catégories : les **mots rares**, qui ont une fréquence inférieure à 20 et les **mots fréquents**, qui ont une fréquence supérieure à 20.

Les seuils fournis dans Lexique 3 sont cependant des approximations, usuellement utilisées par les auteurs de la base de données. Pour déterminer par la suite les fréquences des lemmes qui permettent le mieux de définir la complexité lexicale, nous étudions d'une part les lemmes dont la fréquence est **inférieure** à un seuil X, notée Freq_Lemme$_{\leq X}$ (avec $X \in [0.5, 20]$) et d'autre part les lemmes dont la fréquence est **supérieure** à un seuil X, notée Freq_Lemme$_{\geq X}$ (avec $X \in [20, 50]$).

La proportion de lemmes ayant une fréquence inférieure à 20 reflète la proportion de mots rares contenus dans le texte. La présence de mots rares contribuant à complexifier le vocabulaire, plus la proportion est élevée, plus le document est compliqué en termes de vocabulaire. De façon analogue, la présence de mots fréquents simplifie le vocabulaire : quand la proportion de lemmes avec une fréquence supérieure à 20 augmente la complexité du vocabulaire diminue.

Distribution de la fréquence d'occurrences Lorsque nous étudions une distribution, les percentiles sont les valeurs de la variable qui divisent la population en 100 groupes égaux en nombre. Par exemple, le $25^{ième}$ percentile correspond à la valeur de la variable pour laquelle 25% de la population a une valeur inférieure à cette valeur (représenté en bleu sur la figure 5.1) et 75% de la population a une valeur supérieure.

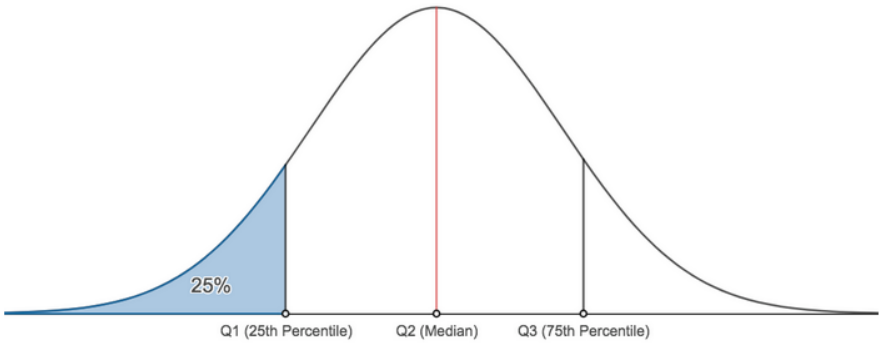

FIGURE 5.1 – Illustration de percentiles

Les percentiles de la fréquence des lemmes donnent eux aussi une information sur la proportion de mots rares et de mots fréquents. Par exemple, un dixième percentile ayant une valeur de 50 signifiera que 10% des mots présents dans l'extrait auront un lemme ayant une fréquence

comprise entre 0 et 50. Tandis qu'un dixième percentile ayant une valeur de 5 signifiera que 10% des mots auront un lemme dont la fréquence sera comprise entre 0 et 5. Dans ce cas, cela voudra dire que l'extrait contient au moins 10% de mots rares, voire très rares. Ainsi, des valeurs de percentiles de plus en plus basses mettent en avant la présence de mots très rares dans l'extrait et reflètent une augmentation de la complexité du vocabulaire.

Dans la suite nous étudions les dix premiers percentiles de la fréquence des lemmes, noté Perc_X (avec $X \in [1, 10]$).

Pourcentage de personnes connaissant la définition des lemmes issu de Lexique

En plus de l'étude des lemmes, à l'aide des fréquences lexicales qui sont calculées à partir d'un corpus de films français, nous avons exploité une information complémentaire de Lexique 3 qui est la proportion de personnes qui connaissent les définitions des mots présents dans la base de données. Un sondage intitulé *Combien de mots connaissez-vous ?* a été mis en place. Les participants y indiquent pour chaque mot si oui ou non ils connaissent la définition de son lemme. Ainsi, le pourcentage de participants ayant répondu par l'affirmative, désigné comme fréquence *subjective objective*, a été ajouté comme information. Cette information permet de prendre en considération une autre vision de la fréquence. Interroger des personnes qui ne sont pas forcément du domaine de la linguistique sur leur connaissance des mots permet ainsi de savoir s'ils y sont fréquemment exposés.

Pour chaque texte traité, j'ai ainsi extrait le pourcentage moyen des personnes ayant répondu qu'elles connaissaient la définition des lemmes qui y sont présents.

Proportion de mots inconnus

Bien que Lexique 3 soit basé sur une vaste base de données de sous-titres de films, il reste cependant possible que certains mots ne soient pas connus de la base de données lexicales Lexique 3. Il peut s'agir de :

- néologismes, mots-valises, acronymes,
- mots du registre familier ou vulgaire, issus du langage urbain (verlan),
- mots du registre soutenu,
- mots en langue étrangère,
- noms propres qui correspondent à des villes ou de quartiers, ou encore des marques...

Dans l'étude que nous avons menée auprès d'enseignants de FLE, et décrite dans la partie 1 de ce document, il est apparu dans les commentaires des annotateurs que des noms communs inconnus peuvent être des mots qui sont perçus comme compliqués par les apprenants, et d'un point de vue didactique, leur présence augmente la complexité lexicale. Par exemple des mots comme *garde-chiourmes* présents dans notre propre corpus ne font pas partie de Lexique 3.

Les noms propres sont intéressants à quantifier, car ils font intervenir la notion de contexte (personne, lieu, événement... évoqués dans le texte traité) et de culture (éléments de contexte général censés être plus ou moins connus). La connaissance de ces noms permet de comprendre le contexte. Ceci a été confirmé par les annotateurs de notre corpus, la présence de mots nécessitant une référence culturelle étant un facteur de complexité du vocabulaire. Ne pas connaître les noms propres dont il est question influence la compréhension de l'apprenant. Ainsi, la présence de beaucoup de noms propres augmente la complexité du vocabulaire. Trois paramètres liés aux mots inconnus et aux noms propres sont calculés (les mots de la *stoplist* sont ignorés). Il s'agit, pour un texte donné de :

— la proportion de mots inconnus,
— la proportion de ces mots qui ne sont pas des noms propres,
— la proportion de noms propres.

5.3.2.2 Richesse lexicale

Comme expliqué dans la première partie, un texte sera considéré comme riche s'il contient un grand nombre de mots différents. L'étude de la richesse lexicale, qui est un « concept quantitatif » [Thoiron and Arndaud, 1992], se décline en deux branches : la diversité et la densité lexicale.

Diversité lexicale

La diversité lexicale est parfois assimilée à la richesse lexicale par les linguistes, et elle est considérée comme un indicateur efficace de la difficulté d'un texte. Cette mesure prend en considération tous les mots distincts présents dans un segment de texte traité, qu'il s'agisse de mots pleins ou de mots grammaticaux. Parmi les mesures les plus connues de diversité lexicale, il est possible de citer :

— Le Type-Token Ratio (TTR) : il s'agit de la mesure la plus traditionnelle de diversité lexicale [Templin, 1957]. Il se calcule à l'aide de la formule :

$$\mathrm{TTR} = \frac{V}{N} \tag{5.1}$$

avec V le nombre de mots **distincts** (Types) et N le nombre total de mots (Tokens). Cette mesure pouvant être très influencée par la longueur du texte traité, des alternatives ont été proposées.

— Le Root Type-Token Ratio (RTTR) ou index de Guiraud [Guiraud, 1954]. Cette mesure vise à répondre aux problèmes liés aux effets de longueur du TTR et se calcule avec la formule suivante :

$$\mathrm{RTTR} = \frac{V}{\sqrt{N}} \tag{5.2}$$

Nous pouvons travailler avec des textes de longueur variable. Pour créer une mesure cohérente de la complexité du vocabulaire qui soit indépendante de la longueur des documents traités, nous avons fait le choix par la suite d'utiliser l'index de Guiraud comme paramètre pour représenter de la diversité lexicale.

Densité lexicale

Contrairement à la diversité lexicale, la densité se focalise uniquement sur les mots **lexicaux**, c'est-à-dire ayant un sens *plein* comparativement aux mots grammaticaux. Après avoir retiré les mots de la *stoplist*, la densité lexicale (DL) est calculée avec la formule :

$$\mathrm{DL} = \frac{V}{N} \tag{5.3}$$

Dans mon étude, j'utilise une variante de la mesure de densité lexicale en calculant le pourcentage de mots lexicaux distincts.

5.3.2.3 Niveau CECRL

Dans l'analyse des commentaires des enseignants ayant participé à l'expérimentation, le niveau CECRL des mots employés est un indicateur intéressant. De leur point de vue, la présence de mots à partir du niveau B1 est une source de difficulté pour les apprenants.

Partant de ce constat, j'ai calculé la proportion de mots de niveaux A1 ou A2. Pour cela, j'ai utilisé une liste des mots français de niveau A1 et A2 qui a été construite à partir de manuels de référence [Beacco and Porquier, 2007, Beacco et al., 2004] répertoriant l'ensemble des mots devant être acquis à ces niveaux d'apprentissage. Ceci permet indirectement d'obtenir la proportion de mots qui sont à un niveau CECRL supérieur et d'estimer la difficulté potentielle liée à la présence de ces mots.

5.3.2.4 Registre de langue

Dans les commentaires des annotateurs de l'expérience, l'utilisation de registres de langue différents du registre courant est évoquée à plusieurs reprises comme un facteur de complexité lexicale. Pour inclure les facteurs qui y sont liés, plusieurs paramètres sont calculés. À partir du Wiktionnaire il est possible de calculer le poids de chaque registre de langue (courant, familier, etc.) dans le texte traité. Pour chaque mot considéré, nous considérons le pourcentage de définitions qui sont classées comme appartenant à chacun des registres. Il est possible ensuite de calculer le poids de chaque registre en sommant le ratio des définitions qui lui correspondent. Par exemple pour la phrase *Le chien aboie.*, le mot outil *Le* étant ignoré, le calcul portera sur les mots *chien* et *aboie*. Admettons que le nom *chien* compte trois définitions dont deux relevant du registre courant et une du registre familier, alors leur ratio respectif sera de 2/3 et de 1/3. Si le verbe *aboie* a deux définitions (une dans le registre familier et une autre dans le registre courant), alors leur poids respectif sera de 1/2 et le poids du registre familier, pour l'ensemble du texte traité, sera de 0,83.

Ainsi, pour chaque texte traité, nous pouvons extraire six paramètres représentant le poids :

— du registre courant (normalisé par le nombre total de mots pour minimiser l'influence de la longueur du texte traité sur le calcul du paramètre),

— du registre désuet,
— du registre familier,
— du registre argotique,
— du registre vulgaire,
— des registres familier, vulgaire et argotique (représenté par le super registre *fva*).

5.3.2.5 Longueur des mots

La longueur des mots peut donner une indication quant à la complexité lexicale de celui-ci. Les mots courts étant les plus fréquents, nous partons du principe que plus un mot sera long, plus il sera complexe dans sa forme, mais aussi dans son sens. Deux paramètres, en lien avec la longueur des mots, ont été considérés (les mots appartenant à la *stoplist* sont ignorés) :

— le nombre moyen de syllabes par mot : la base de données Lexique donnant une information sur le nombre de syllabes de chacun des mots, il est possible de calculer le nombre moyen de syllabes par mot,
— le nombre moyen de lettres par mot : la longueur moyenne des mots peut-être mesurée simplement en récupérant la longueur de chaque mot en nombre de lettres et en la divisant par le nombre total de mots.

5.3.2.6 Prédictibilité des mots

Il est possible de retrouver dans la base de données lexicales Lexique 3 deux informations nous donnant une indication sur la prédictibilité des mots :

— la position du point d'unicité orthographique (calculée sur la base des lemmes),
— le nombre de voisins orthographiques (calculé à l'aide de toutes les entrées de Lexique).

Ceci nous permet de calculer deux autres paramètres.

Position moyenne du point d'unicité orthographique

La position moyenne du point d'unicité orthographique des mots du texte traité a été calculée avec la formule suivante :

$$\text{Point\_unicité\_moyen} = \frac{\sum_{i=1}^{n} \text{Point\_unicité\_mot}_i}{N} \tag{5.4}$$

Nombre moyen de voisins orthographiques

À partir du nombre de voisins orthographiques de chaque mot, nous calculons, pour chaque texte traité, le nombre moyen de voisins orthographiques par mot considéré.

Plus les deux paramètres précédemment présentés sont élevés, plus le lexique est considéré comme complexe, car il contient des mots difficiles à prédire.

5.3.2.7 Polysémie

Un mot ayant plusieurs sens peut être plus compliqué à traiter qu'un mot ayant une seule définition. Allouer le sens correct à un mot polysémique dans un contexte spécifique peut dépendre de la connaissance des diverses définitions de leur adéquation avec le contexte courant.

À partir des informations du Wiktionnaire deux paramètres permettant d'associer au texte traité un certain degré de polysémie, après avoir retiré du texte les mots de la *stoplist* et avoir ramené lorsque c'était possible chaque mot du texte à son lemme, ont été calculés :

— le nombre moyen de définitions par lemme,
— il est possible qu'il y ait des cas de redondance de mots, mais un mot polysémique, une fois que son sens a été désambiguïsé dans le contexte, ne sera plus une source de difficulté : nous avons donc également calculé le nombre moyen de définitions par lemme **distinct**.

Nous avons identifié un ensemble de 21 paramètres susceptibles de rentrer en jeu dans l'estimation de la complexité du vocabulaire. Ils sont tous répertoriés dans le tableau 5.1.

5.3.3 Paramètres liés à la complexité grammaticale

La complexité grammaticale est évaluée en considérant les aspects morphologiques (liés à la forme des mots) et syntaxiques (liés à la structure des phrases).

5.3.3.1 Complexité morphologique

En français, les verbes possèdent un grand nombre de formes, dépendant du temps et du mode. L'apprentissage de la langue commence par les temps simples de l'indicatif pour aller vers des temps composés et des modes plus complexes. Pour prendre en compte l'influence des temps verbaux sur la complexité grammaticale d'un texte traité, nous avons calculé dans un premier temps :

— le nombre de temps verbaux différents,
— la fréquence des temps verbaux simples,
— la proportion de verbes conjugués au temps du passé.

La proportion de verbes conjugués à un temps spécifique est aussi calculée, en se basant sur le référentiel de l'Alliance Française [Française and Chauvet, 2008]. Pour avoir des informations sur les temps verbaux, il est possible d'exploiter trois informations de Talismane. La première information est le « tag » donné par Talismane qui donne une indication sur le mode, s'il s'agit d'un verbe. La seconde information peut être trouvée dans les informations morphosyntaxiques complémentaires, où il est possible de connaître le temps verbal utilisé, enfin une information sur la présence de l'auxiliaire est également donnée par Talismane, en indiquant soit qu'il s'agit d'un auxiliaire au présent (« aux_tps »), soit d'un auxiliaire pour la voix passive (« aux_pass »).

TABLE 5.1 – Paramètres et mesure de la complexité du vocabulaire

Phénomène	Description du paramètre	Notation
Fréquence lexicale	Proportion de mots distincts avec un lemme de fréquence $< X (X \in [0.5, 20])$	$\text{Freq\_Lemme}_{\leq X}$
	Proportion de mots distincts avec un lemme de fréquence $> X (X \in [20, 50])$	$\text{Freq\_Lemme}_{\geq X}$
	$X^{\text{ième}}$ percentile de la fréquence des lemmes ($X \in [1, 10]$)	Perc_X
	Pourcentage de personnes connaissant la définition d'une lemme	$\%_{\text{Lemmes\_Connus}}$
	Pourcentage de mots inconnus	$\%_{\text{Mots\_Inconnus}}$
	Pourcentages de mots inconnus qui ne sont pas des Noms Propres (NP)	$\%_{\text{Mots\_Inconnus\_SansNP}}$
	Pourcentage de Noms Propres	$\%_{\text{NP}}$
Diversité lexicale	Index de Guiraud (Root Type Token Ratio)	RTTR
Densité lexicale	Pourcentage de mots lexicaux distincts	$\text{DL}_{\text{sans répétitions}}$
Niveau CECRL	Pourcentage de mots de niveau A1 ou A2	$\%_{\text{A1-A2}}$
Registre de langue	Poids du registre désuet	$\text{Poids}_{\text{Désuet}}$
	Poids du registre familier	$\text{Poids}_{\text{Familier}}$
	Poids du registre argotique	$\text{Poids}_{\text{Argot}}$
	Poids du registre vulgaire	$\text{Poids}_{\text{Vulgaire}}$
	Poids des registres familier, vulgaire et argotique	$\text{Poids}_{\text{FVA}}$
Longueur des mots	Nb moyen de syllabes par mot	$\text{L}_{\text{Syllabe}}$
	Nb moyen de lettres par mot	L_{Lettre}
Prédictibilité des mots	Position moyenne du point d'unicité orthographique par mot	PU_{Orth}
	Nb moyen de voisins orthographiques	$\text{Voisins}_{\text{Ortho}}$
Polysémie	Nb moyen de définitions par lemme	$\text{Moyenne}_{\text{Déf}}$
	Nb moyen de définitions par lemme distinct	$\text{Moyenne}_{\text{Déf\_Distinct}}$

En didactique des langues et notamment en FLE, les temps verbaux simples sont considérés comme étant :

— le présent de l'indicatif,
— l'impératif présent,
— le passé composé.

À partir des informations combinées fournies par Talismane, il est possible de déterminer le temps individuel de chacun des verbes présents dans un texte étudié et de calculer :

— le nombre de temps verbaux différents (nb_temps),
— la proportion de temps verbaux simples,
— la proportion de verbes conjugués au temps du passé.

La proportion de verbes conjugués à des temps simples est calculée de la façon suivante :

$$\text{Temps}_{\text{Simple}} = \frac{\text{nombre verbes conjugués à un temps simple}}{N_{\text{Verbes}}} \tag{5.5}$$

N_{Verbes} correspond au nombre de verbes conjugués.

Pour calculer les verbes conjugués au temps du passé, nous ne prenons pas en compte le passé composé puisqu'il n'a parfois pas la valeur du passé (ce sera le cas pour la phrase : « Aujourd'hui, il a tout perdu.»). La proportion de verbes conjugués à des temps du passé est calculée avec :

$$\text{Temps}_{\text{Passé}} = \frac{\text{nombre verbes conjugués à un temps passé}}{N_{\text{Verbes}}} \tag{5.6}$$

Nous ajoutons aussi aux paramètres le pourcentage de verbes conjugués dans chacun des temps verbaux que nous pouvons identifier à partir des tags de Talismane. Pour prendre en compte l'effet des morphèmes sur la complexité morphologique, nous avons calculé le nombre moyen de morphèmes par mot. Un morphème correspond à la plus petite unité constituant un mot, comme la racine, la partie flexionnelle (terminaison d'une conjugaison ou marque du genre et du nombre) et/ou une partie dérivationnelle (préfixe, suffixe). Par exemple, le mot *immangeables* compte 4 morphèmes : *in-, mange-, -able, -s*. L'information sur les nombres de morphèmes est obtenue via la base de données lexicales Lexique 3 pour chacun des mots (ou forme lexicale) qui y sont répertoriés. Le **nombre moyen de morphèmes par phrase** qui peut alors être calculé permet de caractériser également le texte traité. Un texte contenant en moyenne des phrases longues et beaucoup de morphèmes est peut-être considéré comme complexe du point de vue morphosyntaxique.

5.3.3.2 Complexité syntaxique

La complexité syntaxique désigne la complexité des phrases en termes de structure. Étudier la syntaxe revient à étudier les aspects structurels d'un texte et des phrases (ordre des mots, formulations, nombres de propositions, connecteurs...) [Akinci, 2005, Blache, 2010].

Dans [De Clercq, 2016], la complexité syntaxique est mesurée selon deux points de vue. L'auteur parle de complexité absolue par opposition à complexité relative. La première prend en compte la nombre d'éléments par phrase (longueur) ainsi que le nombre et du type des relations qui existent entre ces éléments (structure). La seconde est plutôt en lien avec la capacité de l'apprenant ou le la méthodologie d'apprentissage. Nous nous plaçons ici au niveau global en considérant l'ensemble du document.

La longueur des phrases peut se calculer de plusieurs façons, la plus utilisée, proposée par Hunt [Hunt, 1965] est le nombre de mots par phrase. Le paramètre de la complexité syntaxique associé sera **le nombre moyen de mots par phrase** dans le texte traité. Plus la moyenne est élevée, plus la complexité est considérée comme importante. Une seconde possibilité est de calculer la longueur des phrases en comptant le nombre de caractères [Ausloos, 2008].

C'est également une mesure très utilisée pour estimer la lisibilité, si le nombre de lettres excède un certain seuil, la phrase sera considérée comme complexe. Ainsi, **le nombre moyen de caractères par phrase** peut être un indicateur pertinent de la complexité grammaticale du texte traité, qui sera aussi considéré comme paramètre candidat par la suite.

Il y a également plusieurs façons d'analyser la structure d'une phrase et sa complexité, comme évoqué dans la première partie du manuscrit 4.3.2.2, il est possible d'étudier le nombre de propositions dans les phrases, mais aussi d'étudier les arbres syntaxiques.

Profondeur de l'arbre syntaxique

Dans la sortie de Talismane, le gouverneur d'un mot donné est indiqué par un nombre entier qui correspond à la position du gouverneur du mot dans la phrase. Si le mot est lui-même la racine de l'arbre, alors le nombre du gouverneur sera fixé à 0. À partir de cette indication, il est possible de récupérer le nombre de gouverneurs distincts dans une phrase. Nous pouvons ainsi mesurer la profondeur d'un arbre syntaxique en calculant **le nombre moyen de gouverneurs par phrase**.

Plus le nombre moyen de gouverneurs par phrase augmente, plus l'arbre syntaxique est profond et plus sa complexité augmente.

Nombre de propositions par phrase

Une phrase comptant beaucoup de propositions pourrait être perçue comme étant plus complexe. Il existe plusieurs types de propositions : les juxtapositions, les coordinations et les subordinations. Je me concentre sur les propositions les plus faciles à repérer grâce aux mots outils qui les introduisent : les coordinations et les subordinations. Pour pouvoir repérer les mots outils, il est nécessaire de traiter les textes sans ôter les mots de la *stoplist*.

À l'aide de la sortie de Talismane, il est possible d'obtenir une approximation du nombre de propositions coordonnées et subordonnées. Le « tag » donné à un mot par Talismane indiquera si un mot est une conjonction de coordination (CC) ou une conjonction de subordination (CS). Ces informations permettent de calculer **le nombre moyen de coordonnées et de subordonnées par phrase**. Plus la moyenne de coordonnées et de subordonnées par phrase est élevée, plus le texte est susceptible d'être complexe du point de vue grammatical.

Emploi de la voix passive

Avec l'étiquette « aux_pass », Talismane donne une information qui permet de savoir si la voix passive est utilisée. En décomptant, pour le texte traité, le

nombre de fois où cette information est donnée par Talismane, il est possible de calculer le **nombre moyen de voix passives par phrase** en calculant le nombre de fois que la voie passive est utilisée dans un texte avant de normaliser par le nombre total de phrases.

Les 11 paramètres en lien avec la complexité grammaticale sont répertoriés dans le tableau 5.2.

TABLE 5.2 – Paramètres et mesure de la complexité grammaticale

Phénomène	Description du paramètre	Notation
Temps verbaux	Nb temps verbaux différents	Nombre_Temps_Verbaux
	Proportion de verbes conjugués à un temps verbal simple	$\text{Temps}_{\text{Simple}}$
	Proportion de verbes conjugués au temps du passé	$\text{Temps}_{\text{Passé}}$
	Proportion de verbes conjugués à un temps verbal spécifique : TEMPS $\in$ {Présent, Imparfait, Passé_Composé, Passé_Simple, Conditionnel, Futur}	V_{TEMPS}
Morphèmes	Nb moyen morphèmes par mot	$\text{Morphèmes}_{\text{Mots}}$
	Nb moyen morphèmes par phrase	$\text{Morphèmes}_{\text{Phrase}}$
Longueur phrases	Nb moyen mots par phrases	L_{Mots}
	Nb moyen caractères par phrase	$L_{\text{Caractères}}$
Profondeur arbres syntaxiques	Nb moyen gouverneurs par phrase	$\text{Nombre}_{\text{Gouverneurs}}$
Nombre propositions	Nb moyen propositions subordonnées et coordonnées par phrase	$\text{Nombre}_{\text{Coord-Sub}}$
Voie passive	Nb moyen voies passives par phrase	$\text{Nombre}_{\text{Passive}}$

5.3.4 Cohérence du discours : marqueurs

Les marqueurs du discours permettent de consolider la structure d'un discours en ajoutant des relations explicites entre plusieurs éléments ou idées. Dans la littérature [Camiciottoli, 2004, Chaudron and Richard, 1986], il est dit que ces marqueurs favorisent une meilleure compréhension orale et qu'une mesure permettant de les dénombrer peut constituer un paramètre pertinent pour prédire la difficulté de la compréhension orale. Pour l'étude, je n'ai pas calculé le nombre total de marqueurs du discours, mais un nombre moyen dans l'ensemble de l'extrait. Pour ce calcul, nous avons utilisé LEXCONN. Pour faciliter le traitement, parmi ces connecteurs, nous avons uniquement conservé les marqueurs composés d'un seul mot et nous avons calculé le nombre moyen de marqueurs du discours par phrase en ne prenant en compte que les mots n'appartenant pas à

la stoplist. Plus la moyenne du nombre de marqueurs de discours est élevée, plus le discours est structuré, ce qui doit diminuer la difficulté globale du document considéré.

5.4 Paramètres liés à l'intelligibilité de la parole

Dans cette section, je présente les paramètres qui peuvent intervenir dans la mesure de l'intelligibilité et qui sont en lien avec les phénomènes cités dans la première partie de ce manuscrit. Pour la suite, la composante audio sera étudiée sous divers niveaux de granularité :

- en considérant la totalité du signal audio,
- par pseudo-tours de parole (ou sous-titres par sous-titres), en se basant sur la transcription exacte de la modalité texte et les timecodes de début et fin de chaque sous-titre, les pauses entre sous-titres étant ignorées, nous obtenons ainsi des mesures plus fines et nous nous assurons de se placer dans un cas idéal où toute la parole est considérée, ça n'est pas forcément le cas si nous détectons les segments de paroles avec un outil de détection d'activité vocale dont les performances peuvent être altérées dans le cas d'un environnement bruité.
- par segments de parole détectés à partir du signal audio avec de la VAD si nous nous référons aux sous-titres, nous ignorons les temps de pauses et de silences présents dans les zones délimitées avec les timecodes, l'utilisation de VAD, sensible à ces phénomènes, permet d'être encore plus fin même si des erreurs peuvent exister dans les résultats obtenus.

Pour chacune des granularités considérées, le choix a été fait d'appliquer les algorithmes et calculs présentés sur des fenêtres glissantes d'une seconde, avec un chevauchement de moitié (0,5 seconde). Avec toutes les mesures obtenues, nous calculons des statistiques descriptives : moyenne, médiane, écart-type, minimum et maximum. Si nécessaire, nous ajouterons à la nomenclature des paramètres calculés : le niveau de granularité (si le paramètre n'a pas été calculé à partir de l'intégralité du document) : ST, SEG ; et une information sur la statistique calculée notée $STAT$ (avec $STAT \in \{MOY, MED, MIN, MAX, ET\}$).

5.4.1 Rapport signal sur bruit

Le bruit environnant est une source sonore qui influe sur l'intelligibilité de ce qui est dit. Quand il s'agit d'évaluer le niveau de bruit, nous pensons immédiatement à calculer le rapport signal sur bruit.

5.4.1.1 Méthode classique : rapport des puissances

L'approche la plus traditionnelle qui a été testée consiste à se servir de la formule suivante :

$$\mathrm{SNR}_{\mathrm{RMS}} = \frac{\mathrm{Puissance}_{\mathrm{Signal}}}{\mathrm{Puissance}_{\mathrm{Bruit}}} = \left(\frac{A_{\mathrm{Signal}}}{A_{\mathrm{Bruit}}}\right)^2 \tag{5.7}$$

avec A_{Signal} la moyenne quadratique (RMS pour « Root Mean Square ») de l'amplitude du signal et A_{Bruit} la moyenne quadratique de l'amplitude du bruit.

Pour pouvoir calculer ce rapport, il est nécessaire d'extraire le bruit du signal. Nous utilisons la méthode de Ephraim et Malah [Ephraim and Malah, 1985], implémentée en Python[6]. Le bruit est obtenu en soustrayant le signal débruité au signal d'origine. La connaissance du signal de bruit, permet de calculer la puissance du bruit et la puissance du signal bruité et d'en déduire le rapport signal sur bruit.

5.4.1.2 Méthode du NIST

La première approche nécessite de faire des traitements sur le signal d'entrée pour extraire le signal de bruit. Cela passe par une phase de débruitage du signal qui peut être une source d'erreur pour le calcul final du SNR.

Une seconde approche que nous avons testée est le calcul du SNR avec l'outil STNR développé par le NIST [Wierzynski and Fiscus, 2000]. N'ayant pas accès au signal de bruit, le programme STNR du NIST ne va pas essayer de l'extraire, mais va faire une estimation du niveau de la parole et du niveau du bruit : cela permet de ne pas réaliser de modification sur le signal d'entrée au risque de le dégrader. Le programme va estimer le SNR d'un signal d'entrée à l'aide du rapport suivant :

$$\text{SNR}_{\text{NIST}} = 10 \log \frac{\text{Pic puissance parole}}{\text{Puissance moyenne bruit}} \tag{5.8}$$

où la puissance fait référence à la variance du signal (ou RMS) calculé sur des fenêtres glissantes de 20 ms avec un chevauchement de 10 ms.

L'estimation du niveau de parole et de bruit se fait en se basant sur des observations faites sur les distributions typiques de la puissance du bruit et de la parole. L'hypothèse qui est retenue est que la distribution de la puissance du signal complet est un mélange de deux distributions : une pour le bruit et une pour la parole (voir figure 5.2). Plus de détails sur le calcul peuvent être trouvés dans la thèse de Xavier Anguera[7].

5.4.2 Réverbération

La réverbération, évoquée dans la section 4.4.2.2, est un phénomène acoustique qui peut contribuer à diminuer l'intelligibilité de la parole. Ce phénomène peut être plus ou moins contrôlé en fonction des conditions de prise du son d'un signal audio, mais il n'y a pas de garantie que cela soit fait de manière systématique, c'est pour cela que la révérbération est prise en compte dans la suite de notre étude. Pour mesurer la réverbération au sein du signal audio,

6. https://pypi.org/project/logmmse/
7. http://www.xavieranguera.com/phdthesis/node49.html

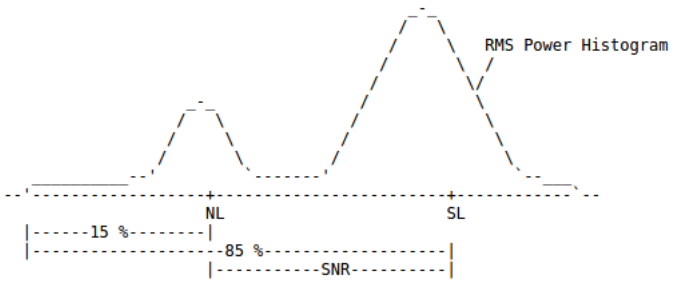

FIGURE 5.2 – Exemple de distribution de la puissance du bruit (NL) et de la parole (SL), issu de [Wierzynski and Fiscus, 2000]

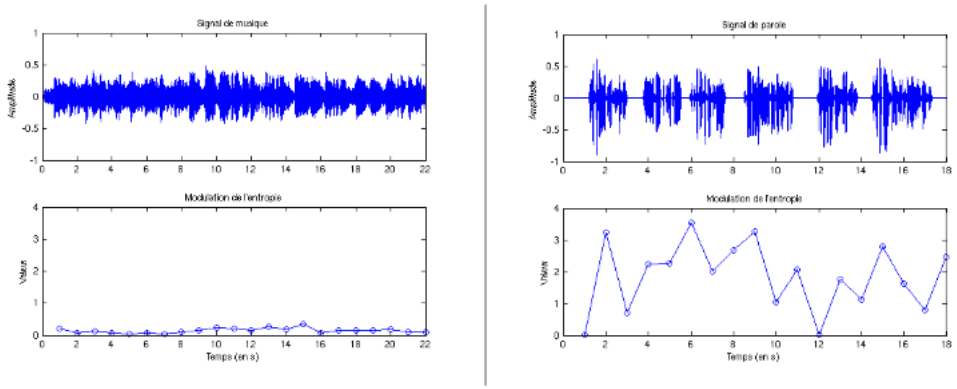

FIGURE 5.3 – Modulation de l'entropie pour la musique (à gauche) et pour la parole, issue de [Pinquier et al., 2003]

j'ai exploité la métrique SRMR (*Speech-to-Reverberation Modulation energy Ratio*) [Falk et al., 2010]). Il s'agit d'une mesure non intrusive qui permet d'évaluer la qualité de la parole et l'intelligibilité, en se basant sur la représentation spectrale de la modulation d'un signal qui contient de la parole. La réverbération s'obtient en calculant le rapport entre l'énergie dans les fréquences de modulation supérieures et l'énergie dans les fréquences de modulation inférieures.

5.4.3 Pureté de la parole

L'étude de Pinquier [Pinquier, 2004] met en avant que le signal de parole est plus désordonné que le signal de musique. En d'autres termes, le signal de parole contient plus de variations (voir figure 5.3). L'ordre du signal peut se mesurer en calculant son entropie, elle permet d'obtenir une information sur la pureté de la parole.

L'entropie du signal audio, notée Entropie, est calculée en se basant sur la

méthode utilisée dans [Pinquier et al., 2002]. Le signal est découpé en trames de 16 ms (sans recouvrement), et pour chaque trame l'entropie est estimée à partir de son histogramme, à l'aide d'un estimateur biaisé [Moddemeijer, 1989]. À la fin du processus, un signal audio est représenté par N valeurs d'entropie à partir desquelles nous pouvons calculer des statistiques descriptives pour décrire la pureté globale d'un signal (moyenne, médiane, maximum, minimum et écart-type). Nous considérons que plus la parole sera pure, plus le signal sera désordonné et son entropie élevée.

5.4.4 Débit de parole

Le débit de parole est un phénomène mesurable, qui se calcule communément à partir du nombre de syllabes prononcées par le locuteur sur la durée de l'énoncé. En nous basant sur cette définition, j'ai utilisé trois méthodes de calcul du débit de parole, une reposant à la fois sur les informations textuelles et sur le signal et deux mesures qui utilisent intégralement le signal audio.

5.4.4.1 Approximation par les informations textuelles

Il s'agit de faire une approximation du débit de parole en utilisant uniquement les informations issues de la composante texte : le nombre de syllabes (obtenu avec Lexique 3) et le nombre de lettres de la transcription exacte considérée dans son intégralité. Une telle approximation basée sur la composante textuelle permet d'obtenir ce que je désignerai par la suite comme un « pseudo-débit » car un débit se calcule traditionnellement en se basant intégralement sur le signal audio. Il s'agit d'une approximation qui peut être utile dans le cas où une estimation rapide du débit de parole veut être établie sans analyser le signal audio.

$$\text{Pseudo\_Debit}_S = \frac{\text{Nombre syllabes}}{D_{\text{totale}}} \tag{5.9}$$

$$\text{Pseudo\_Debit}_L = \frac{\text{Nombre lettres}}{D_{\text{totale}}} \tag{5.10}$$

Considérer la durée totale du document audiovisuel amène à inclure des temps de pauses et de silences dans le calcul, cela amène de l'erreur dans les résultats. Nous proposons donc aussi une méthode alternative qui va permettre d'obtenir des résultats plus précis en se basant sur la **durée totale de parole contenue dans le signal**. La durée de parole est obtenue en se basant sur les segments de parole du signal.

Pour trouver ces segments de parole, nous avons fait le choix de travailler avec l'outil de l'INA appelé `inaSpeechSegmenter`[8] [Doukhan et al., 2018]. Le système, utilisant des réseaux de neurones convolutionnels, a été entraîné sur des documents audio en français pour reconnaître et classer des segments homogènes de parole, de musique et de bruit (ce sont les segments qui ne seront ni de la musique, ni de la parole). La particularité de ce système est qu'il est capable de

8. `https://pypi.org/project/inaSpeechSegmenter/`

détecter la parole même si elle est superposée à la voix ou au bruit. Dans les contenus audiovisuels et notamment les contenus de fiction, les cas de parole et musique superposées et de parole dans du bruit sont fréquents. Compte tenu de notre cas applicatif, la capacité du système de l'INA à détecter la parole dans ce genre de circonstances m'a motivée à opter pour ce système de détection de la parole. Après avoir déterminé l'ensemble des segments contenant de la parole, il est possible de calculer le débit de parole (ou débit syllabique) ainsi :

$$\text{Debit}_{\text{INA}} = \frac{\text{Nombre Syllabes}}{\sum_i D_{\text{segment}}} \tag{5.11}$$

avec i le nombre total de segments contenant de parole.

5.4.4.2 Prise en compte des pseudo syllabes

Une deuxième approche fait intervenir la méthode développée au sein de l'équipe SAMoVA et qui repose sur la détection dans un signal de parole, d'une unité élémentaire appelée pseudo syllabe [Farinas et al., 2005]. Cette unité se définit par un nombre indéfini de consonnes (C) suivi par une voyelle (V). Une pseudo syllabe pourra par exemple avoir la forme simple `CV` ou des formes plus complexes comme `C...CV` (ou les trois petits points correspondent à un nombre indéfini de C). En se basant sur cette unité, **un nombre moyen de pseudo syllabes par seconde** sera calculé et noté Debit_{PS}.

5.4.4.3 Prise en compte des noyaux syllabiques

Une troisième approche est de calculer le débit de parole en repérant les noyaux syllabiques dans le signal [De Jong and Wempe, 2009]. Dans les signaux audio contenant de la parole, les pics d'intensité (dB) qui sont suivis et précédés par des baisses d'intensités peuvent considérés des noyaux syllabiques potentiels. Le logiciel `Praat` dédié à l'étude de la phonétique et de la linguistique à travers l'analyse de de fichiers audio[9] [Styler, 2013] permet notamment de détecter ces noyaux syllabiques en procédant aux étapes suivantes :

- Étape 1 : extraction de l'intensité du signal audio,
- Étape 2 : détection des pics d'intensité excédant un seuil donné. Ces pics sont considérés comme des noyaux syllabiques potentiels,
- Étape 3 : exclusion des pics qui ne sont pas précédés par une baisse d'intensité entre 2dB et 4dB par rapport au pic d'intensité considéré,
- Étape 4 : exclusion des pics non voisés,
- Étape 5 : sauvegarde des pics restants comme Noyaux Syllabiques (NS).

Le **débit de parole** est alors calculé ainsi :

$$\text{Debit}_{\text{NS}} = \frac{\text{Nombre NS}}{D_{\text{totale}}} \tag{5.12}$$

9. `https://praat.fr.softonic.com/`

5.4.5 Mesures d'intelligibilité

Si les paramètres précédents permettent d'obtenir des informations sur le contexte dans lequel la parole est énoncée (environnement sonore) et qui a un impact sur la perception de la parole par l'auditeur, une autre famille de paramètres peut être extraite de la composante audio pour qualifier l'intelligibilité de la parole elle-même.

5.4.5.1 Qualité de la prononciation

Pour estimer la qualité de la prononciation, nous nous positionnons au niveau des sous-titres qui correspondent à la transcription exacte d'un document audiovisuel. Nous utilisons :

— la transcription phonétique exacte du sous-titre $\text{Transcription}_{\text{exacte}}$,
— une transcription phonétique (*Speech-To-Phonemes* ou $S2P$) automatique du segment audio aligné au sous-titre considéré $\text{Transcription}_{\text{kaldi}}$.

Pour réaliser le $S2P$, nous avons utilisé une approche TDNNF, détaillée dans [Gelin et al., 2021], utilisant la boîte à outils kaldi[10] [Povey et al., 2011], et entraînée sur 150 heures de parole d'adultes du corpus Common Voice[11].

La distance de Levenshtein [Levenshtein, 1965] donne une mesure de la différence entre deux chaînes de caractères. Les chaînes de caractères comparées ici correspondent à des suites de phonèmes, où chaque phonème est représenté par un caractère unique.

Soit A et B deux chaînes de phonèmes, la distance de Levenshtein correspond au nombre minimal d'opérations nécessaires pour passer de la chaîne A à la chaîne B. Chaque opération (insertion, suppression, substitution de phonèmes) a un coût de 1. La distance correspond alors aux coûts cumulés.

Cette distance va donner une estimation de la différence entre ce qui est réellement dit (transcription exacte) et ce qui a été décodé par le système automatique. L'hypothèse qui est faite est que plus la prononciation est dégradée, plus la distance entre la transcription exacte et la transcription issue de Kaldi est élevée. Cette distance sera notée comme suit :

$$\text{Dist}_{\text{Lev}} = d(\text{Transcription}_{\text{exacte}}, \text{Transcription}_{\text{kaldi}}) \tag{5.13}$$

où d correspond à la distance de Levenshtein.

Une fois la distance de Levenshtein calculée pour chaque sous-titre, elle est normalisée par le nombre total de caractères (phonèmes) de la transcription exacte.

5.4.5.2 Fiabilité de la transcription automatique

Certains outils de transcription Speech-to-Text (ou STT) fournissent une mesure de leur taux de confiance quant à la qualité de la transcription qui a été

10. https://kaldi-asr.org/
11. https://commonvoice.mozilla.org/fr

produite. Par exemple, l'API Google Speech Recognition[12] permet d'avoir un score de confiance ente 0 et 1 sur l'ensemble de la transcription produite à partir d'un signal audio en entrée. Ce score de confiance, noté $\text{Conf}_{\text{Google}}$ est exploité ensuite comme un paramètre susceptible de prédire la qualité de l'intelligibilité de la parole. Nous avons supposé que le score de confiance était corrélé positivement à la qualité de la prononciation : plus la parole sera distincte, plus le score de confiance augmentera.

Pour pallier le bruit éventuel susceptible d'affecter la tâche de STT, le traitement a été réalisé avec et sans nettoyage du signal. Ce pré-traitement consiste à soustraire le bruit du signal d'origine. Ceci a été fait à l'aide de l'algorithme de `logmmse`, cité plus tôt [Ephraim and Malah, 1985], qui permet de réduire le bruit du signal et d'améliorer la parole.

5.4.5.3 Taux de parole détectée

Les documents audiovisuels qui nous intéressent incluent un ou plusieurs locuteurs. Nous considérons que le signal audio contient des temps de pause, mais minoritaires par rapport aux segments de parole. Sachant cela, il est possible d'estimer la qualité de la parole en estimant la quantité de parole détectée dans un signal audio, en se servant d'un outil de détection d'activité vocale (Voice Activity Detector ou VAD). Si le taux de parole détectée est faible, cela voudra alors dire que l'outil n'est pas arrivé à discriminer la parole dans le signal et donc que celle-ci n'aura pas été suffisamment intelligible.

Dans un premier temps, l'outil utilisé est celui de détection de parole/musique/bruit (PMB) de l'IRIT [Pinquier, 2004], pour réaliser la discrimination des zones de paroles. La méthode de modulation d'énergie à 4 Hertz est utilisée pour détecter les zones de parole : un signal de parole est caractérisé par un pic de modulation en énergie autour de la fréquence syllabique de 4 Hertz (une personne prononçant en moyenne quatre syllabes par seconde).

$$\text{Taux}_{\text{Parole}} = \frac{D_{\text{parole détectée}}}{D_{\text{totale}}} \tag{5.14}$$

Le taux de parole détecté augmentera avec la qualité de la parole dans le signal audio. L'outil utilisé comprend une phase de segmentation réalisée avec l'algorithme « Forward-Backward Divergence » [André-Obrecht and Jacob, 1997]. Le label parole, musique ou bruit est attribué à chacun des segments délimités lors de cette phase. Les segments de parole trouvés sont donc dépendants de la qualité de la segmentation et pour obtenir des segments fins, il est nécessaire de faire varier dans un premier temps les paramètres de l'algorithme de segmentation.

Il est intéressant d'utiliser un outil réalisant en simultané la délimitation des segments et leur étiquetage. Pour cela, l'outil WEBRTC[13] a également été testé. Il s'agit d'un outil de VAD gratuit et simple d'utilisation, et il est également possible de modifier la finesse de la détection en réglant l'agressivité de l'outil

12. `https://cloud.google.com/speech-to-text/docs`
13. `https://webrtc.org/`

(1 à 3) qui sert de filtre pour exclure les segments de non-parole. Avec une agressivité faible, le système a tendance à ne pas détecter les segments de silence, qui peuvent correspondre à une pause ou encore une hésitation, et ceux-ci seront inclus dans un segment de parole. Si l'agressivité est élevée, ces zones de silence seront prises en compte par le système. Pour un signal audio donné en entrée de l'outil, la différence entre la durée de parole détectée quand l'agressivité est à 1 et quand l'agressivité est à 3 est calculée (le système étant le plus exigeant quand l'agressivité est fixée à 3). Cette valeur sera notée $\text{Duree}_{\text{RTC}}$. Plus la différence entre les deux durées de parole tend à augmenter, plus la qualité de l'intelligibilité va diminuer.

5.4.6 Prédictibilité phonologique des mots

5.4.6.1 Point d'unicité phonologique

Dans le cadre de l'étude sur l'intelligibilité, et de façon analogue à l'étude des facteurs affectant la complexité lexicale, j'ai souhaité intégrer une mesure permettant de donner une indication sur la prédictibilité moyenne des mots à travers la suite de phonème qui représente leur prononciation : le **point d'unicité phonologique**, qui donne une information sur la position du phonème qui permettra de discriminer et d'identifier un mot entendu parmi plusieurs candidats possibles. [Dufour et al., 2002] permet de voir que, globalement, le nombre de mots candidats possibles a tendance à chuter considérablement quand le quatrième phonème est atteint, à partir de la cinquième syllabe le point d'unicité phonologique peut être atteint.

Dans la base de données lexicales Lexique 3, il est possible de connaître le point d'unicité phonologique des mots calculé **sur la base des lemmes**, une partie des mots peut donc être ignorée si les formes orthographiques considérées ne sont pas des lemmes, de plus. À noter que cette mesure se place dans le cas idéal : elle repose sur l'hypothèse que le lemme est prononcé correctement, ce qui n'est pas forcément le cas dans la réalité (la prononciation perçue étant affectée par l'accent, le débit de parole...). Avec cet ensemble d'approximations, nous obtenons une estimation de la position moyenne du point d'unicité phonologique du document traité. Sachant la position du point d'unicité phonologique de chaque lemme présent dans la partie textuelle associée au document audio à traiter, la **position moyenne du point d'unicité phonologique des lemmes** dans l'intégralité du document traité, notée par la suite PU_{phon}. Dans le cas où un mot est absent de la base de données lexicales la position moyenne du point d'unicité phonologique calculé sur l'ensemble de la base de donnée lexicale est utilisée.

Plus le point d'unicité des lemmes se situe loin dans le signal, plus leur identification est longue. Nous pouvons supposer que plus la position moyenne du point d'unicité est élevée, plus cela aura un impact négatif sur l'intelligibilité : le temps de décodage d'un mot est autant de temps perdu pour le décodage des mots suivants et l'interprétation du message.

5.4.7 Redondance son et image

Dans la section 4.4.5 de la première partie, l'impact de l'image sur la compréhension orale a été souligné, et cela a amené à inclure un paramètre permettant d'évaluer l'influence de la redondance entre le son et l'image sur l'intelligibilité. Cette mesure est basée sur des informations bimodales issues de l'image et de l'audio. Ces informations permettent de calculer le pourcentage d'images qui contiennent des visages qui coïncident avec des zones de parole. Nous supposons que les zones de parole où il est possible d'avoir accès à des visages sont potentiellement plus intelligibles. Les mouvements du visage (notamment les mouvements labiaux) ainsi que les expressions faciales pouvant faciliter la compréhension du message prononcé par un interlocuteur. Nous supposons que plus il y a de superpositions entre des zones de parole et des images contenant des visages, plus la parole est intelligible. Pour alléger les traitements, les zones de paroles considérées sont celles obtenues à l'aide des *timecodes* des sous-titres correspondant à la transcription exacte de la parole, en faisant ce choix des visages peuvent se superposer à des zones de silence ignorées. Il est également possible que des visages soient détectés, mais qu'ils ne correspondent pas à ceux du (des) locuteur(s) : nous pouvons avoir ce cas pour une *voix-off* par exemple. J'ai fait le choix d'intégrer ces approximations en considérant que la mesure telle quelle pouvait constituer un estimateur suffisant pour la prédiction de l'intelligibilité.

La détection de visages dans les images se fait avec l'outil YOLO (You Only Look Once) V3 [Redmon et al., 2016], qui a été entraîné pour la tâche de détection de visages sur le corpus `WIDER FACE`[14] [Yang et al., 2016]. Le but ici est de détecter si une image contient au moins un visage occupant au minimum 10% de l'image. Ce seuil, fixé arbitrairement, considère qu'un visage doit être assez grand et visible à l'écran pour que le spectateur puisse exploiter des informations telles que les expressions faciales et les mouvements labiaux, pour l'aider dans sa compréhension orale. Le nombre total d'images contenant au moins un visage en présence de parole est normalisé par le nombre total d'images de la séquence vidéo traitée. Le paramètre ainsi obtenu sera noté dans la suite $\text{Visage}_{\text{Parole}}$

Les 15 paramètres calculés en lien avec l'intelligibilité sont répertoriés dans le tableau 5.3.

5.5 Paramètres complémentaires liés à la compréhensibilité globale d'un document

Certains paramètres identifiés dans la première partie sont susceptibles d'influer sur la compréhensibilité globale d'un document sans être liés à la complexité du vocabulaire, de la grammaire ou à l'intelligibilité. Cette section est consacrée à la présentation de ces paramètres.

14. `http://shuoyang1213.me/WIDERFACE/`

TABLE 5.3 – Paramètres et mesure d'intelligibilité

Phénomène	Description du paramètre	Notation
Rapport Signal sur Bruit	Rapport des puissances	SNR_{RMS}
	Rapport avec l'outil STNR du NIST	SNR_{NIST}
Débit de parole	Pseudo-débit : Nb moyen syllabes par seconde	$Pseudo\_Debit_S$
	Pseudo-débit : Nb moyen lettres par seconde	$Pseudo\_Debit_L$
	Nb moyen syllabes par seconde, durée de parole calculée à l'aide de l'outil de VAD de l'INA	$Debit_{INA}$
	Nb pseudo-syllabes par seconde	$Debit_{PS}$
	Nb noyaux syllabiques par seconde	$Debit_{NS}$
Qualité prononciation	Distance de Levenshtein entre la transcription exacte et la transcription kaldi	$Dist_{Lev}$
Fiabilité transcription	Niveau de confiance de transcription du STT de Google	$Conf_{Google}$
Taux parole	Proportion de parole dans l'ensemble du signal	$Taux_{Parole}$
	Différence de durée de parole détectée par l'outil RTC VAD de Google en fonction du réglage de l'agresivité (1 et 3)	$Diff_{RTC}$
Prédictibilité parole	Position moyenne du point d'unicité phonologique	PU_{phon}
Pureté parole	Entropie de la parole	Entropie
Réverbération	Réverbération	SRMR
Redondance audio/image	Nombre moyen de visages superposés à des zones de parole	$Visage_{Parole}$

5.5.1 Voisinage phonologique

Dans le paragraphe 4.4.3, il est expliqué qu'un nombre de voisins phonologiques élevé peut nuire à la tâche de compréhension orale. Le nombre de voisins phonologiques par mot est renseigné dans la base de données lexicales Lexique 3. Nous calculons à partir de ces informations **le nombre moyen de voisins phonologiques par mot**. L'hypothèse est la suivante : si le nombre moyen de voisins phonologiques augmente, alors le temps accordé à identifier chaque mot tend à augmenter, nuisant à la compréhension globale du message.

5.5.2 Longueur du document traité

Dans la sous-section 4.5.1, la longueur d'un document traité est évoquée comme une source potentielle de difficulté. Plus il sera long, plus il peut être perçu comme globalement compliqué. La longueur des documents traités doit donc être incluse parmi les paramètres étudiés. Ainsi, pour chacune des composantes la longueur sera représentée par :

— la durée (en secondes) (D_{totale}) pour la composante audio,
— le nombre total de phrases ($N_{phrases}$), de mots (N_{mots}) et de phonèmes ($N_{phonemes}$) pour la composante texte,
— le nombre total d'images (N_{images}) pour la composante vidéo.

5.5.3 Quantité d'informations visuelles

Dans la sous-section 4.5.3, nous avons évoqué la quantité d'informations visuelles comme une source potentielle de difficulté globale à cause de la charge cognitive qu'une forte quantité d'informations peut amener chez le spectateur. Pour étudier ceci, nous nous sommes concentrée sur trois aspects :

— la quantité d'objets présents dans l'image.
L'outil YOLO a été utilisé pour déterminer dans un premier temps le nombre de personnes ($N_{personnes}$) et le nombre d'objets (c'est-à-dire tous les objets détectés qui ne sont pas des personnes) (N_{objets}) présents dans une image.
— le taux d'occupation spatiale de l'image.
C'est le rapport entre la surface totale occupée par des objets dans l'image et la surface de l'image, notée $Taux_{occupation}$, cette mesure prend en compte à la fois les personnes et les objets détectés à l'écran à l'aide de YOLO. Un taux d'occupation faible peut indiquer que les éléments à l'écran occupent très peu de place et/ou qu'ils sont petits (ceux-ci seront plus durs à exploiter pour le spectateur). Un fort taux d'occupation peut suggérer soit que les éléments à l'écran sont peu nombreux, mais très volumineux soit qu'au contraire il y en a beaucoup de tailles variables (dans ce cas, ce sera plus compliqué à comprendre pour le spectateur).
— l'étude du flot optique.
Elle consiste à étudier des images consécutives dans une vidéo pour analyser le mouvement des objets présents. Il a été calculé à l'aide de l'outil OpenCV [Bradski and Kaehler, 2000], il sera noté $Flux_{opt}$. La connaissance du flot optique de tous les objets présents à l'image permet de calculer des statistiques sur le flot optique global dans la vidéo. Nous supposons que plus les objets se déplacent dans la vidéo, plus celle-ci est difficile. En supposant que beaucoup de mouvements à l'écran peut nuire au traitement de l'information chez le spectateur, l'étude des mouvements peut être un apport pour l'estimation de la difficulté globale d'un document audiovisuel.

Pour chacun des aspects étudiés, nous allons nous intéresser aux statistiques

descriptives obtenues à partir de l'ensemble des données extraites, à savoir : **la moyenne, la médiane, le minimum, le maximum et l'écart-type**.

L'ensemble des 40 paramètres qui ont été présentés dans cette section sont récapitulés dans le tableau 5.4.

TABLE 5.4 – Paramètres et mesure du niveau de compréhensibilité.

Phénomène	Description du paramètre	Notation
Voisinage phonologique	Nb moyen voisins phonologiques par mot	$\text{Voisins}_{\text{Phon}}$
Longueur extraits	Durée (en secondes)	D_{totale}
	Nb phrases	$\text{N}_{\text{phrases}}$
	Nb mots	N_{mots}
	Nb phonèmes	N_{phons}
	Nb images dans la vidéo	N_{images}
Quantité informations visuelles	Nb objets à l'écran	N_{objets}
	Nb personnes à l'écran	$\text{N}_{\text{personnes}}$
	Taux d'occupation spatiale	$\text{Taux}_{\text{occupation}}$
	Flux optique	Flux_{opt}

5.6 Bilan

Dans ce chapitre, nous avons présenté l'ensemble des paramètres qui ont été calculés soit directement à partir des modalités audio, vidéo et texte, soit qui en sont dérivés. ces paramètres sont susceptibles de rentrer en jeu dans un modèle de prédiction de la compréhensibilité globale.

La première partie de notre étude avait permis de conclure que la complexité du vocabulaire et de la grammaire et l'intelligibilité de la parole étaient des éléments influençant la difficulté de compréhension globale de documents authentiques. Nous avions également identifié un ensemble de phénomènes et de facteurs qui pouvaient entrer en jeu dans la complexité de chacune de ces dimensions. Dans cette seconde partie, notre première contribution a été de traduire ces différents facteurs en familles de paramètres calculables à partir des modalités analysées, et ce, à différents niveaux de granularité. Ces différentes familles de paramètres sont synthétisés dans les tables 5.1, 5.2, 5.3 et 5.4, en fonction de la dimension considérée.

Notre objectif est maintenant d'utiliser ces jeux de paramètres pour des tâches de prédiction du niveau de compréhensibilité. Une nouvelle fois, nous aborderons la question en considérant les différentes dimensions que nous retrouvons tout au long de ce manuscrit : complexité du vocabulaire, complexité grammaticale et intelligibilité de la parole. Nous avons choisi de comparer deux approches : une approche dite *interprétable*et une approche dite *neuronale*. Concernant la première approche, notre objectif est de démontrer que les ré-

sultats obtenus restent explicables et que de ce fait, des compléments d'informations sur les causes impactant le niveau de compréhensibilité peuvent être identifiées et communiqués aux utilisateurs de nos futurs outils. C'est ce qui fait l'objet du chapitre suivant.

Chapitre 6

Mesure objective du niveau de compréhensibilité : approche interprétable

Sommaire

6.1 Introduction

Nous avons envisagé le calcul de la mesure du niveau de compréhensibilité sous deux angles, il s'agit de deux approches prédictives fondées sur des méthodes d'apprentissage machine : une approche *classique* basée sur un ensemble de paramètres extraits des contenus étudiés. Cette **approche** *classique* est désignée comme ***interprétable***, car nous souhaitons pouvoir interpréter et comprendre comment la mesure est construite. Ce chapitre est dédié à cette première approche. Dans un but de comparaison des performances, une deuxième approche dite *neuronale* reposant sur l'utilisation de représentations issues de réseaux de neurones sera étudiée dans le chapitre suivant.

Dans le chapitre précédent, nous avons identifié plusieurs catégories de paramètres. Avec des données numériques continues fournies en entrée (valeurs de paramètres et valeurs de la vérité terrain issue du corpus ESCAL), nous nous plaçons dans une problématique de **régression** où l'objectif est de prédire a mieux le niveau de compréhensibilité, que ce soit au niveau global (également désignée dans notre étude par le terme difficulté globale) ou au niveau des différentes dimensions que sont la complexité du vocabulaire, la complexité de la grammaire et l'intelligibilité de la parole.

Dans ce chapitre nous nous intéresserons en premier lieu à l'ensemble des étapes préliminaires à réaliser avant de procéder à la construction des modèles : la sélection de paramètres à utiliser et la gestion de la multicolinéarité. Dans un second temps nous expliquerons quelle est la méthode de construction de modèle qui a été choisie avant de présenter les divers modèles nécessaires pour aboutir à un modèle de prédiction objective du niveau de compréhensibilité. Enfin, nous procéderons à l'application aux données du corpus ESCAL.

6.2 Sélection des paramètres pour la prédiction

Certains des paramètres identifiés dans le chapitre précédent peuvent s'avérer inutiles pour réaliser les prédictions. Il nous a donc semblé nécessaire d'en étudier la pertinence. Il existe deux façons possibles d'éliminer les paramètres qui n'apportent aucune information pour les prédictions :

— réaliser une réduction de dimensions,
— réaliser une sélection de paramètres.

La réduction de dimension consiste en « la production d'un nouvel ensemble de paramètres à partir des paramètres d'origine » [Cunningham, 2008] *via* une combinaison et une projection de ces paramètres d'origine vers une espace de dimension plus petite. À l'issue d'une réduction de dimension, nous obtenons un sous-ensemble de paramètres qui ne peuvent pas être expliqués de manière qualitative. Dans le cas de l'analyse en composante principale (ou ACP) [Ringnér, 2008], qui est une des méthodes classiques de réduction de dimension, chaque paramètre obtenu est une combinaison linéaire des paramètres d'origine. Comme, à terme, nous voulons être capable d'interpréter les résultats des

modèles en fonction des paramètres conservés et caractériser les prédictions obtenues nous nous sommes orientée vers des méthodes de sélection de paramètres qui permettent de conserver la nature des paramètres sélectionnés.

Parmi ces méthodes, nous trouvons celles qui peuvent être réalisées de manière supervisée et celles qui peuvent être réalisées de manière non supervisée. Dans le premier cas, nous avons une vérité terrain et savons ce que nous souhaitons obtenir, dans le second cas, ce que nous cherchons à obtenir est inconnu. La constitution du corpus ESCAL nous permet d'avoir une vérité terrain et de nous situer dans le cas d'une **sélection des paramètres supervisée**. Cela nécessite d'enlever les variables non pertinentes, ce qui peut être fait soit en se basant sur les méthodes de type *filter* ou *wrapper* [Chandrashekar and Sahin, 2014].

Les approches *filter* sont des méthodes d'élimination des paramètres indépendantes du modèle auquel nous voulons aboutir et vont servir de pré-traitement. Elles permettent d'évaluer les relations entre les variables d'entrées et les variables cibles à l'aide d'un score, basé sur une mesure statistique, qui va servir de critère pour filtrer les variables à garder. Parmi les méthodes *filter*, nous pouvons citer une des plus traditionnelles qui est la *Correlation-based feature selection* ou *CFS* [Hall, 1999].

Les approches *wrapper* sont dépendantes du type de modèle que nous voulons obtenir. Elles permettent de trouver la combinaison optimale de paramètres en testant plusieurs modèles qui sont créés à partir de différents sous-ensembles de paramètres. Les divers modèles sont ensuite évalués à partir d'une métrique qui servira à déterminer quel est le sous-ensemble de paramètres qui maximise les performances du modèle. Deux approches classiques de *wrapper* sont le *Stepwise Regression* [Efroymson, 1960] (ou régression pas-à-pas) et le *Recursive Feature Elimination* ou RFE [Guyon et al., 2002]. Ces deux approches vont permettre d'aboutir à une sélection des variables explicatives en suivant une **procédure automatique**.

Les méthodes de type *filter* considèrent chaque variable explicative séparément en les comparant à la variable expliquée, mais cela amène à négliger les potentielles interactions qui existent entre toutes les variables X_i. En effet, il est possible que certaines variables explicatives soient fortement **corrélées** à d'autres ou encore que des variables explicatives soient des combinaisons linéaires d'autres variables explicatives, dans ce cas-là, nous sommes face à une problématique de **multicolinéarité** où la relation entre les variables explicatives est décrite par une relation du type :

$$X_j = \sum_i \lambda_i X_i + \beta \tag{6.1}$$

avec λ_i et β des constantes.

Les méthodes *wrapper* sont également susceptibles d'aboutir à des modèles pour lesquels les paramètres conservés sont fortement corrélés entre eux, voire multicolinéaires.

Dans l'introduction de cette partie, nous avions présenté comme un de nos objectifs de produire un modèle pour lequel nous pourrions décrire, expliquer et expliciter les relations entre les paramètres gardés dans le modèle et les prédictions qui sont réalisées. Quelques modèles ont été explorés tels que les forêts d'arbres de décision (*Random Forest*, RF) [Goel et al., 2017], les machines à vecteurs de supports (*Support Vector Machine*, SVM) [Hearst et al., 1998b] et des régressions linéaires multiples [Aiken et al., 2003]. Les RF et SVM ne permettent cependant pas d'aboutir à un modèle pour lequel il est possible d'expliquer et quantifier le rôle des paramètres dans la prédiction. La **régression linéaire multiple** permet de répondre au mieux à notre besoin puisque qu'elle permet de modéliser une **relation linéaire** entre N variables indépendantes et la variable dépendante Y. Cette relation peut être représentée par la relation mathématique suivante :

$$Y = a_1X_1 + a_2X_2 + ... + a_iX_i + ... + a_{n-1}X_{n-1} + a_nX_n + b \tag{6.2}$$

avec :
- a_i : le $i^{ème}$ coefficient de régression,
- X_i : la $i^{ème}$ variable indépendante.

Dans la suite, nous utiliserons la méthode des moindres carrés pour définir la valeur des coefficients de régression. Celle-ci consiste à minimiser la quantité :

$$S = \sum_i e_i^2 \tag{6.3}$$

où :

$$e_i = y_i - \hat{y}_i \tag{6.4}$$

avec :
- y_i : la $i^{ème}$ valeur observée,
- $\hat{y}_i$: la $i^{ème}$ prédiction,
- e_i : écart entre l'observation et la prédiction.

L'équation qui sera retenue pour expliciter la relation entre variable dépendante et variables indépendantes sera celle qui aura permis de minimiser S.

Faire appel à l'approche *wrapper* nous a semblé plus pertinente puisqu'elle permet d'obtenir à un modèle de régression linéaire tout en réalisant la sélection des paramètres. Mais, à terme, nous devons pouvoir comprendre quels sont les paramètres qui ont été conservés dans le modèle et pour quelles raisons ils l'ont été. Les méthodes *Stepwise Regression* et RFE ont en commun de toutes les deux réaliser une sélection « pas-à-pas » des paramètres : la première va conserver les paramètres en fonction de leur *pvalue*, qui permet de quantifier la significativité statistique d'un choix (ce qui veut dire que les paramètres conservés à l'issue de cet algorithme sont les plus « adéquats » pour prédire la variable dépendante [Efroymson, 1960], et la seconde va classer les paramètres en fonction de leur apport dans le modèle et éliminer ceux classés en dernier. Le RFE

nécessite de renseigner en entrée le nombre de variables explicatives à conserver, et comme nous n'avons pas cette information au préalable, des traitements supplémentaires pour déterminer le nombre de paramètres idéal à garder devraient être réalisés, chose qui n'est pas nécessaire pour le *Stepwise Regression*. De plus, RFE est déjà un algorithme qualifié de « gourmand » parce qu'il teste tous les sous-ensembles possibles de paramètres, nous avons décidé d'utiliser le *Stepwise Regression* pour **la sélection des paramètres et l'obtention du modèle**.

Nous avons opté pour la réalisation de modèles de régression linéaire multiple, mais la multicolinéarité peut être problématique dans le cas où nous souhaitons estimer et interpréter un modèle linéaire car s'il y a colinéarité, nous ne pouvons pas nous assurer de la fiabilité des coefficients de chaque variable explicative. Certains paramètres gardés, s'il y a colinéarité, sont susceptibles de mesurer le même phénomène[1]. De plus, la valeur des *pvalues* des variables colinéaires n'est pas fiable alors que c'est important pour pouvoir appliquer le *Stepwise Regresssion*. Les **problématiques de multicolinéarité** doivent donc être traitées **en amont avant** d'appliquer l'algorithme du *Stepwise Regression*.

Les différentes étapes qui permettent d'aboutir à la construction du modèle de régression linéaire multiple, sont détaillées dans la figure 6.1.

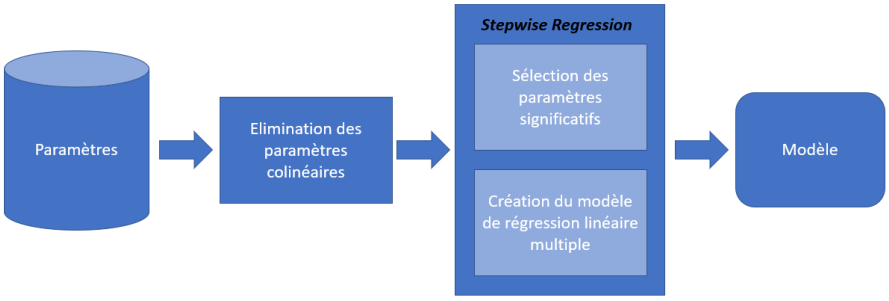

FIGURE 6.1 – Étapes pour la construction d'un modèle de prédiction

6.3 Gestion de la multicolinéarité

Comme expliqué dans la section précédente, une multicolinéarité entre les variables quand nous souhaitons réaliser une régression linéaire va amener de l'incertitude quant à la fiabilité des coefficients de variables explicatives et leur *pvalue*. Éliminer les problèmes de multicolinéarité va permettre de construire des modèles de régression linéaire fiables et interprétables à partir de l'algorithme de *Stepwise Regression*.

1. http://larmarange.github.io/analyse-R/multicolinearite.html

Pour identifier les variables colinéaires, nous utilisons la méthode présentée dans [Welch et al., 1994] qui consiste à construire une matrice de corrélations des variables, à partir de la mesure de corrélation de Pearson (le coefficient de corrélation est traditionnellement noté r) entre chaque paire de variable indépendante. C'est une des méthodes les plus anciennes pour détecter une relation linéaire entre deux variables X, Y continues[2].

Plusieurs forces de corrélations sont identifiées, en fonction de la valeur absolue de r :

— Entre 0 et 0,2 : X et Y sont très faiblement corrélées,
— Entre 0,2 et 0,4 : X et Y sont faiblement corrélées,
— Entre 0,4 et 0,6 : X et Y sont moyennement corrélés,
— Entre 0,6 et 0,8 : X et Y sont fortement corrélés,
— Supérieur à 0,8 : X et Y sont très fortement corrélées.

Si r excède un seuil α, nous considérons qu'un cas de colinéarité est présent. Le seuil α varie selon les études : certains le fixent à 0,5 [Donath et al., 2012]. Dans [Dormann et al., 2013] il est indiqué que si r excède le seuil de $0,7$ alors la colinéarité entre les variables est élevée, la multicolinéarité peut alors devenir problématique. Cependant, une valeur plus commune est 0,8 : c'est la valeur de seuil utilisée dans la suite[Berry et al., 1985]. Ainsi, si deux variables X et Y sont corrélées avec un $r \geq 0,8$ alors nous éliminons une des deux variables pour diminuer la multicolinéarité parmi les variables [Kumari, 2008].

Une fois la problématique de multicolinéarité considérée et gérée, il est possible de réaliser l'étape suivante qui est la construction d'un sous-ensemble de paramètres pertinents.

6.4 Sélection des paramètres et construction du modèle avec le *Stepwise Regression*

Pour trouver les paramètres pertinents, nous avons fait le choix d'appliquer l'algorithme de sélection des paramètres « Stepwise » [Hocking, 1976].

Les méthodes de sélection pas-à-pas permettent de construire un modèle de régression tout en choisissant en parallèle un sous-ensemble de variables explicatives parmi un ensemble de variables candidates. La sélection des paramètres peut se faire soit :

— par élimination de variable : il s'agit alors d'une méthode descendante,
— par ajout de variable : c'est une méthode ascendante.

La méthode *Stepwise*, développée dans cette section, est une combinaison des méthodes descendante et ascendante : une fois qu'une variable est ajoutée, comme l'ajout d'une nouvelle variable significative peut influer sur la significativité des autres variables, nous vérifions si l'ensemble des variables déjà incluses sont toujours significatives.

2. http://www.biostat.ulg.ac.be/pages/Site_r/corr_pearson.html

L'ajout ou le rejet d'une variable explicative dans le modèle repose sur sa significativité. Pour la déterminer, nous pouvons nous référer soit à la *pvalue* de la variable explicative, soit au coefficient de détermination R^2 du modèle de régression. La *pvalue* quantifie la significativité statistique d'un choix en estimant la probabilité qu'une variable indépendante soumise à un test spécifique doive son coefficient au hasard. Si la *pvalue* dépasse un seuil donné (en général ce seuil est fixé à 0,05), alors, il est possible que le coefficient obtenu pour la variable étudiée soit dû au hasard, et nous considérerons qu'elle n'est pas significative. Le coefficient de détermination R^2 (ou coefficient de corrélation multiple) est un estimateur de la qualité d'un modèle de régression qui indique la proportion de variance « expliquée » par le modèle de régression [Nagelkerke et al., 1991]. L'apport d'une variable à un modèle peut être estimée à la manière dont le R^2 va évoluer lors de son ajout ou de sa suppression. Le critère d'élimination le plus usité dans les logiciels statistiques (du type IBM SPSS[3] ou encore STATA[4]) est la *pvalue*, c'est le critère qui a été utilisé ici : une variable sera conservée dans le modèle si sa *pvalue* est inférieure ou égale à 0,05.

L'algorithme de sélection de paramètres *Stepwise Regression* fonctionne ainsi :

1. **Initialisation :** l'algorithme est initialisé avec un ensemble vide de variables explicatives. La première variable incluse dans le modèle de régression est celle qui propose le meilleur coefficient de détermination R^2,
2. **Itération :** à chaque étape la séquence suivante est réalisée :
   (a) une variable est ajoutée dans le cas où elle est significative,
   (b) un contrôle est réalisé pour vérifier si toutes les variables déjà incluses dans le modèle sont toujours significatives.
3. **Arrêt de l'algorithme :** le traitement s'arrête si un critère d'arrêt pré-défini est atteint ou s'il ne reste plus de variables à tester.

Le critère d'arrêt peut être de différents types [Legrand, 2004], soit :
— nous prédéfinissons un nombre de variables explicatives à sélectionner,
— nous fixons un nombre d'itérations maximal à réaliser,
— la suppression ou l'ajout de variable ne permet pas d'obtenir de modèle plus performant,
— le sous-ensemble est évalué comme le meilleur possible.

Dans notre cas, aucun critère d'arrêt n'est défini : l'algorithme s'arrête quand il n'y a plus de variables à tester.

Pour pouvoir appliquer l'algorithme de sélection de paramètres, nous devons nous baser sur un modèle de régression à partir duquel nous allons évaluer la significativité des paramètres. Notre objectif *in fine* est d'aboutir à un modèle de régression linéaire multiple, c'est donc à partir de ce modèle que nous évaluons

3. https ://www.ibm.com/fr-fr/analytics/spss-statistics-software
4. https ://www.stata.com/

la significativité des paramètres. À la fin de l'algorithme *Stepwise Regression*, nous obtenons en sortie **un modèle de régression linéaire multiple avec l'ensemble des variables explicatives qui ont été conservées**. Pour s'assurer de la pertinence du modèle, il est ensuite nécessaire de s'assurer de sa robustesse quand il est confronté à de nouvelles données. Pour se faire, nous utilisons une méthode de validation croisée.

6.4.1 Validation croisée

6.4.1.1 Principe

Lorsque nous construisons des modèles, il faut s'assurer que les modèles ne réalisent pas de sur-apprentissage et qu'ils sont capables de faire des prédictions sur de nouvelles données. Plusieurs manières de valider les modèles existent. Il est possible de séparer les données en deux parties : le modèle est appris sur les données d'entraînement et les performances sont évaluées sur les données de test à l'aide de métriques d'évaluation.

Cependant, lorsque nous disposons de peu de données il est intéressant de réaliser une **validation croisée** [Moore, 2001], qui permettra d'exploiter toutes les données disponibles. Les données sont découpées en k parties équilibrées (*folds* en anglais), et à chaque itération, une de ces parties est utilisée comme jeu de validation. Les autres parties sont utilisées comme données pour l'entraînement. Avec cette approche, l'intégralité des données est évaluée exactement une fois, en k itérations.

Si l'ensemble des données est vraiment restreint, nous procédons à une validation croisée en *Leave-One-Out*. À chaque étape, une seule donnée est utilisée pour la validation et le reste est utilisé pour l'apprentissage.

À l'issue de chaque étape de la validation croisée, les prédictions sont récupérées et stockées. Les métriques d'évaluation sont ensuite moyennées à partir de l'ensemble des prédictions.

6.4.1.2 Métriques d'évaluation

Pour évaluer la performance des modèles avec la validation croisée, nous utilisons deux métriques :

— le coefficient de corrélation de Pearson : r Présentée dans la section 6.3, dans le cas de l'évaluation de la qualité d'un modèle de prédiction, la corrélation de Pearson est calculée entre les valeurs prédites et les valeurs réelles. Notre intérêt est de choisir les modèles qui permettent d'obtenir le r le plus élevé possible. En effet, nous considérons que plus la valeur du r est élevée plus le modèle est pertinent.

— l'erreur quadatrique moyenne : $RMSE$ (*Root Mean Square Error*) Elle permet d'évaluer la qualité de prédiction d'un modèle sur des données inconnues, en donnant une indication sur l'amplitude des écarts entre les

valeurs prédites et les valeurs de référence :

$$RMSE = \sqrt{\frac{S}{n}} \tag{6.5}$$

où n désigne le nombre d'éléments, et S est présenté dans l'équation 6.3. L'objectif est de minimiser la valeur du $RMSE$: plus elle se rapproche de zéro, meilleure est la prédiction du modèle.

6.4.1.3 Règle de base pour évaluer la pertinence d'un modèle de régression

En plus des métriques r et $RMSE$ mesurées en *Leave-One-Out* pour évaluer les performances du modèle, nous considérons également la règle de base (ou *rule of thumb*) appelée « *one in ten* » [Harrell Jr et al., 1984] pour s'assurer que les modèles issus du *Stepwise Regression* sont pertinents. Cette règle dit que lorsque nous construisons un modèle de régression, il faut au maximum une variable indépendante pour 10 observations. Au-delà, il y a un risque que le modèle fasse du sur-apprentissage : en dehors des données d'entraînement, le modèle réalisera de mauvaises prédictions. En sachant que nous avons au total 55 observations issues de notre corpus ESCAL, un modèle issu du *Stepwise Regression* devrait compter cinq variables indépendantes pour être conservé. Ainsin dans le cas où le nombre de variables gardées serait supérieures à cinq, il faudrait vérifier si le r en *Leave-One-Out* et le r obtenu sur les données d'entraînement ont un écart conséquent. Par exemple, si sur les données d'entraînement le r vaut 0,8 et chute à 0,2 en *Leave-One-Out*, nous sommes probablement confrontée à un cas de sur-apprentissage.

6.5 Modèles de prédiction du niveau de compréhensibilité en fonction des modalités

Nous considérons la mesure de la compréhensibilité selon quatre points de vue en lien avec les différentes modalités. D'un côté nous souhaitons produire une mesure globale de la compréhensibilité, de l'autre nous voulons faire un focus sur chacune des dimensions identifiées : la complexité du vocabulaire, de la grammaire et l'intelligibilité de la parole. Pour chacune de ces dimensions, nous avons mené une étude sur la prédiction en considérant une modalité (texte, audio, vidéo/image) ou une combinaison de ces modalités.

Nous avons quatre dimensions à évaluer et trois modalités à notre disposition. Il est possible d'aborder la prédiction de chacune des dimensions soit en exploitant :

— les paramètres liés à une seule dimension et calculés à partir d'une seule modalité (par exemple, pour prédire la complexité du vocabulaire, nous pouvons nous baser uniquement sur des paramètres issus du texte),

— les paramètres liés à une seule dimension et calculés à partir de plusieurs modalités (par exemple, pour prédire l'intelligibilité de la parole, nous

pouvons exploiter des paramètres issus de la modalité audio, mais aussi de la modalité vidéo),
— les paramètres issus de plusieurs modalités même s'ils ne sont *a priori* pas liés à la dimension que nous cherchons à évaluer (par exemple, pour prédire la complexité grammaticale, nous pourrions faire appel à des paramètres issus du texte, mais aussi de la modalité audio),
— tous les paramètres issus de toutes les modalités.

La figure 6.2 permet de résumer les différentes étapes qui vont aboutir à la génération d'un modèle en fonction de la dimension à prédire.

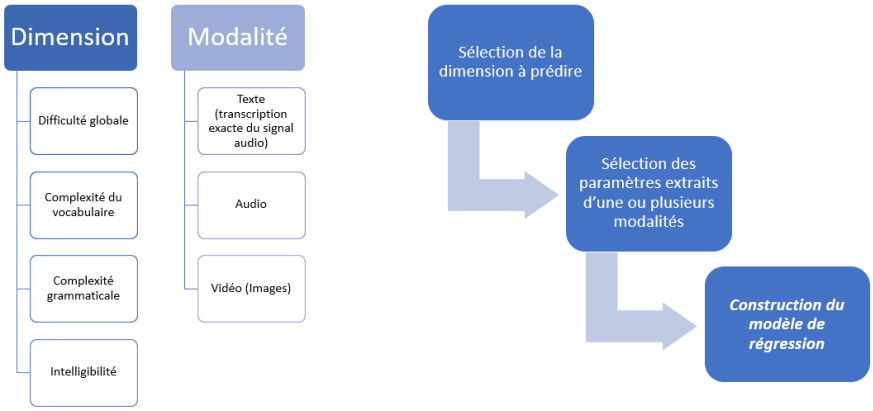

FIGURE 6.2 – Résumé de l'approche de prédiction. La construction du modèle de régression reprend les étapes présentées dans la figure 6.1.

6.6 Application aux données du corpus ESCAL

Après avoir décrit les principes de l'approche proposée, nous avons réalisé une étude des résultats sur le corpus ESCAL décrit dans le chapitre 3.

Pour rappel, le corpus comprend au total 55 documents qui correspondent chacun à un extrait de film, ces extraits ont été choisis pour se focaliser sur des situations d'interaction. Les 55 extraits ont été présentés sous différentes modalités ou combinaisons de modalités à des enseignants de FLE pour être évalués en termes de difficulté globale, de complexité du vocabulaire, de complexité grammaticale et d'intelligibilité de la parole. Cela nous a permis d'obtenir une vérité terrain qui va pouvoir être comparée aux prédictions des différents modèles construits.

Dans cette section seront présentés les meilleurs modèles qui ont été obtenus pour chacune des dimensions considérées. Au vu de la taille modeste du corpus,

la méthode choisie pour l'évaluation des modèles est le *Leave-One-Out*, présenté dans la section 6.4.1.1.

6.6.1 Prédiction de la complexité du vocabulaire

Pour prédire le niveau complexité du vocabulaire, nous avons dans un premier temps considéré les paramètres issus de la modalité texte qui sont uniquement en lien avec cette dimension.

Dans le cadre de l'évaluation réalisée par les enseignants de langue pour la constitution du corpus ESCAL, il est possible que des facteurs en lien avec la complexité grammaticale aient été implicitement pris en compte par les annotateurs pour prédire la complexité du vocabulaire. L'évaluation du vocabulaire a pu être affectée. Aussi, nous avons comparé les résultats en considérant plus largement les paramètres issus de la modalité texte qui sont liés à la fois à la complexité du vocabulaire et à la complexité grammaticale.

Les deux cas permettent d'aboutir au même modèle de régression linéaire multiple dans lequel seuls des paramètres liés à la complexité du vocabulaire sont conservés. Nous pouvons en déduire, que l'aspect grammatical n'a pas influencé les annotateurs lors l'évaluation du vocabulaire. L'évaluation avec le *Leave-One-Out* nous donne : $r = 0,64$ et $RMSE = 13,31$. La valeur de r met en avant une corrélation forte entre les valeurs prédites et les valeurs réelles, cette relation linéaire est mise en avant dans la figure 6.3. Sur cette figure, comme sur les suivantes, l'axe des abscisses correspond aux prédictions et l'axe des ordonnées aux valeurs réelles issues de la vérité terrain. Chaque point représente donc un des 55 extraits du corpus ESCAL.

6.6.1.1 Meilleurs paramètres

Sur les 21 paramètres correspondant à cette catégorie, 4 ont été retenus pour construire la régression linéaire multiple et sont présentés dans la table 6.1. Ils y sont classés par ordre décroissant, du plus influent au moins influent sur la prédiction de la difficulté du vocabulaire. L'influence est estimée à partir des coefficients de régression standardisés[5].

TABLE 6.1 – Paramètres pour la prédiction de la complexité du vocabulaire

Paramètres retenus	Facteurs représentés
RTTR	Diversité lexicale
$\text{Freq}_{\text{Lemme}_{\leq 19}}$	Fréquence lexicale
Perc_1	Fréquence lexicale
$\text{L}_{\text{Syllabes}}$	Longueur des mots

Le paramètre ayant le plus d'influence sur la complexité du vocabulaire est l'index de Guiraud (RTTR), qui permet de mesurer la diversité lexicale. Les

5. http://w3.uohpsy2.univ-tlse2.fr/UOHPsy2/index.php?option=com_content&task=view&id=186&Itemid=30&limit=1&limitstart=4

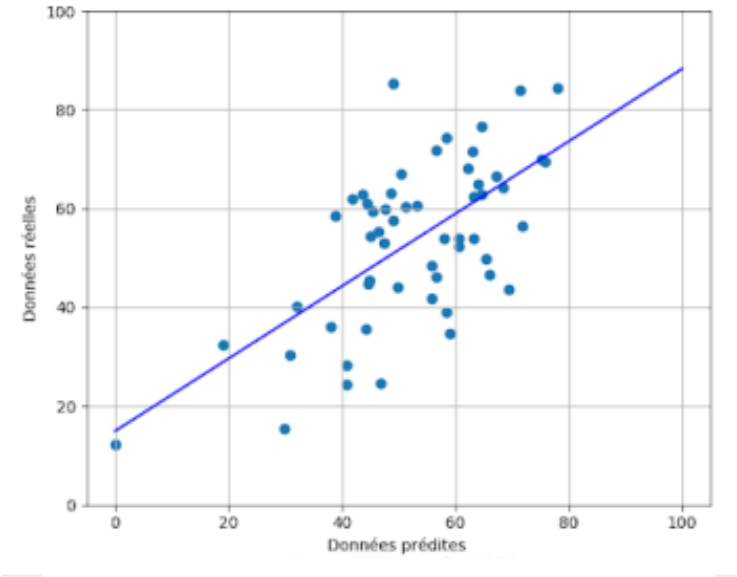

FIGURE 6.3 – Nuage de points mettant en lien les prédictions de complexité du vocabulaire et les valeurs réelles (Modalité texte, Paramètres en lien avec le vocabulaire)

deux paramètres suivant sont tous les deux en lien avec la fréquence lexicale :

- le pourcentage de mots dont le lemme a une fréquence inférieure à 19 dans la base de données Lexique. Cela signifie que le pourcentage de mots peu fréquents influe sur la complexité du vocabulaire,
- le premier percentile de la fréquence des lemmes. Il correspond à la valeur de fréquence pour laquelle 1% des mots ont un lemme de fréquence inférieure à cette valeur. Plus cette valeur est faible, plus les mots présents sont susceptibles d'être rares. Nous voyons ainsi l'impact de la rareté du vocabulaire sur la complexité du vocabulaire.

Le dernier paramètre qui a été gardé correspond au nombre moyen de syllabes par mot, ce qui veut dire que la longueur des mots utilisés dans un extrait joue un rôle sur la complexité lexicale, même s'il a moins d'influence que la diversité et la fréquence lexicale.

6.6.2 Prédiction de la complexité de la grammaire

De la même façon, nous avons d'abord considéré les paramètres issus de la modalité texte en lien avec la complexité grammaticale, avant d'élargir à l'ensemble des paramètres issus de la modalité texte (vocabulaire et grammaire), pour couvrir les cas où les annotateurs auraient été influencés par la dimension lexicale lors de leur évaluation de la grammaire.

Avec seulement les paramètres liés à la complexité grammaticale, nous obtenons, en *Leave-One-Out*, $r = 0,17$ et $RMSE = 11,68$. Quand les paramètres liés à la complexité du vocabulaire sont pris en compte, nous atteignons alors $r = 0,53$ et $RMSE = 9,38$. Nous passons ainsi d'une corrélation faible à une corrélation moyenne entre les valeurs prédites et les valeurs réelles.

Cette différence dans la qualité des prédictions se reflète lorsque nous comparons les nuages des points obtenus avec les modèles des deux régressions (figures 6.5 et 6.4). En effet, bien que les deux prédictions soient toujours comprises dans le même intervalle de valeurs (entre 20 et 60), nous remarquons que la droite de régression a une pente plus conséquente dans le second graphique ce qui démontre une meilleure qualité du modèle de régression. Sur la figure 6.5, les points sont plus concentrés et suivent la droite de régression tandis que pour la figure 6.4, nous remarquons une forte dispersion autour de la droite de régression, ce qui reflète une mauvaise qualité de la régression.

Modalité texte - Nuage de points mettant en lien les prédictions de complexité grammaticale et les valeurs réelles

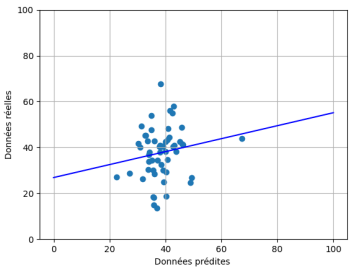

FIGURE 6.4 – Paramètres en lien avec la grammaire

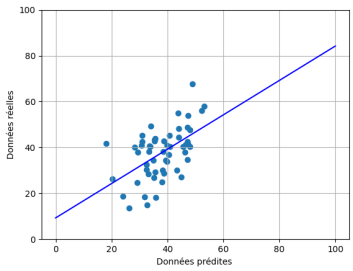

FIGURE 6.5 – Paramètres en lien avec la grammaire et le vocabulaire

6.6.2.1 Meilleurs paramètres

La table 6.2 répertorie les 2 paramètres qui ont été retenus, sur les 11 calculés. Le paramètre qui a le plus d'influence est une nouvelle fois l'index de Guiraud. Ce paramètre étant en lien avec la diversité lexicale, et étant le paramètre ayant le plus d'influence sur la complexité du vocabulaire, cela signifie que lors de l'expérience, les annotateurs ont été fortement influencés par l'aspect lexical lorsqu'ils ont évalué la complexité grammaticale. Le second paramètre qui a été

gardé correspond au pourcentage de verbes qui ont été conjugués avec un temps verbal difficile, c'est-à-dire tous les temps verbaux hors présent de l'indicatif, impératif présent et passé composé.

TABLE 6.2 – Paramètres pour la prédiction de la complexité grammaticale

Paramètres retenus	Facteurs représentés
RTTR	Diversité lexicale
Difficulté_Temps_Verbaux	Complexité morphologique (complexité verbale)

Ainsi, pour évaluer la complexité de la grammaire d'un extrait, deux aspects ont été pris en compte par les annotateurs : la diversité lexicale et la complexité des temps verbaux utilisés.

6.6.3 Prédiction de l'intelligibilité de la parole

Dans l'état de l'art (voir section 2.4), il a été vu que l'intelligibilité dépend de ce qui est entendu, mais des informations non verbales (gestuelle, expressions faciales) peuvent aussi influencer la compréhension d'une situation. Comme les documents avec lesquels nous travaillons sont des documents audiovisuels, nous nous intéressons aux paramètres en lien avec l'intelligibilité qui sont issus des modalités audio **et** vidéo.

Notre modèle obtient, en *Leave-One-Out*, $r = 0,48$ et $RMSE = 17,61$. Le r met en avant qu'il existe une corrélation moyenne entre les valeurs prédites et les valeurs réelles. La figure 6.6 reflète la relation linéaire existante entre les valeurs prédites et les valeurs réelles, cependant, la dispersion des points autour de la droite de régression montre que le système manque de précision.

6.6.3.1 Meilleurs paramètres

La table 6.3 liste l'ensemble des 6 paramètres qui ont été conservés pour la prédiction de l'intelligibilité, sur les 15 qui en sont extraits. Nous constatons que seuls des paramètres liés à la modalité audio ont été conservés. Bien que deux annotateurs aient évoqué que l'absence de redondance entre l'audio et l'image pénalisait l'intelligibilité de certains documents, cela n'est apparemment pas rentré en jeu dans la prédiction du niveau d'intelligibilité de la parole.

Si nous regroupons les paramètres gardés par catégorie, nous remarquons que les paramètres qui ont joué un rôle sur l'intelligibilité sont les suivants :

1. la qualité de prononciation,
2. le débit de parole,
3. le bruit,
4. le taux de parole détectée.

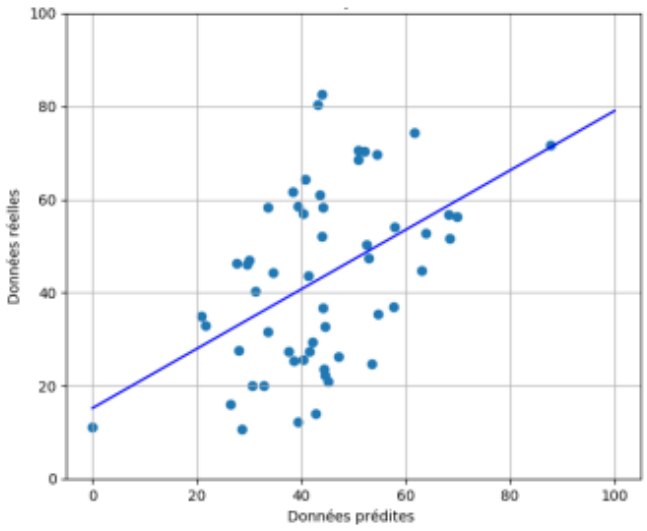

FIGURE 6.6 – Nuage de points mettant en lien les prédictions d'intelligibilité et les valeurs réelles (Modalité Audio et Texte - Paramètres liés à l'intelligibilité)

TABLE 6.3 – Paramètres pour la prédiction de l'intelligibilité

Paramètres retenus	Facteurs représentés
D_Lev_{Med}	Qualité de prononciation
$Débit_{INA\_ST\_Moy}$	Débit de parole
$SNR_{RMS\_ST\_EIQ}$	Bruit
$Durée\_RTC_{Moy}$	Taux de parole détectée
D_Lev_{Max}	Qualité de prononciation
$SNR_{NIST\_ST\_Min}$	Bruit

La règle de base présentée dans 6.4.1.3 n'est pas respectée : en effet, six paramètres ont été conservés à l'issue du *Stepwise Regression*. Il faudrait normalement que cinq paramètres soient conservés, nous devons donc vérifier la validité du modèle en comparant ses performances en *Leave-One-Out* et sur les données d'entraînement. Lorsqu'il est calculé à partir des données d'entraînement, le r atteint une valeur de 0,77 contre 0,48 en *Leave-One-Out*. Nous passons d'une corrélation forte à une corrélation moyenne entre les données prédites et les données réelles, il est donc possible que le modèle fasse du surapprentissage. La différence entre le nombre de variables indépendantes conservées pour ce

modèle et le nombre idéal de paramètres à conserver d'après la règle de *one in ten* est très faible mais il faudra toutefois considérer pour la suite que le modèle ne sera pas forcément robuste.

6.6.4 Prédiction du niveau de compréhensibilité global

Pour prédire le niveau de compréhensibilité globale d'un document (ou difficulté globale), plusieurs stratégies de fusion sont possibles :

Fusion tardive : en combinant les prédictions directement issues des modèles obtenus pour les différentes dimensions ;

Fusion intermédiaire : en associant les paramètres qui ont été retenus par la sélection pas-à-pas, pour chacun des modèles ;

Fusion précoce : en combinant l'ensemble des paramètres qui ont été extraits et qui sont liés au vocabulaire, à la grammaire, à l'intelligibilité et à la difficulté globale.

Nous avons mené une étude pour comparer les résultats de prédiction obtenus selon l'une de ces trois stratégies de fusion et déterminer quelle est celle qui est la plus adaptée pour prédire la difficulté globale d'un document audiovisuel.

6.6.4.1 Fusion tardive

À partir des paramètres liés à chacune des dimensions et par le biais de l'algorithme *Stepwise*, nous obtenons quatre modèles de régression linéaire multiple. Ces quatre modèles permettent d'obtenir des prédictions pour chaque dimension et pour chacun des 55 extraits du corpus ESCAL. Ces décisions permettent de réaliser la **fusion tardive** : pour chaque document, les prédictions réalisées en *Leave-One-Out* sur chacune des dimensions sont exploitées pour déterminer sa difficulté globale (voir figure 6.7). Elle peut être déterminée de plusieurs façons, voici celles que nous avons retenues :

— en moyennant les prédictions,
— en calculant la médiane des prédictions,
— en gardant la prédiction maximum,
— en gardant la prédiction minimum.

Avec ces quatre mode de calcul nous obtenons en *Leave-One-Out*, les résultats présentés dans la table 6.4. Les performances sont maximisées par la méthode de fusion à partir du calcul de la moyenne des prédictions issues des modèles de régression linéaire multiple. Les prédictions de complexité du vocabulaire, de complexité de la grammaire, d'intelligibilité de la parole et de difficulté globale jouent un rôle à parts égales pour déterminer la difficulté globale. Cette méthode aboutit, en *Leave-One-Out*, à un coefficient de corrélation r égale à 0,4 et un RMSE égale à 18,26.

La valeur du coefficient r de Pearson reflète l'existence d'une relation linéaire faible entre la moyenne des prédictions issues des différents modèles de régression et la variable expliquée. Le $RMSE = 18,26$ signifie que la variance atteint environ 36% de la moyenne des observations qui est égale à 50,98 et qui

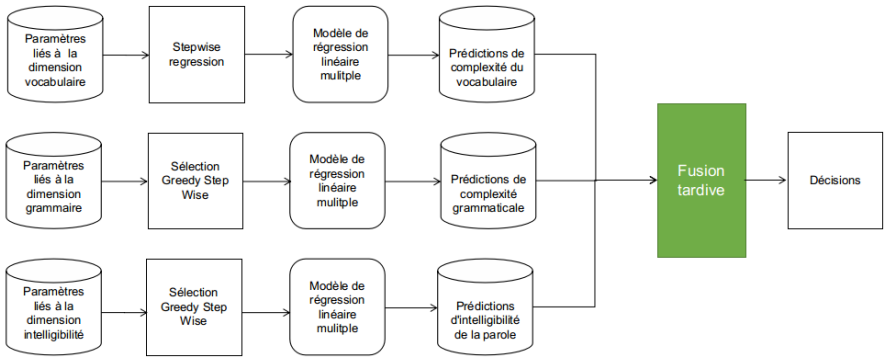

FIGURE 6.7 – Schéma décrivant l'approche de prédiction de la difficulté globale pour une fusion tardive

TABLE 6.4 – Performance des méthodes de fusion tardive en *Leave-One-Out* (*$pvalue \leq 0,05$, **$pvalue \leq 0,01$)

Méthode de fusion	R	RMSE
Moyenne	0,4**	18,26
Médiane	0,29*	19,55
Minimum	0,35**	24,59
Maximum	0,34*	19,45

correspond à la moyenne des valeurs de vérité terrain de la dimension difficulté globale, cette variance moyenne nous indique que le modèle n'est pas satisfaisant en terme de précision. Si nous nous intéressons à la figure 6.8, qui représente le nuage de points mettant en lien les valeurs prédites et les valeurs réelles, la dispersion des points autour de la droite de régression supporte l'argument du manque de précision de notre modèle.

6.6.4.2 Fusion intermédiaire

Pendant les phases de construction des modèles de prédiction pour les quatre dimensions étudiées, un ensemble de paramètres pertinents a été retenu pour chacune des dimensions étudiées. L'approche de fusion dite « intermédiaire » consiste à se servir de ces ensembles de paramètres, mais également des paramètres en lien avec la difficulté globale comme variables candidates pour réaliser la prédiction de la difficulté globale (voir figure 6.9).

Au total, nous mettons en entrée du modèle 22 paramètres candidats dont :
— quatre en lien avec la complexité du vocabulaire (voir section 6.6.1.1),
— deux en lien avec la complexité grammaticale (voir section 6.6.2.1),
— six en lien avec la complexité de l'intelligibilité (voir section 6.6.3.1),

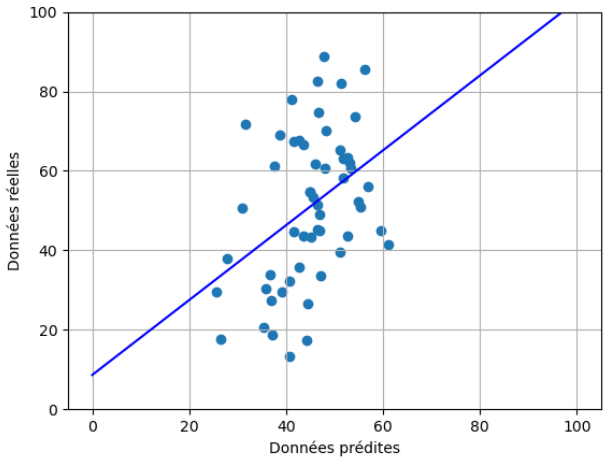

FIGURE 6.8 – Nuage de points mettant en lien les prédictions de difficulté globale et les valeurs réelles en fusion tardive

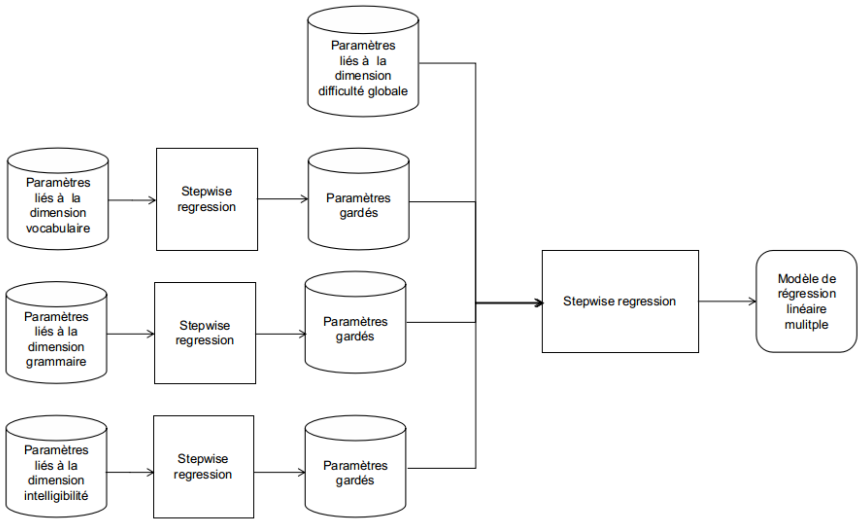

FIGURE 6.9 – Schéma décrivant l'approche de prédiction de la difficulté globale pour une fusion intermédiaire

— et dix paramètres en lien avec la difficulté globale (voir section 5.5).

Cette approche aboutit à un modèle qui en *Leave-One-Out* atteint un r de 0,38 et une $RMSE$ de 17,88. La valeur du coefficient de corrélation r reflète

une corrélation faible entre les prédictions et les valeurs réelles. Le modèle ainsi construit a également des difficultés à réaliser de bonnes prédictions dans le cadre du *Leave-One-Out* : en sachant que la moyenne des observations (vérité terrain) est de $50,98$, le $RMSE = 17,88$ signifie que la variance du modèle atteint entre 35% et 36% de la moyenne des observations. Cette variance moyenne témoigne d'un manque de précision des prédictions du modèle. La figure 6.10 montre que les points ne suivent pas la tendance de la droite de régression : il y a une dispersion des points autour de la droite et ils en sont éloignés. Tout ceci nous permet de conclure que la relation linéaire entre les valeurs prédites et les valeurs réelles existe, mais elle faible et le modèle n'est pas précis.

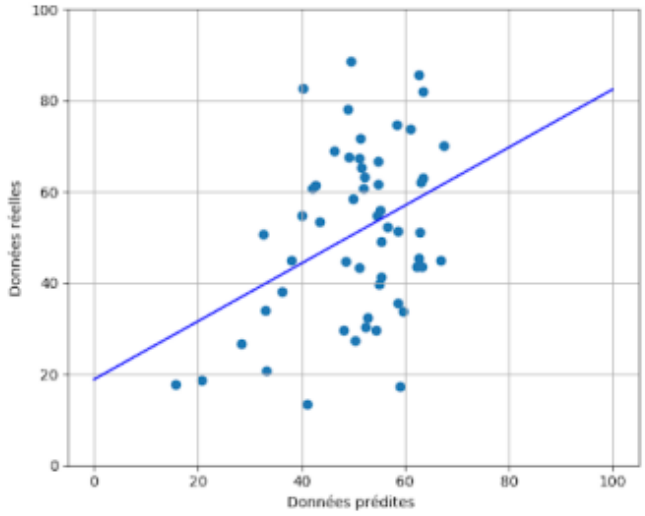

FIGURE 6.10 – Nuage de points mettant en lien les prédictions de difficulté globale et les valeurs réelles en fusion intermédiaire

La table 6.5 montre les paramètres qui ont été conservés pour prédire la difficulté globale. La première chose à noter est que les seuls paramètres qui ont été conservés sont liés à la complexité du vocabulaire : cela met une nouvelle fois en avant que c'est l'aspect lexical qui a la plus forte influence sur la perception de la difficulté globale. Les deux paramètres qui ont été gardés font d'ailleurs partie des trois premiers paramètres les plus influents sur la complexité du vocabulaire. Le premier percentile de la fréquence des lemmes a un lien direct avec la fréquence lexicale : comme expliqué plus tôt, si le premier percentile de la fréquence des lemmes diminue, cela veut dire que le document présentera des

mots de plus en plus rares, ce qui affectera donc la difficulté globale. L'index de Guiraud est un indicateur de la diversité lexicale : l'augmentation du nombre de mots lexicaux présents dans un extrait contribue à modifier la difficulté globale.

TABLE 6.5 – Paramètres inclus dans le modèle de prédiction de la difficulté globale en fusion intermédiaire

Paramètres	Phénomène représenté
Perc_1	Fréquence lexicale
RTTR	Diversité lexicale

6.6.4.3 Fusion précoce

Dans cette approche, l'ensemble des 87 paramètres qui ont été calculés, et qui sont en lien avec toutes les dimensions, sont considérés comme des candidats en tant que variables indépendantes dans la régression linéaire multiple (voir figure 6.11).

Cette approche permet d'obtenir un modèle pour lequel le coefficient de corrélation de Pearson r atteint une valeur de 0,396 et la $RMSE$ est égale à 18,26 en *Leave-One-Out*. Le r met en avant la relation linéaire faible qui existe entre les données prédites et les données réelles, tandis que le $RMSE$ nous indique que la variance représente environ 38% de la moyenne des observations, égale à 50,98. La figure 6.12 montre une droite de régression avec une pente faible, mais aussi une importante dispersion des points autour de la droite de régression, ce qui reflète le manque de précision du modèle.

Toutes les modalités sont entrées en jeu dans la création du modèle, puisque parmi les paramètres conservés à l'issue du *Stepwise Regresssion* nous retrouvons des paramètres issus du texte, de l'audio, mais aussi un paramètre extrait du flux vidéo. Cependant, nous ne pouvons pas considérer ce modèle. En effet, la quantité de paramètres conservés par la méthode de sélection des paramètres est conséquente : pour 55 observations, nous comptons 14 paramètres retenus. Or, d'après la règle de base présentée dans 6.4.1.3, un nombre trop important de paramètres a été conservé après le *Stepwise Regression* : pour qu'il soit pertinent il devrait y avoir cinq paramètres de gardés, nous en comptons quasiment le triple. Le modèle obtenu à l'issue de la fusion précoce est susceptible de faire du sur-apprentissage : il sera trop spécialisé sur les données d'entraînement. Effectivement, nous notons que lorsqu'il est calculé sur l'ensemble des données d'entraînement $r = 0,92$ (ce qui reflète une corrélation linéaire quasiment parfaite), mais, en *Leave-One-Out* le r diminue drastiquement pour chuter à une valeur de 0,42. Nous sommes face à un cas de sur-apprentissage : le modèle de régression n'est pas capable de prédire correctement les données extérieures aux données d'entraînement.

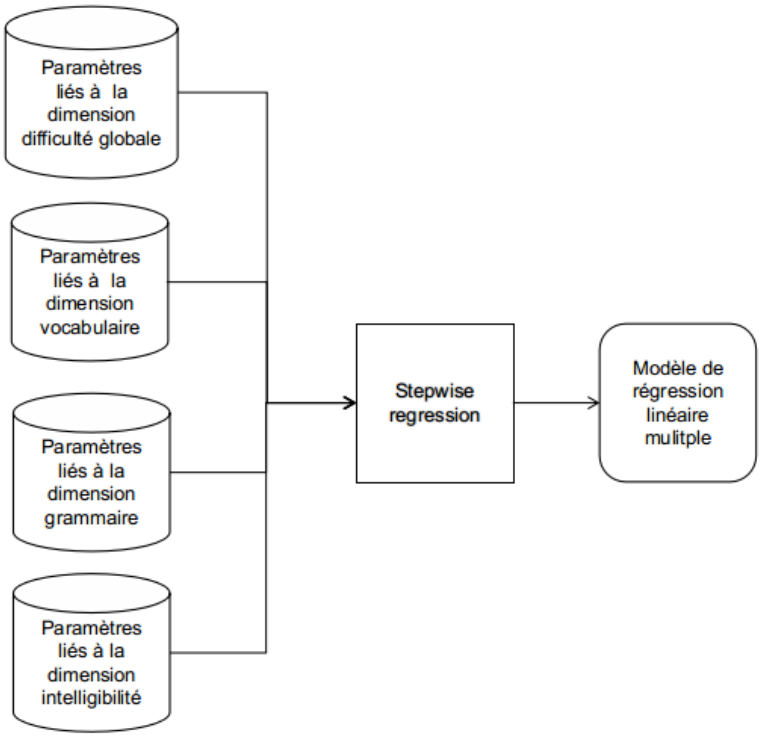

FIGURE 6.11 – Schéma décrivant l'approche de prédiction de la difficulté globale pour la fusion précoce

6.6.4.4 Comparaison des approches de fusion

Comme le modèle issu de la fusion précoce ne répond pas à la règle de base, nous ne le prenons pas en compte et nous comparons uniquement les modèles issus de la fusion tardive et de la fusion intermédiaire. Les résultats sont présentés dans la table 6.6.

TABLE 6.6 – Comparaison des stratégies de fusion pour la prédiction de la difficulté globale

Stratégie	R	RMSE
Fusion tardive	0,4	18,26
Fusion intermédiaire	0,38	17,88

Ces deux stratégies permettent d'obtenir des valeurs relativement proches, que ce soit pour le r ou le $RMSE$: le modèle obtenu en fusion tardive maximise le r et le modèle obtenu en fusion intermédiaire minimise le $RMSE$. Dans un premier temps, nous analysons si les coefficients de corrélation r sont significati-

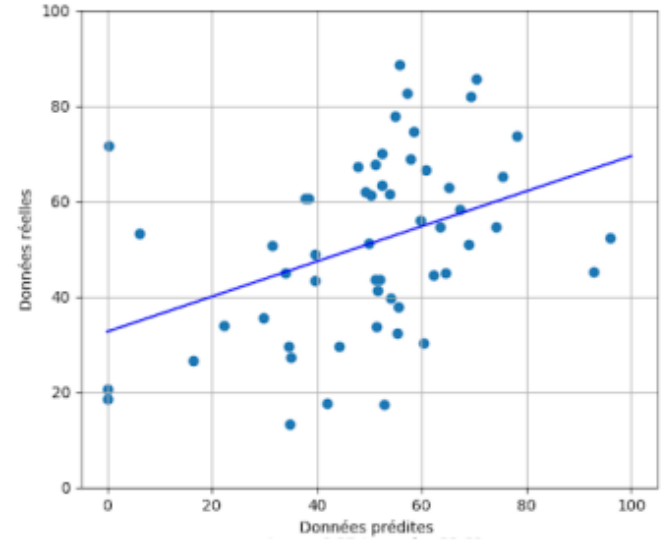

FIGURE 6.12 – Nuage de points mettant en lien les prédictions de difficulté globale et les valeurs réelles en fusion précoce

vement différents. Si c'est bien le cas, cela nous permettra d'identifier si l'un des deux modèles est meilleur que l'autre pour réaliser la prédiction de la difficulté globale. Nous avons comparé les corrélations à l'aide de l'outil COCOR présenté dans [Diedenhofen and Musch, 2015]. Il permet de déterminer que les deux coefficients de corrélation ne sont pas significativement différents. Pour choisir le modèle à conserver pour la suite nous nous sommes donc basée sur la valeur des $RMSE$. Si les deux $RMSE$ obtenus ont un faible écart l'un par rapport à l'autre, pour une exploitation future, nous préférerons cependant le modèle avec le $RMSE$ le plus faible pour ainsi minimiser la variance et favoriser une meilleure précision.

6.7 Bilan

Dans ce chapitre, nous avons modélisé chaque tâche de prédiction et nous avons identifié les paramètres ayant le plus d'influence. A partir des données du corpus ESCAL nous avons construit un modèle pour chaque dimension. Une méthode de validation croisée de type *Leave-One-Out*, adaptée à la taille du corpus ESCAL, a été utilisée pour réaliser ces prédictions et évaluer les performances des différents modèles. La prédiction de la difficulté globale, c'est-à-dire

du niveau de compréhensibilité, est obtenue en appliquant trois stratégies de fusion : fusion précoce effectuées au niveau des paramètres, fusion intermédiaire qui ne tient compte que des paramètres pertinents identifiés dans une phase de sélection et la fusion tardive exploitant les prédictions issues de chaque dimension.

La table 6.7 récapitule les résultats des modèles les plus performants et ce pour chacune des dimensions : vocabulaire, grammaire, intelligibilité et difficulté globale.

Le meilleur modèle en termes de r est celui obtenu pour la complexité du vocabulaire. Nous arrivons à atteindre une corrélation forte entre les valeurs prédites et les valeurs réelles. Les paramètres conservés mettent en avant le fait qu'un vocabulaire diversifié ainsi que la présence des mots rares jouent un rôle prépondérant dans la perception de la difficulté du vocabulaire.

Le modèle avec le $RMSE$ le plus faible correspondant à la meilleure précision, est celui dédié à la prédiction de la complexité grammaticale.

TABLE 6.7 – Performances des modèles selon la méthode *Leave-One-Out*

Dimension considérée	Modalité (Paramètres ou stratégie de fusion)	r	$RMSE$
Vocabulaire	Texte (Vocabulaire)	**0,64**	13,31
Grammaire	Texte (Vocabulaire + Grammaire)	0,53	**9,38**
Intelligibilité	Audio + Vidéo (Intelligibilité)	0,48	17,61
Difficulté globale	Audio + Texte + Vidéo (Fusion intermédiaire)	0,38	17,88

Le modèle obtenu pour l'intelligibilité permet d'aboutir à une corrélation moyenne entre les valeurs de la vérité terrain et les valeurs prédites et donne lieu à une erreur assez élevée.

Le modèle de prédiction de la difficulté globale donnant les résultats les plus intéressants est obtenu par fusion intermédiaire avec $r = 0,38$ et $RMSE = 17,88$.

Pour résumer, les observations qui ont pu être faites pour chacun des modèles construits, permettent de faire plusieurs constats :

— certains modèles sont sujet au surapprentissage, notamment dans le cas de la prédiction de l'intelligibilité (cf. section 6.6.3.1. Dans des études ultérieures, il faudra s'assurer de la capacité de ces modèles à traiter de nouveaux documents, en dehors du corpus ESCAL, et à effectuer des prédictions efficaces.

— les résultats obtenus pour la prédiction de l'intelligibilité et la difficulté globale montrent une faible corrélation entre les variables prédites et les

valeurs réelles. Les modèles obtenus sont susceptibles de ne pas réaliser des prédictions satisfaisantes sur de nouvelles données. Il faut confronter ces modèles à la perception humaine, et donc à de nouvelles données, pour déterminer leur robustesse ;
- les paramètres retenus pour l'intelligibilité (la qualité de la prononciation, le débit de parole, les disfluences, le bruit) sont en accord avec l'étude de la littérature et les phénomènes relevés dans la première partie de ce manuscrit.
- aucun des paramètres conservés n'est en lien avec la vidéo. Il est possible que bien que les modalités aient joué un rôle pour faciliter la perception de l'intelligibilité, nous n'ayons pas considéré dans nos paramètres ce qui est réellement entré en jeu pour simplifier la compréhension de la parole.
- Le modèle obtenu par fusion intermédiaire, ne retient seulement que deux paramètres, tout deux en lien avec la complexité du vocabulaire, ce qui atteste de l'importance de l'aspect lexical sur le niveau de compréhensibilité perçu pour un document audiovisuel.

Dans ce chapitre nous voulions étudier le caractère *interprétable* des modèles construits sur un ensemble bien défini de paramètres. La connaissance des paramètres retenus comme influençant le plus le niveau de compréhensibilité permet de faire un retour vers un utilisateur humain, en l'occurrence l'enseignant ou l'apprenant de FLE, qui ait du sens. Nous voulions que la relation entre les paramètres gardés et les prédictions réalisées soient explicables et quantifiables : pour que l'utilisateur confronté au modèle soit capable, à partir des paramètres, d'identifier quels phénomènes sont entrés en jeu dans la prédiction (par exemple la réverbération ou la fréquence lexicale) et dans quelle mesure ils influencent la dimension considérée grâce au coefficient de régression normalisé associé à chacun des paramètres. Les paramètres candidats pour construire les modèles de régression linéaires ont été choisis en se basant sur la littérature, il est maintenant intéressant de voir ce qu'il se passe quand nous exploitons des représentations issues de réseaux de neurones pour alimenter des modèles « non interprétables » (approche dite *neuronale*), c'est-à-dire des modèles pour lesquels nous ne saurons pas, de par la méthode utilisée et/ou le type de modèle de régression choisi, quels phénomènes jouent un rôle dans la prédiction.

Chapitre 7

Mesure objective de la compréhensibilité : approche basée sur les *représentations neuronales*

Sommaire

7.1 Introduction

Nous avons choisi d'aborder la tâche de prédiction du niveau de compréhensibilité de documents audiovisuels selon deux approches. Dans le chapitre précédent, nous avons présenté une approche reposant sur des modèles construits à partir de paramètres candidats issus de connaissances en didactique des langues étrangères et en traitement automatique des langues. L'intérêt de cette première approche est que les paramètres sélectionnés à l'issue de ce processus de modélisation restent interprétables. Ce qui permet d'expliquer ce qui a le plus joué dans la prédiction du niveau des documents audiovisuels traités. De ce fait, le nombre et le type de paramètres extraits dépendent de choix ou de connaissances humaines. Nous nous sommes interrogée sur ce que nous pourrions obtenir à partir de modèles de prédiction construits sur des représentations produites par des réseaux de neurones pré-entraînés. Cette seconde approche pourrait apporter une meilleure prédiction, tout en nous privant de l'aspect interprétable. En effet, en utilisant de telles représentations, nous n'avons pas connaissance de la manière dont chaque réseau a été construit et entraîné [Zhang et al., 2020].

L'objectif de ce chapitre est donc de comparer les performances obtenues à partir de modèles neuronaux par rapport à l'approche « interprétable ». Nous verrons ainsi si les approches sont équivalentes ou s'il existe un avantage à privilégier l'une des approches plutôt que l'autre.

Les modalités audio, texte et vidéo sont exploitées dans de nombreux travaux pour réaliser des tâches très diverses à partir de réseaux de neurones (reconnaissance, détection...). Ceux-ci peuvent être exploités de diverses manières. Ils peuvent être utilisés tels quels, après une phase d'apprentissage réalisée à partir d'un jeu conséquent de données annotées lié à une ou plusieurs tâches ciblées. Dans ce cas, nous parlons de modèles appris par *apprentissage supervisé*. Ces réseaux peuvent être réentraînés, notamment les dernières couches, à partir d'un jeu de données spécifique permettant de les spécialiser pour une tâche donnée répondant à de nouveaux besoins. Nous parlons alors de *d'apprentissage par transfert (ou transfer learning*. Il est également possible d'utiliser le réseau uniquement dans le but d'en extraire des représentations sous forme de vecteurs ou de matrices suivant le cas et correspondant aux valeurs numériques d'une certaine couche du réseau. Ces représentations peuvent à leur tour servir à alimenter un nouveau réseau de neurones ou servir d'entrée à l'apprentissage d'un modèle. Notre objectif n'étant pas de réaliser une optimisation de réseaux de neurones profonds déjà existants pour réaliser nos prédictions, c'est cette dernière option qui nous avons choisie. Nous avons donc travaillé à partir de systèmes existants pour en extraire les représentations et construire de nouveaux modèles de régression.

Nous avons procédé en exploitant les modalités soit individuellement soit simultanément, dans les deux cas, les représentations permettront de prédire une seule dimension.

Dans un premier temps, nous avons identifié et sélectionné des réseaux de neurones profonds adaptés à notre problématique de prédiction du niveau de compréhensibilité de documents audiovisuels, que ce soit au niveau global ou

selon les différentes dimensions identifiées et liées à la complexité du vocabulaire, de la grammaire et à l'intelligibilité.

7.2 Réseaux de neurones utilisés

Les réseaux de neurones, inspirés du fonctionnement des neurones biologiques, sont des solutions utilisées pour l'apprentissage qui se sont popularisées.

Expliqué succintement, le neurone (au coeur de la solution) constitue un opérateur mathématique qui va appliquer une fonction algébrique non-linéaire sur des variables en entrée et ainsi produire une valeur en sortie (voir 7.1).

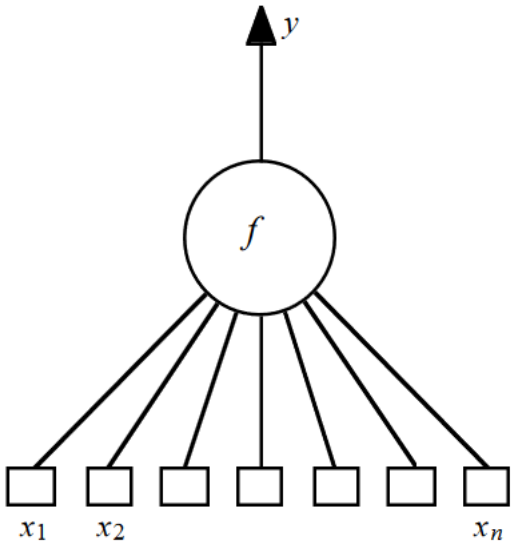

FIGURE 7.1 – Exemple de neurone. Les x_i sont les variables en entrée, y est la sortie

Le réseau de neurone va combiner un ensemble de neurones qui vont être organisés en une ou plusieurs couches (voir 7.2). La couche en entrée récupère les variables en entrée, la couche en sortie, constituée de un ou plusieurs neurones, va produire ce que l'on appelle dans la suite **des représentations**, les couches intermédiaires sont appelées couche cachées.

C'est de l'agencement de ces couches ainsi que des fonctions mathématiques choisies pour chaque neurone dont va dépendre la capacité de l'ensemble du réseau à « apprendre ». Le réseau va être spécialisé pour identifier un ensemble d'informations saillantes à partir des variables en entrée.

Cette section vise à présenter les trois architectures de réseaux de neurones profonds utilisés pour chacune des modalités : texte, audio et vidéo. Pour cha-

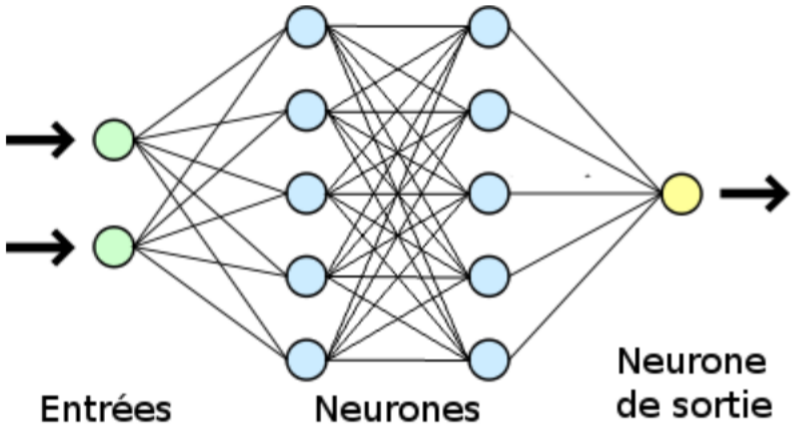

FIGURE 7.2 – Schéma simplifié d'un réseau multicouche

cune d'elles, nous exploitons un modèle pré-entraîné pour en extraire des représentations produites par les différents réseaux de neurones.

7.2.1 Modalité texte : CamemBERT

Le Traitement Automatique du Langage Naturel (TALN) s'est tourné, depuis plusieurs années, vers les réseaux de neurones. Ceci afin de créer et manipuler des représentations numériques qui correspondent aux différents constituants d'une langue (mots, phrases, textes), et ce pour différents types de tâches (transcription de la parole [Yu et al., 2012], traduction [Zbib et al., 2019], chatbot [Serban et al., 2017]...).

Les réseaux de neurones sont capables de capter de multiples informations qui permettent de dégager les caractéristiques d'un langage donné, comme par exemple la façon dont il se structure ou encore les liens sémantiques entre les mots. L'utilisation de modèles de langage construits à partir de réseaux de neurones peut optimiser la réalisation de tâches telles que l'étiquetage morphosyntaxique, la classification de texte, la traduction automatique [Goldberg, 2017]...

7.2.1.1 BERT, RoBERTa et CamemBERT

En 2018, Google a développé un modèle de langage appelé BERT (*Bidirectional Encoder Representations from Transformers*) [Devlin et al., 2018]. Il s'agit d'un réseau de neurones qui a comme particularité majeure de se différencier des autres avec son architecture. Il permet de prendre en compte les relations contextuelles entre les unités choisies (mots ou pseudo-mots *ie* des chaînes de caractères ressemblants à des mots, mais qui ne sont pas porteurs de sens) qui lui sont donnés en entrée en les traitant simultanément et non pas séquentielle-

ment : le réseau ainsi créé peut considérer le contexte de gauche à droite, mais aussi de droite à gauche (aspect bidirectionnel). Pour réaliser un traitement sur un mot à une position donnée, le réseau va donc pouvoir considérer les mots précédents **et** les mots suivants [Vaswani et al., 2017].

Après la création de BERT, un réseau visant à être plus robuste a été développé. L'objectif étant d'améliorer les performances de BERT. En a découlé la création de RoBERTa [Liu et al., 2019]. Ceci a permis d'aboutir à un réseau intéressant pour nous : CamemBERT [Martin et al., 2019], spécialisé pour la langue française.

BERT et RoBERTa sont des réseaux de neurones pré-entraînés pour deux types de tâches :

— le « masked language modeling » qui permet de retrouver la valeur d'un mot masqué dans une phrase,
— le « next sentence prediction » qui permet de déterminer si, dans une paire de phrases, la deuxième phrase est bien consécutive à la première.

CamemBERT est un modèle de langage du français qui a été pré-entraîné uniquement pour la tâche de prédiction de mots masqués, à partir du sous-corpus français du corpus multilingue OSCAR [Ortiz Suárez et al., 2020].

7.2.1.2 Utilisation de CamemBERT

CamemBERT est un modèle de référence pour le français qui, sur plusieurs tâches d'évaluation, permet d'obtenir de meilleures performances comparé à des approches antérieures [Martin et al., 2019]. C'est pour cette raison que nous avons choisi d'utiliser les représentations issues de CamemBERT pour notre problématique. Sa capacité à prédire des mots masqués, en prenant en compte simultanément le contexte gauche et droit permet au modèle d'obtenir des scores intéressants quand il est utilisé et entraîné pour des tâches d'étiquetage morphosyntaxique et d'analyse syntaxique [Martin et al., 2020]. Ceci montre que CamemBERT doit être capable de capter dans un document texte des informations liées à la grammaire et au vocabulaire pour pouvoir prédire correctement le mot masqué.

7.2.2 Modalité audio : PASE+

De part la nature des documents que nous traitons et qui constituent le corpus ESCAL, nous traitons des extraits de films qui correspondent à des interactions entre personnages : les signaux audio avec lesquels nous travaillons contiennent donc de la parole qui peut être dans un environnement bruité et multilocuteurs... L'utilisation d'un outil qui permet de détecter des informations liées à la parole (notamment à la prosodie) mais aussi aux locuteurs est donc intéressante. Pour le traitement de l'analyse des signaux audio, la capacité des réseaux de neurones à trouver des caractéristiques saillantes dans un signal permet de répondre à un très grand nombre de problématiques. Celles qui nous

intéressent sont directement en lien avec le traitement de la parole telles que la détection de la parole, la reconnaissance de locuteur, la transcription de parole en texte... Nous nous sommes donc intéressée à des réseaux de neurones qui ont été spécifiquement développés pour ce type de tâches.

7.2.2.1 PASE

PASE [Pascual et al., 2019] est l'acronyme de *Problem Agnostic Speech Encoder*. Le réseau de neurones profond PASE est entraîné pour la détection de parole, il est donc capable de capter des informations en lien avec la parole du type empreinte vocale ou encore phonèmes. Ce réseau de neurones va dans un premier temps extraire d'un signal audio une représentation encodée. Cette représentation servira à alimenter plusieurs réseaux de neurones simples, chacun constitué d'une seule couche cachée et spécialisé dans **une** tâche **non supervisée** (qui ne nécessite donc pas d'annotation manuelle pour la vérité terrain).

PASE est pré-entraîné pour prédire des paramètres basiques comme les coefficients cepstraux (MFCC), des paramètres liés à la prosodie (PROSO) comme la fréquence fondamentale ou l'énergie, mais aussi des paramètres spécifiques aux locuteurs (LIM). Le modèle PASE génère ainsi des représentations de signaux de parole. Ces représentations permettront de réaliser des tâches telles que la reconnaissance de locuteurs ou la reconnaissance automatique de la parole.

7.2.2.2 PASE+

PASE+ [Ravanelli et al., 2020] est une version améliorée du réseau de neurones PASE, auquel ont été rajoutés des réseaux de neurones spécialisés dans de nouvelles tâches. Il a été conçu pour réaliser de la reconnaissance de la parole dans des environnements **complexes**. Pour rendre le système plus robuste face aux conditions sonores dégradées, une augmentation des données a été réalisée. PASE+ a donc été entraîné sur des signaux de parole contenant de la réverbération, du bruit et de la parole superposée... Chaque signal reçoit des modifications différentes, plusieurs types de modifications pouvant être apportées simultanément. Cela permet d'augmenter la quantité de données d'entraînement tout en y apportant de la diversité.

7.2.2.3 Utilisation de PASE+

Une façon d'exploiter PASE+ est de s'en servir pour extraire des paramètres spécifiques à la parole. C'est l'utilisation qui est faite dans cette thèse, où les représentations sont utilisées pour alimenter des modèles de régression qui permettent de prédire l'intelligibilité de la parole, mais aussi la difficulté globale. De plus, comme PASE+ est entraîné pour être robuste à l'environnement sonore, le réseau peut extraire des informations liées aux bruits et à la réverbération, qui entrent en jeu dans l'intelligibilité et qui sont considérés dans notre étude.

7.2.3 Modalité vidéo : ResNet 3D

La détection et la reconnaissance de personnes et d'objets présents au sein d'une interaction peuvent constituer un apport pour comprendre la situation de communication et donc diminuer la difficulté globale perçue d'un document audiovisuel. Le traitement de l'image est l'un des premiers domaines qui a bénéficié de l'utilisation de réseaux de neurones dont une des applications populaires est la classification. L'évolution des réseaux de neurones a permis d'aboutir à des réseaux capables de capter et d'analyser des informations dynamiques en analysant des images successives. C'est ce type de réseau qui nous intéresse dans cette section.

7.2.3.1 CNN et CNN 3D

Un réseau de neurones convolutionnel (CNN pour *Convolutional Neural Network*) permet d'extraire les paramètres les plus pertinents d'une image et la réduire à une représentation ne contenant que ces paramètres [Albawi et al., 2017]. Le CNN prend en compte la dimension spatiale des images en entrée et considère la hauteur et la largeur des données. Le CNN 3D [Ji et al., 2012] fonctionne d'une façon analogue, à la différence qu'il considère la hauteur, la largeur, mais aussi la profondeur des données et qu'il exploite cette information supplémentaire pour fournir une représentation 3D des données contenant les informations pertinentes. La prise en compte de cette troisième dimension permet d'étudier des images en considérant le contexte précédent et suivant d'une image donnée. L'accès à des informations contextuelles rend l'exploitation des CNN 3D intéressante pour l'étude de vidéos.

7.2.3.2 ResNet 3D

Les ResNets constituent une des dernières avancées pour les tâches liées à l'étude des images et des vidéos. Ils sont utilisés pour la détection d'activités anormales [Dubey et al., 2019] mais aussi pour la classification d'objets [Ioannidou et al., 2019]. L'architecture du réseau permet d'analyser les objets et de classer leur position dans des séquences d'images en prenant en compte les images adjacentes pour ensuite interpréter ou prédire le mouvement. C'est à partir de l'analyse de ce mouvement que l'action réalisée dans la vidéo sera déterminée. Nous nous intéressons à ce réseau de neurones profonds de par sa capacité à détecter et classifier les objets. C'est le réseau configuré et entraîné pour la reconnaissance d'action [Hara et al., 2017] qui est utilisé ici.

7.2.3.3 Utilisation de ResNet 3D

Le réseau est capable d'extraire des informations concernant les mouvements et les actions au sein des vidéos qui mettent en scène des interactions. Dans notre première approche, dite *interprétable*pour prédire la difficulté, les mouvements ont été pris en compte (via les paramètres liés avec la quantité de mouvement), cependant, aucun phénomène en lien direct avec les actions réalisées pendant

les interactions n'avait été considéré. ResNet 3D peut obtenir des informations plus poussées concernant les mouvements ou les actions présents dans le flux vidéo : ces informations peuvent être un apport dans la prédiction de la difficulté globale et/ou de l'intelligibilité.

7.3 Exploitation des réseaux de neurones

Nous voulons utiliser les représentations issues des réseaux de neurones précédemment identifiés, pour alimenter des modèles de régression et comparer cette approche avec l'approche «interprétable», en analysant les métriques produites par les différents modèles.

7.3.1 Extraction des représentations

Pour pouvoir fusionner aisément les représentations issues de nos trois réseaux, chacun dédié à une modalité, il est nécessaire de travailler au même niveau, c'est-à-dire choisir une unité commune. Les représentations ont donc été extraites de façon similaire, sur des segments de même niveau ou même granularité (la granularité pouvant être le fichier complet ou des segments plus petits correspondant à des sous-titres par exemple).

Or, ces réseaux étant entraînés sur des segments de longueur limitée, il n'est pas judicieux de leur donner en entrée la totalité de l'extrait. Dans l'approche «interprétable» nous avons travaillé à différents niveaux de granularité (cf. sectionn 5). Pour garder le même alignement et pour que les résultats soient comparables entre les deux approches, la granularité choisie ici est donc le niveau du sous-titre.

Chaque extrait est segmenté en sous-titres, et pour chaque sous-titre les extractions de représentations se font sur le texte du sous-titre, ainsi que sur le signal audio et la succession d'images du segment temporel correspondant (voir figure 7.3). Ainsi, chaque sous-titre d'un extrait donné a trois représentations : Texte (par CamemBERT), Audio (par PASE+) et Vidéo (par ResNet 3D).

7.3.2 Fusion des représentations

Pour un extrait composé de n sous-titres, nous avons N représentations par modalité. Pour pouvoir réaliser une prédiction pour **l'ensemble de l'extrait**, nous pouvons soit procéder par modalité ou alors en fusionnant les représentations issues des différents réseaux de neurones pour réaliser une fusion intermédiaire et obtenir une prédiction relative à la dimension choisie.

Nous avons deux cas possibles :

— le cas monomodal, où nous cherchons à prédire une dimension avec des représentations issues d'une seule modalité. Par exemple, pour prédire la complexité du vocabulaire ou la complexité grammaticale, nous utilisons les représentations issues de la modalité texte.

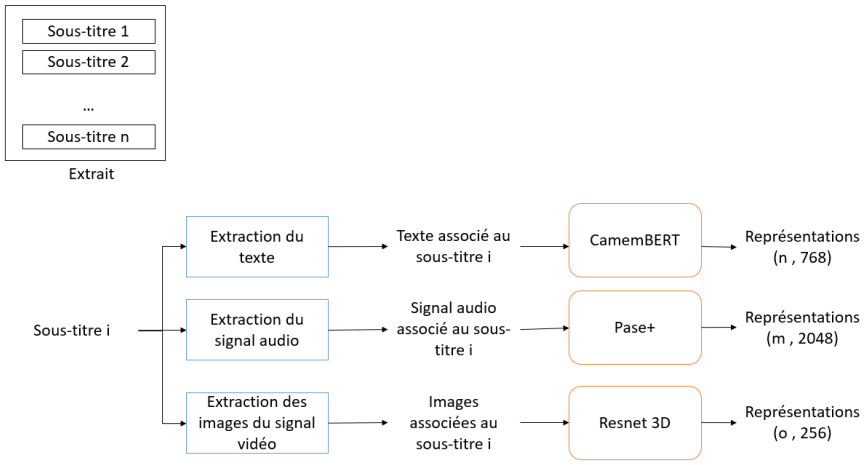

FIGURE 7.3 – Description de la chaîne de traitement par sous-titre

— le cas multimodal, où nous cherchons à prédire une dimension avec des représentations issues de plusieurs modalités. Par exemple, pour prédire la difficulté globale, nous utilisons toutes les modalités (audio, vidéo, texte).

7.3.2.1 Cas monomodal

Dans le cas monomodal, toutes les représentations ont une dimension commune fixe, que nous allons noter m. Chaque réseau de neurones traite la modalité considérée en la découpant en plusieurs parties, appelées *trames* pour la suite. La dimension m dépend du nombre de caractéristiques extraites par le réseau pour une trame donnée. Les représentations ont également une dimension variable qui est dépendante de la longueur du sous-titre lui-même (en termes de nombre de mots pour la modalité texte ou de durée pour la modalité audio).

Nous ramenons ces représentations à un seul vecteur obtenu comme suit :

1. les N représentations sont concaténées pour former une matrice de dimension (N, m), avec N qui correspond au nombre de vecteurs de représentations obtenus sur l'ensemble de l'unité traitée. Par exemple, en fonction de la modalité considérée, si le sous-titre considéré comporte 10 mots, le segment audio 50 trames et le clip vidéo 30 images, nous aurons respectivement 10, 50 et 30 représentations à concaténer,
2. à partir de la matrice ainsi obtenue, trois vecteurs (moyen, médian et écart-type) sont calculés, chacun avec une dimension $(1, m)$,
3. en concaténant ces trois vecteurs, nous obtenons un vecteur de dimension $(1, 3 * m)$.

L'ensemble de ces étapes est résumé dans la figure 7.4.

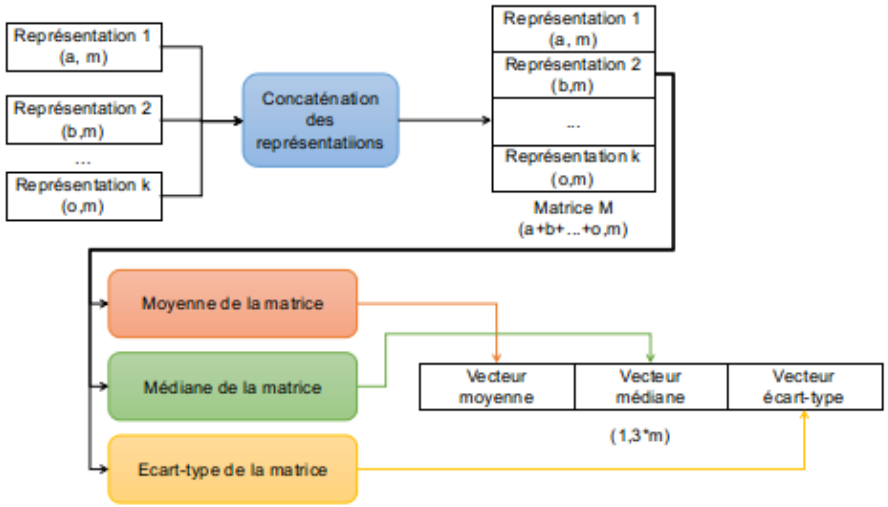

FIGURE 7.4 – Fusion des représentations pour le cas monomodal

7.3.2.2 Cas multimodal

Dans le cas où il y a plusieurs modalités exploitées, nous commençons par réaliser le prétraitement cité précédemment pour que chaque modalité ne soit plus représentée que par un vecteur de dimension $(1, 3*m)$. Ensuite, ces différents vecteurs sont concaténés à leur tour pour garder une représentation vectorielle de l'extrait. Nous avons :

- N_T valeurs pour la représentation Texte (CamemBERT),
- N_A valeurs pour la représentation Audio (PASE+),
- N_V valeurs pour la représentation Vidéo (ResNet 3D)).

Ainsi, la modalité audio donne lieu à un vecteur de dimension $(1, 3 * N_A)$, la modalité vidéo à un vecteur de dimension $(1, 3 * N_V)$ et la modalité texte à un vecteur de dimension $(1, 3 * N_T)$.

7.3.3 Prédictions

Chaque document est représenté par un vecteur qui a été construit à partir d'une ou plusieurs modalités. Nous nous retrouvons alors dans un cas similaire à l'approche décrite dans le chapitre 6. La différence repose sur le fait que ces vecteurs ne sont pas interprétables en termes de facteurs qui vont influencer le niveau de compréhensibilité du document traité. L'ensemble des vecteurs obtenus permet de réaliser des prédictions. Pour cela, nous allons nous intéresser à une famille de modèles, les étudier et comparer la qualité des prédictions

obtenues. Les modèles testés sont parmi ceux les plus traditionnellement utilisés pour répondre à des problématiques de régression :

— les machines à vecteurs supports (SVM),
— les K-plus proches voisins (KNN pour *K Nearest Neighbours*),
— les forêts d'arbres décisionnels (RF),
— le perceptron.

Quelques éléments caractéristiques de chacune des méthodes testées sont détaillés ci-après.

7.3.3.1 Machines à vecteurs supports pour la régression

Dans sa version la plus simple, un SVM, utilisé pour un problème de classification, sert à trouver un hyperplan qui permet de séparer au mieux deux classes dans un espace de représentation [Hearst et al., 1998a]. Suivant la complexité des données, il faut recourir à l'utilisation de noyau de type linéaire, Gaussien (ou rbf) pour changer d'espace de représentation et se positionner dans un espace de plus grande dimension pour y trouver un hyperplan séparateur. Le type de noyau choisi a un impact sur la façon dont les données sont séparées. Dans le cadre de cette étude, nous avons choisi d'utiliser le noyau le plus classique : le noyau gaussien, qui permet de décrire implicitement les données de l'espace de départ dans un espace de dimension infinie.

Initialement conçues pour des tâches de classification, il est possible d'étendre l'utilisation des SVM pour des problèmes de régression dans le cas où nous souhaitons réaliser la prédiction de valeurs continues [DJEFFAL, 2012]. L'application des vecteurs de supports pour la régression revient à un problème de recherche d'une fonction f. Sachant que la fonction f est une fonction linéaire du type $f = (w, x) + b$ avec w un vecteur et b un scalaire, et qu'un hyperplan se caractérise par un couple $(w*, b*)$, trouver la fonction f revient à trouver l'hyperplan qui va permettre de minimiser l'écart entre les $f(x_i)$ (fonction f appliquée sur les données en entrée x_i) et les y_i qui sont les valeurs d'entraînement à trouver.

La figure 7.5 illustre le fonctionnement des SVM pour la régression.

Soit y l'hyperplan, encadré de frontières séparées de ϵ. Ces frontières constituent un **hypertube** de largeur 2ϵ. Les points situés dans l'hypertube correspondent à l'ensemble des exemples d'entraînement. En sachant que, comme dans la problématique de classification, plusieurs hypertubes candidats existent, l'objectif est de fixer le ϵ optimal tel que le meilleur hypertube soit celui pour lequel la distance des exemples d'entraînement à l'hyperplan est maximisée. De la même façon que pour la classification, il est possible d'appliquer des noyaux pour transformer l'espace de départ et se trouver dans le cas d'une régression linéaire et chercher la solution dans ce nouvel espace.

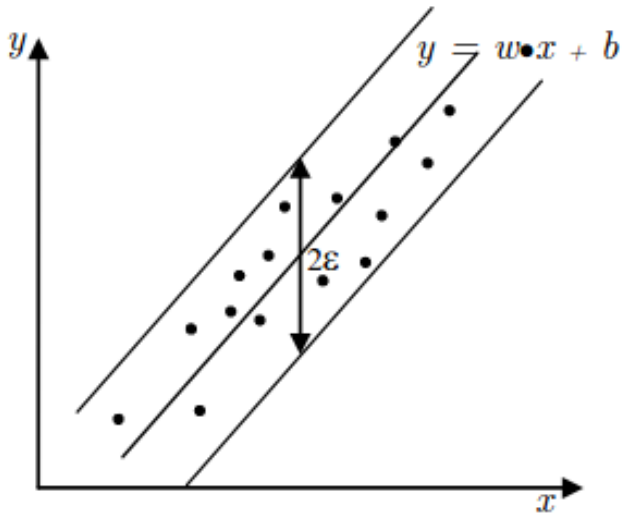

FIGURE 7.5 – Illustration du fonctionnement des SVM pour la régression, issue de [DJEFFAL, 2012]

7.3.3.2 Les K plus proches voisins

Le KNN est usuellement utilisé pour des problématiques de classification. Les classes sont représentées par leur coordonnées (x_i, y_i) dans un même plan. Soit un nouveau point X de coordonnées (x, y), le choix de la classe d'appartenance se fait à partir des K points les plus proches (plus proches voisins vis-à-vis d'une distance) du point X [Peterson, 1883]. La classe attribuée au point X dépend du nombre de voisins les plus proches considérés.

Dans le cas de la régression, nous cherchons non plus la classe pour un objet en entrée, mais sa valeur. Comme pour le problème de classification, l'algorithme prend d'abord en entrée un ensemble de données qui correspondent aux échantillons d'apprentissage. Pour prédire la valeur de la nouvelle donnée, l'algorithme étudie l'ensemble des échantillons d'apprentissage et les K plus proches voisins sont les K échantillons les plus similaires à la donnée en entrée, la valeur de l'objet est alors la moyenne des valeurs de ses K plus proches voisins. Dans le cadre de cette étude, la valeur de K est déterminée automatiquement à l'aide d'une validation croisée.

7.3.3.3 Perceptron

Le perceptron monocouche constitue le réseau de neurones artificiel le plus simple [Noriega, 2005]. Il s'agit d'un type de réseau initialement destiné à faire de la classification binaire et se limite au cas où les données sont linéairement séparables. Composé d'une couche d'entrée avec N neurones et une couche de

sortie avec un seul neurone, le réseau alimenté en *feed-forward* : les informations circulent de la couche d'entrée à la couche de sortie.

Pour des cas non linéaires, il est nécessaire d'utiliser des perceptrons multicouches qui comptent une ou plusieurs couches situées entre la couche d'entrée et la couche de sortie : ces couches sont appelées *couches cachées* et chaque couche cachée peut contenir un nombre de neurones différents.

Pour les perceptrons monocouches **et** multicouches, chaque liaison interneuronale a un poids. L'ensemble de ces poids affecte la décision finale du système pour la classification. L'erreur apportée par chaque neurone est minimisée en modifiant les poids grâce une propagation de l'information par rapport à l'erreur de la couche de sortie vers la couche d'entrée.

Si les données sont continues, un perceptron peut être utilisé pour réaliser une régression, la prédiction en sortie est alors calculée en utilisant une fonction de transfert qui a une sortie linéaire. Les poids des liaisons synaptiques sont modifiés pour minimiser l'erreur de la prédiction. Dans les problématiques de régression, le nombre de couches cachées va avoir une influence sur la qualité des prédictions. Dans notre étude, nous avons testé des perceptrons à plusieurs couches cachées (1, 2 et 3).

7.3.3.4 Forêt d'arbres décisionnels

Arbre de décision

Un arbre de décision est un modèle représenté sous forme d'arbre qui sert à réaliser des tâches de classification ou de régression [Breiman, 2001]. Il permet de faire un choix entre plusieurs alternatives possibles en fixant des critères pour discriminer un ensemble de données à chaque embranchement de l'arbre de décision (appelé nœud), les feuilles de l'arbre représentant le choix possible à l'issue d'une succession de branches spécifiques.

Forêt d'arbres décisionnels

Les *Random Forest* combinent plusieurs arbres de décision. Les résultats obtenus par chaque arbre de décision sont associés pour pouvoir obtenir un modèle de classification **ou** de régression plus fiable. La décision finale pour classer ou déterminer la valeur d'une donnée inconnue pourra être prise :

- en réalisant un vote majoritaire à partir du résultat donné par chaque arbre s'il s'agit d'une classification,
- en faisant la moyenne de la prédiction de tous les arbres s'il s'agit d'une régression.

Nombre d'arbres de décision

Le nombre d'arbres de décision qui compose une forêt d'arbres décisionnels est variable (une forêt peut compter plus d'une centaine d'arbres de décision) et va avoir une influence sur la qualité du modèle final. Le choix du nombre d'arbres va dépendre du problème. Plusieurs études ont été menées pour déterminer le nombre d'arbres de décisions à utiliser lorsque nous voulons nous servir d'une

forêt d'arbres décisionnels. Ici, le nombre d'arbres à tester a été fixé à partir de l'étude réalisée dans [Oshiro et al., 2012]. Les nombres d'arbres testés par la suite sont : 64, 100 et 128.

7.3.3.5 Synthèse des systèmes de prédiction utilisés

Au total huit classifieurs sont testés :
— SVM, avec noyau gaussien,
— KNN, avec recherche automatique du meilleur K,
— 3 Perceptrons, avec respectivement une, deux et trois couches cachées,
— 3 Random Forest, avec respectivement 64, 100 et 128 arbres de décision.

7.4 Meilleurs classifieurs de l'approche *neuronale*

Pour comparer les performances entre les classifieurs, nous utilisons les mêmes métriques que dans le chapitre précédent : le r (coefficient de Pearson) et la $RMSE$ (erreur quadratique moyenne). Nous analysons ces mesures lorsque nous réalisons un *Leave-One-Out*. Les meilleurs classifieurs pour chacune des dimensions considérées sont présentés dans cette section.

7.4.1 Représentation de la Modalité Texte

Le réseau de neurones profond CamemBERT fournit en sortie des représentations de dimension (n, N_T) avec n dépendant du **nombre de mots** en entrée du réseau de neurones et $N_T = 768$ qui correspond au nombre de neurones de la couche cachée du réseau. Lorsque nous travaillons au niveau de l'extrait, nous obtenons donc un vecteur issu de la concaténation de 3 vecteurs correspondant respectivement à la moyenne, la médiane et l'écart-type des vecteurs représentant chaque sous-titre, soit une dimension $(1, 3 * 768)$.

7.4.2 Complexité du vocabulaire

Le SVM appliqué à la régression (SVR) permet de maximiser la valeur du coefficient r de Pearson et de minimiser la $RMSE$ (avec $r = 0,53$ et $RMSE = 14,5$). La valeur de r met en avant l'existence d'une corrélation positive moyenne entre les valeurs prédites et les valeurs réelles. Ce résultat montre que le réseau de neurones CamemBERT génère des représentations dans lesquelles se trouvent certaines informations textuelles en lien avec la complexité du vocabulaire.

Utiliser des représentations issues de CamemBERT permet d'obtenir des résultats proches de ceux obtenus avec le meilleur modèle *interprétable* de régression linéaire multiple, qui avait un $r = 0,64$ et un $RMSE = 13,31$. L'étude de la différence des corrélations conclut au fait que la différence des coefficients n'est pas significative, mais nous notons cependant que le SVM pour la régression aboutit à un modèle qui est moins précis que le modèle *interprétable*.

7.4.3 Complexité de la grammaire

Le SVM pour la régression est également la méthode qui permet d'obtenir les meilleurs r et $RMSE$ pour la prédiction de la complexité grammaticale. Avec $r = 0,56$ et $RMSE = 9,22$, nous notons une corrélation moyenne entre les données prédites et les données réelles, mais aussi de faibles erreurs de prédiction. Le réseau de neurones profond CamemBERT est capable d'identifier des informations pertinentes en lien avec la syntaxe et la morphologie. Il n'est cependant pas possible de savoir si, comme pour le meilleur modèle obtenu pour le cas de l'approche *interprétable*, l'aspect lexical est également entré en jeu dans la prédiction de la complexité grammaticale.

Dans le cas de l'approche *interprétable*, nous obtenions, à partir de paramètres en lien conjointement avec la complexité lexicale et la complexité grammaticale, un modèle avec un $r = 0,53$ et un $RMSE = 9,38$. L'absence de différence significative entre les valeurs des coefficients de corrélation r ainsi que le faible écart entre les $RMSE$ montre que les deux modèles sont équivalents, ce qui nous ne nous permet pas de conclure sur la meilleure approche pour réaliser la prédiction de la complexité de la grammaire.

7.4.4 Intelligibilité de la parole

Dans le cadre de l'approche *interprétable*, les prédictions ont été réalisées en considérant simultanément des paramètres issus de la dimension audio et des paramètres issus de la dimension vidéo. Pour que les résultats soient comparables, pour l'approche *neuronale*, les classifieurs ont été entraînés :

— avec les représentations sonores, issues de PASE+,
— avec les représentations sonores et visuelles, issues respectivement de PASE+ et ResNet3D.

En exploitant conjointement les représentations des deux réseaux de neurones, nous nous plaçons dans le cas multimodal. Les représentations issues de PASE+ sont de dimension (n, N_A) avec n dépendant du nombre de trames traitées (PASE+ traite un signal en entrée en le découpant en segments, une trame correspond donc à un segment audio d'une durée de 0,01s) et $N_A = 256$, dimension de la représentation d'une trame. Pour ResNet3D nous acons un vecteur de dimension (m, N_V), avec m qui correspond au nombre de trames traitées par le réseau ResNet3D (une trame ayant une durée de 0,66s) et $N_V = 2048$, dimension de la représentation de chaque trame.

Nous calculons tout d'abord, pour chacune de ces deux matrices, un vecteur moyen, un vecteur médian et vecteur correspondant à l'écart-type. En concaténant ces 3 vecteurs nous obtenons une nouvelle représentation vectorielle de chaque sortie, de dimensions $(1, 3 * 256)$ et $(1, 3 * 2048)$ respectivement. Ce deux vecteurs sont à leur tour concaténés pour donner une représentation finale de dimension $(1, 3 * 256 + 3 * 2048)$ alors fournie en entrée du classifieur considéré.

Tous les classifieurs testés amènent à des modèles ayant des performances insatisfaisantes avec une corrélation r très faible entre les données prédites et

les données réelles et un $RMSE$ élevé. L'approche *neuronale*, fondée sur les représentations conjointes sonores et visuelles est inappropriée pour réaliser la prédiction de l'intelligibilité de la parole. Le meilleur modèle *interprétable*, avec $r = 0,48$ et un $RMSE = 17,61$ est plus pertinent.

7.4.5 Difficulté globale

Pour réaliser la prédiction de la difficulté globale, les représentations exploitées sont issues des trois réseaux de neurones profonds à notre disposition. Après avoir constitué pour chaque réseau le vecteur (moyenne, médiane, écart-type), nous avons trois vecteurs de dimension :

— $(1, 3 * 768)$ pour le texte,
— $(1, 3 * 256)$ pour l'audio,
— $(1, 3 * 2048)$ pour la vidéo.

Pour alimenter les classifieurs, nous concaténons ces trois vecteurs : les vecteurs en entrée d'un classifieur sont de dimension $(1, 3 * 768 + 3 * 256 + 3 * 2048)$.

En termes de valeur absolue, le classifieur qui permet d'obtenir le r le plus élevé est le SVR avec $r = -0,89$. Ceci indique une corrélation très forte entre les valeurs réelles et les valeurs prédites. Ainsi, au moins un des trois réseaux de neurones exploité permet de trouver des informations qui ont beaucoup d'influence pour évaluer la difficulté globale d'un extrait. Faire des études complémentaires en faisant varier les données en entrée pourrait permettre d'identifier quel réseau entre PASE+, ResNet 3D et CamemBERT a eu le plus de poids sur la qualité des prédictions de la difficulté globale.

Le SVR donne un RMSE plus élevé $RMSE = 19,13$ comparé à la régression linéaire multiple $RMSE = 17,88$. Cependant, le SVR est meilleur en valeur absolue avec un $r = -0,89$. Ceci met en avant l'existence d'une corrélation forte entre valeurs prédites et valeurs réelles, face à un $r = 0,38$ pour le modèle issu de l'approche *interprétable*. Néanmoins, le signe négatif du coefficient de corrélation signifie que le modèle prédit quasiment systématiquement la valeur opposée à la valeur réelle.

7.5 Bilan des approches *interprétable* et *neuronale*

Deux approches différentes pour la construction de modèles de prédiction du niveau de compréhensibilité ont été explorées :

— la première consistait à choisir les meilleurs paramètres pour la prédiction parmi un ensemble de paramètres sélectionnés à partir de connaissances en didactique et en traitement automatique : l'approche «interprétable» ;
— la seconde consistait à construire des modèles en utilisant des représentations issues de réseaux de neurones : l'approche «neuronales».

Les meilleurs modèles, selon les métriques utilisées r (corrélation de Pearson) et $RMSE$ (erreur quadratique moyenne) obtenues avec la méthode de validation croisée *Leave-One-Out*, issus de chacune des approches sont rappelés à titre de comparaison, dans la table 7.1 et ce, pour chacune des dimensions considérées.

TABLE 7.1 – Comparaison des performances des approches *interprétable* et *neuronale*, en *Leave-One-Out* sur notre corpus ESCAL

Dimension	Approche *interprétable*	Approche « IA »	
	Résultats	Résultats	Modèle
Vocabulaire	$r = \mathbf{0{,}64}$ $RMSE = \mathbf{13{,}31}$	$r = 0,53$ $RMSE = 14,5$	SVR
Grammaire	$r = 0,53$ $RMSE = 9,38$	$r = \mathbf{0{,}56}$ $RMSE = \mathbf{9{,}22}$	SVR
Intelligibilité	$r = \mathbf{0{,}48}$ $RMSE = \mathbf{17{,}61}$	$r = -0,17$ $RMSE = 20,1$	K-NN
Difficulté globale	$r = \mathbf{0{,}38}$ $RMSE = \mathbf{17{,}88}$	$r = -0,89$ $RMSE = 19,13$	SVR

Pour résumer, et en considérant chacune des dimensions qui interviennent dans la prédiction du niveau de compréhensibilité global, nous pouvons tirer les conclusions suivantes :

Complexité du vocabulaire l'approche «interprétable» aboutit à des modèles pour lesquels la relation linéaire entre données prédites et données réelles est plus forte, mais la différence des r n'est cependant pas significative. La valeur du $RMSE$ permet de voir quelle est l'approche qui amène la meilleure précision : avec $RMSE = 13,31$ (contre $14,5$) l'approche «interprétable» améliore comparativement la précision d'environ 9%. Cependant, même si les $RMSE$ n'ont pas un écart important, à r équivalents, nous préférerons l'approche qui va maximiser la précision.

Complexité grammaticale les performances des modèles sont équivalentes avec des r qui ne sont pas significativement différents, et des $RMSE$ très proches. Dans ce cas, nous favoriserons un modèle «interprétable» pour une exploitation industrielle car nous souhaitons que l'utilisateur reçoive une information indiquant quels facteurs sont rentrés en jeu dans la prédiction et dans quelle proportion.

Dans la perspective du développement d'un outil d'aide à l'indexation de contenus authentiques, tel que prévu à l'issue de cette étude, un tel modèle sera plus accessible pour un utilisateur cherchant à comprendre comment les prédictions sont réalisées. Cela lui permettrait de choisir les séquences les plus appropriées au type d'exercices ciblés.

Intelligibilité de la parole l'approche *neuronale* fournit des modèles de mauvaise qualité avec des r très faibles entre valeurs prédites et valeurs réelles. Ce constat nous amène à conclure qu'ici aussi, le modèle issu de l'approche *interprétable* est plus approprié.

niveau de compréhensibilité global le constat plus mitigé. Si l'approche «interprétable» permet d'obtenir un r positif qui reflète une corrélation faible entre données prédites et données réelles, l'approche «neuronale» obtient un r qui démontre une corrélation très forte entre les données prédites et réelles. Cependant, ce coefficient a pour principal désavantage d'être négatif. Le modèle prédira donc quasiment systématiquement la valeur opposée à la valeur réelle. Le modèle obtenu n'est pas utilisable en l'état pour prédire la difficulté globale d'un document audiovisuel. Utilisé tel quel, il n'est pas pertinent. Néanmoins, à des fins applicatives, un post-traitement réalisé sur les prédictions en sortie de ce modèle, permettrait de les corréler positivement aux valeurs réelles. Cela permettrait d'obtenir un prédicteur très performant. Pour la suite de nos travaux, nous continuerons à nous intéresser au modèle de régression linéaire multiple obtenu avec l'approche «interprétable».

Grâce à cette étude comparative, nous avons pu constater que, d'un point de vue quantitatif, les régressions linéaires multiples fournissent les meilleures prédictions. Il nous a semblé intéressant pour clore ce travail de confronter les divers modèles obtenus contribuant à la définition d'une mesure objective du niveau de compréhensibilité, à un panel d'utilisateurs experts pour voir si leur perception du niveau de compréhensibilité est en accord avec les prédictions fournies par les nos modèles. C'est l'objectif du chapitre suivant.

Chapitre 8

Mesure objective de la compréhensibilité : une approche pertinente ?

Sommaire

8.1 Introduction

Les chapitres 6 et 7 ont été consacrés à la prédiction du niveau de compréhensibilité selon deux approches, appliquées chacune sur les données du corpus ESCAL. Les résultats obtenus ont été comparés. En se basant sur le modèle ayant donné les meilleurs résultats de prédiction, obtenu avec l'approche «interprétable», nous allons maintenant étudier son comportement face à un nouveau jeu de données et comparer celui-ci avec le comportement d'utilisateurs ayant participé à une seconde expérimentation mise en place pour collecter une nouvelle série d'évaluations du niveau de compréhensibilité à partir de nouveaux extraits de documents audiovisuels.

8.2 Nouvelle expérimentation

8.2.1 Description des données à évaluer

En restant dans une philisophie identique à celle de la constitution du corpus ESCAL, le nouveau jeu de données est constitué de 10 documents audiovisuels, dont huit sont des extraits des mêmes films que ceux qui avaient été utilisés pour la création du corpus ESCAL. Pour s'assurer que le modèle n'a pas été biaisé par une notion de connaissance des films, deux extraits issus de deux nouveaux films ont également été utilisés. La table 8.1 présente les différents films qui ont été exploités ainsi que la durée des extraits associés aux titres. Cela représente une durée totale de 9 minutes et 36 secondes.

TABLE 8.1 – Corpus de 10 extraits pour la validation de notre modèle de prédiction de la difficulté globale

Film	Durée de l'extrait
Amélie Poulain	57s
Cyrano de Bergerac	54s
Intouchables	1min 43s
La chèvre	17s
La cité de la peur	38s
La folie des grandeurs	24s
La gloire de mon père	43s
Les plages d'Agnès	26s
LOL	1min 2s
Qu'est-ce qu'on a fait au bon Dieu ?	2min 32s

8.2.2 Participants

Comme pour l'expérience ayant abouti à la création du corpus ESCAL (voir chapitre 3), des enseignants de FLE ont été sollicités. Ils ont été recrutés sur la base du volontariat sur des groupes dédiés aux professeurs et apprenants de FLE

sur les réseaux sociaux. Au total, 22 personnes (dont 20 francophones natifs) ont pris part à l'expérience : 7 hommes et 15 femmes, entre 25 et 69 ans (âge moyen : 43,2) et 2 à 40 ans d'expérience d'enseignement du FLE (moyenne : 14,8, écart-type : 11,49). Tous les participants sont normo-entendants (autodiagnostiqués) pour s'assurer que leur perception de la compréhensibilité ne soit pas influencée par des problèmes d'audition.

8.2.3 Protocole expérimental

Pour permettre aux participants d'annoter ce nouveau jeu de données en fonction du niveau de compréhensibilité perçu, une interface d'évaluation a été créée à partir du logiciel Prodigy [1] pour proposer deux tâches aux participants :

Classement : c'est-à-dire classer les dix extraits en fonction de la **difficulté globale** perçue (voir figure 8.1). Les extraits sont présentés avec l'ensemble des modalités disponibles (AVT) et doivent être classés en réalisant un glisser-déposer de chaque bloc contenant l'ensemble vidéo-texte. Une jauge colorée du vert (niveau jugé très facile) au rouge (niveau jugé très difficile) placée sur le côté servant de repère aux participants. Ceux-ci peuvent visionner les extraits autant qu'ils le souhaitent et modifier le classement autant que nécessaire avant de le valider.

Scores détaillés : il s'agit d'attribuer à chacun des dix extraits un **score de difficulté relatif à la grammaire, au vocabulaire et à l'intelligibilité de la parole**. La méthode d'annotation est similaire à celle de la première expérience qui avait permis de créer le corpus ESCAL (cf. chapitre 3), à la différence que la difficulté globale du document n'est pas évaluée et qu'aucune justification n'est demandée aux participants (voir figure 8.2).

Remarque : pour chaque participant, les extraits ont été présentés dans un ordre différent pour s'assurer que les classements ont été constitués indépendamment de l'ordre de présentation des extraits.

À l'issue de cette expérience, nous avons un ensemble d'évaluations réalisées par les 22 participants de cette expérience avec :

— 22 classements des dix extraits par difficulté globale croissante,
— 220 scores de difficulté de la grammaire,
— 220 scores de difficulté du vocabulaire,
— 220 scores de difficulté à percevoir la parole.

Pour chacun des dix documents, nous extrayons l'ensemble des paramètres nécessaires pour alimenter le modèle de régression linéaire multiple issu de l'approche «interprétable» (voir table 6.7). Nous obtenons ainsi dix prédictions objectives du niveau de compréhensibilité.

Nous avons d'une part les évaluations subjectives réalisées par les 22 annotateurs humains et d'autre part l'évaluation objective obtenue avec notre système.

1. https://prodi.gy/

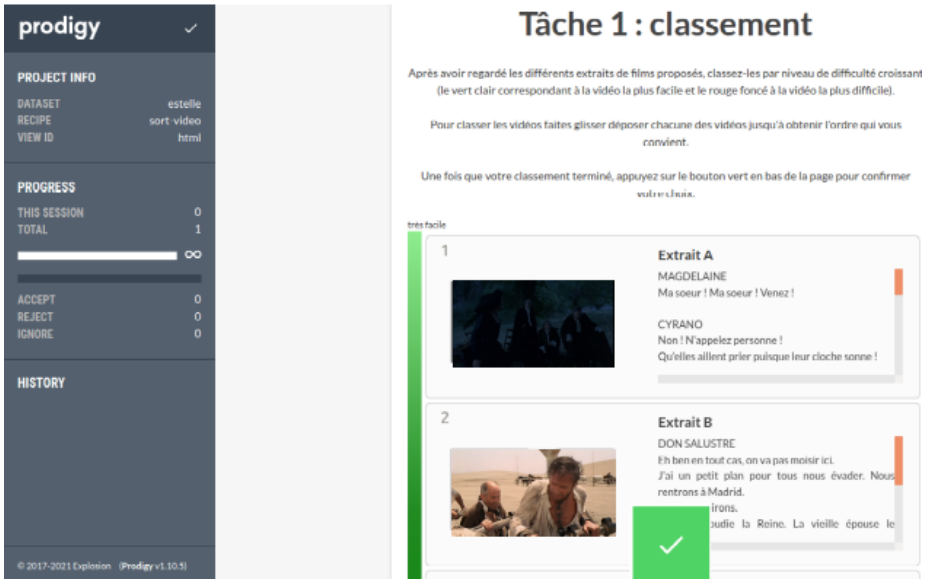

FIGURE 8.1 – Interface de classement

8.3 Analyse des résultats

8.3.1 Clusters de participants

Pour répondre à la première question *Existe-t-il des tendances dans la façon d'annoter des participants de l'expérience ?*, nous appliquons une méthode de clustering en nous basant sur les classements par niveau de difficulté globale croissant atribués par les participants lors de la première tâche. Pour cela, nous utilisons les K-Means [Likas et al., 2003], une méthode de séparation d'éléments en fonction d'une valeur K, un nombre entier positif. Pour un K donné, nous cherchons à diviser un ensemble d'éléments en K groupes en minimisant la distance entre les points appartenant à un même groupe.

Pour estimer le nombre idéal de clusters, nous avons utilisé le coefficient de silhouette [Rousseeuw, 1987]. Ce coefficient est une mesure comprise entre -1 et 1, qui permet d'étudier la qualité d'une partition de données en comparant la distance d'un point par rapport à ses voisins et la distance par rapport aux points des autres clusters. La première distance permet d'évaluer la **cohésion** du cluster, tandis que la seconde évalue la **séparation** des clusters. Le coefficient de silhouette combine les mesures de cohésion et de séparation pour déterminer l'**homogénéité** des clusters en déterminant si les points appartiennent au cluster qui convient le mieux ou s'il y a une ambiguïté concernant la séparation des clusters. Un coefficient proche de 1 indique qu'un point donné est dans le cluster approprié, un coefficient proche de -1 montre que le point est plus proche des points d'autres clusters et qu'il n'est donc pas dans le bon groupe. Il faut donc

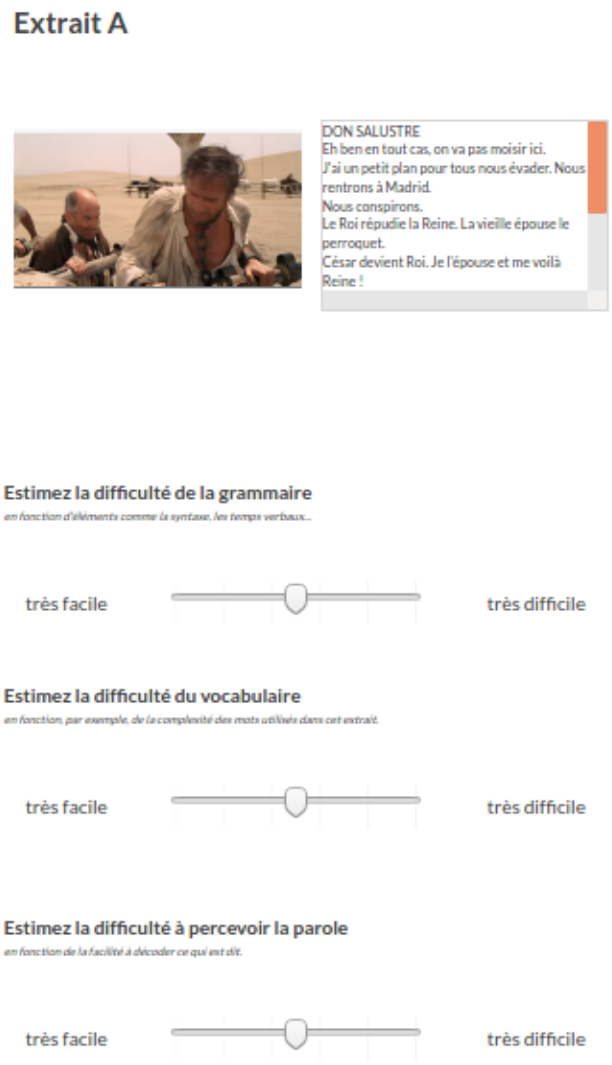

FIGURE 8.2 – Interface d'annotation

idéalement que le coefficient de silhouette se rapproche de 1.

Pour déterminer le nombre optimal de clusters, nous faisons varier K : un coefficient de silhouette moyen est calculé et, à partir d'une représentation graphique, nous prenons connaissance de la largeur (le nombre d'éléments à l'intérieur du cluster) et la hauteur (la valeur du coefficient de silhouette pour un élément du cluster) des clusters. Le nombre optimal est celui pour lequel :

— le score moyen de silhouette est le plus proche de 1 : cela signifie qu'en moyenne tous les points se trouvent dans le cluster approprié et que les

clusters sont bien séparés,
— la largeur des clusters est la plus uniforme possible : cela signifie que les clusters sont équilibrés en termes de nombres d'éléments,
— la hauteur de chaque cluster est au-dessus du coefficient de silhouette moyen.

La figure 8.3 permet de voir les scores moyens de silhouette en fonction du nombre K de clusters et de visualiser de quelle manière varient les clusters en fonction du nombre de clusters, avec K compris entre 2 et 5.

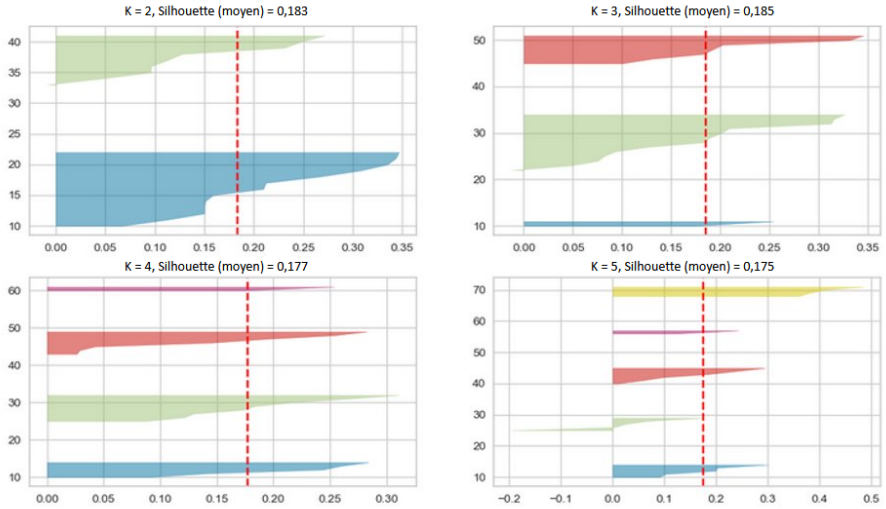

FIGURE 8.3 – Coefficient de silhouette : variation des clusters en fonction de la valeur de K. La ligne verticale rouge en pointillée indique la valeur du coefficient de silhouette moyen.

Pour $K \in \{3, 4, 5\}$, les éléments ne sont pas répartis de façon équilibrée dans les clusters obtenus. Le déséquilibre le plus notable étant pour $K = 3$, avec un cluster contenant un seul élément, bien que cette valeur de K permette de maximiser la valeur du coefficient moyen de silhouette avec une valeur de 0,185. Le déséquilibre est moindre pour $K = 2$, qui permet d'autre part d'obtenir le second meilleur coefficient de silhouette avec une valeur de 0,183. Pour la suite, nous fixons $K = 2$: le premier cluster comporte 8 participants et le second 14.

8.3.2 Analyse des clusters

8.3.3 Comparaison des classements

Suite à l'étape de clustering, nous souhaitons analyser les différences notables dans la manière dont les participants ont perçu la difficulté globale, en nous

référant aux classements par ordre croissant de difficulté des extraits réalisés lors de la première tâche. Pour cela, nous regardons le **classement médian** de chacun des extraits (classés de 1 à 10, avec 1 l'extrait le plus facile et 10 l'extrait le plus difficile) pour chacun des deux clusters formés.

Nous regardons ensuite, pour chacun des dix extraits, s'il existe une différence significative de la valeur médiane de classement entre les deux clusters. Cette analyse se fait à partir du test de Wilcoxon-Mann-Whitney qui sert à estimer si deux distributions sont égales [Nachar et al., 2008]. Pour chaque extrait, nous appliquons le test de Wilcoxon-Mann-Whitney à l'ensemble des positions de l'extrait dans les classements du cluster 1 et à l'ensemble des positions de l'extrait dans les classements du cluster 2. Dans le cas où la distribution n'est pas égale, nous considérons que les participants des deux clusters ont eu une façon différente de percevoir la difficulté de l'extrait, qui a entraîné une différence notable dans la manière de le classer. Les résultats sont présentés dans la table 8.2.

TABLE 8.2 – Comparaison des classements médians des extraits, * indique que les classements médians sont significativement différents

Extrait	Cluster 1	Cluster 2
Amélie Poulain	6,5	7,5
Cyrano de Bergerac	9,0	8,5
La cité de la peur	6,0	5,0
La folie des grandeurs	7,5	7,5
Les plages d'Agnès	2,5	2,0
Intouchables*	9,0	6,0
La chèvre*	2,0	7,0
LOL*	4,5	2,5
Qu'est-ce qu'on a fait au bon Dieu ?*	7,5	2,5
La gloire de mon père*	5,0	8,5

Parmi les dix extraits, nous notons cinq extraits pour lesquels le classement médian est significativement différent (cinq dernières lignes du tableau) : pour ces cinq extraits, les participants des deux clusters ont perçu la difficulté de façon différente, ce qui a mené à une différence notable dans la position de ces extraits dans leurs classements médians. Dans la suite nous nous intéressons à ces cinq extraits pour étudier les phénomènes à l'origine de ces différences.

8.3.4 Étude des extraits significativement différents

La seconde tâche de l'expérience consistait à faire évaluer les trois dimensions : la difficulté du vocabulaire, de la grammaire et la difficulté à percevoir la parole. Pour les extraits dont le classement médian est significativement différent entre les deux clusters (voir tableau 8.2), nous analysons les évaluations faites par les participants pour voir quelles pistes pourraient expliquer la diffé-

rence constatée.

D'une façon analogue à celle présentée dans la sous-section 8.3.3, nous utilisons le test de Wilcoxon-Mann-Whitney pour déterminer si, pour un extrait donné et une dimension donnée, la médiane des scores attribués est significativement différente entre les deux clusters. Si tel est le cas nous nous intéressons au cluster pour lequel le score médian ainsi observé est le plus élevé. Si, dans ce cluster, le document a un classement médian plus élevé, nous pourrons supposer que différence de perception des participants de la dimension concernée a influé sur la perception de la difficulté globale et explique la différence de classement médian du document entre les deux clusters.

En suivant ce principe nous observons que quand les participants du cluster 1 ont perçu un extrait comme globalement moins compréhensible que le cluster 2, la différence de classement ne semble due ni au vocabulaire, ni à la grammaire, ni à l'intelligibilité.

Ceci est le cas des extraits d'*Intouchables* et de *Qu'est-ce qu'on a fait au bon Dieu*, ce qui peut laisser supposer que la différence faite au niveau de difficulté globale est liée à un autre aspect, non évalué dans cette expérience. Pour trouver quelques pistes de réponses, nous sommes revenue vers les constations faites dans les commentaires collectés lors de la constitution du corpus ESCAL. Parmi les 15 participants, 7 ont alors indiqué que la durée du document avait été une source de difficulté pour les extraits de plus d'une minute. Il s'avère que les deux extraits mentionnés ci-dessus, sont effectivement dans ce même cas

Pour les extraits de *La chèvre* et de *La gloire de mon père*, nous notons que quand le classement médian est plus élevé pour le cluster 2 (donc perçu comme globalement plus difficile que pour le cluster 1). Il y a également une différence significative dans le score médian relatif à la difficulté de percevoir la parole pour ce même cluster. Il est possible qu'une différence de sensibilité à l'intelligibilité de la parole explique la différence de perception de difficulté globale entre les deux clusters. Dans l'extrait du film *La Chèvrre*, l'élocution de Gérard Depardieu peut être une source de difficulté de compréhension orale. Dans le cadre de la première expérience, des extraits issus de deux films avec ce même acteur ont été évalués (*Cyrano de Bergerac* et *La Chèvre*). Sept participants ont indiqué dans leurs commentaires que son élocution pouvait rendre la compréhension orale compliquée (cf. section 3.4.5.4). Concernant l'extrait issu de *La gloire de mon père*, la perception de l'intelligibilité de la parole peut être influencée par le fort **accent provençal** de cet extrait. Onze participants sur 15 ayant noté dans les commentaires du corpus ESCAL que le fort accent du Sud pouvait nuire à la compréhension orale sur ce film (cf. section 3.4.5.4). Ces différents éléments peuvent nuire à la facilité à percevoir la parole et ont pu jouer un plus grand rôle dans les annotations attribuées par les participants du cluster 2.

Ces différents éléments, liés à l'intelligibilité (débit de parole, phrasé ou accent régional spécifique) peuvent être effectivement un problème pour les apprenants.

Une tendance mineure se manifeste avec l'extrait de *LOL*, pour lequel nous relevons une différence significative dans le score médian de la difficulté de la

grammaire. Les participants du cluster 1 ont considéré cet extrait plus compliqué d'un point de vue grammatical par rapport au cluster 2, ce qui peut expliquer la différence de classement médian.

Cette expérience et cette comparaison entre les clusters a permis de relever quelques pistes qui pourront être prises en compte dans l'outil d'aide à la didactisation. La durée des extraits proposés aux apprenants ainsi que leur intelligibilité pourra être prise en compte en fonction des niveaux.

Nous venons de comparer les comportements des participants à cette seconde expérimentation pour savoir ce qui pouvait motiver la différence de classement de certains extraits. Une dernière chose à faire est de voir comment notre système automatique se positionne par rapport aux comportements des annotateurs humains et voir de quel groupe il se rapproche le plus.

8.4 Système automatique : un participant comme un autre ?

Suite aux études réalisées dans les chapitres 6 et 7, nous avons proposé un modèle de régression linéaire pertinent pour la prédiction de la difficulté globale de documents audiovisuels. Si nous avons vu que l'approche interprétable permet d'obtenir de meilleurs résultats que l'approche neuronale, nous considérons que les valeurs de r et $RMSE$ ne sont pas suffisamment probantes pour qualifier notre modèle de « précis ». Mais, si le système automatique n'est pas capable de réaliser des prédictions identiques à celles d'experts humains, il est possible qu'il parvienne, comme eux, à déterminer si un extrait est plus compréhensible qu'un autre. C'est pour cela que nous avons voulu savoir comment notre système classait les extraits et s'il avait un comportement similaire à l'un des deux clusters identifiés précédemment.

TABLE 8.3 – Classements médians des extraits et scores automatiques de notre système de prédiction de la difficulté globale

Extrait	Cluster 1	Cluster 2	Système auto
Amélie Poulain	6,5	7,5	54,1
Cyrano de Bergerac	9,0	8,5	90
La cité de la peur	6,0	5,0	25,5
La folie des grandeurs	7,5	7,5	50,98
Les plages d'Agnès	2,5	2,0	46,23
Intouchables*	9,0	6,0	64,2
La chèvre*	2,0	7,0	37,2
LOL*	4,5	2,5	58,8
Qu'est-ce qu'on a fait au bon Dieu*	7,5	2,5	59,3
La gloire de mon père*	5,0	8,5	55,7

À partir du meilleur modèle *interprétable* pour la prédiction de la difficulté, nous calculons la difficulté globale des dix nouveaux documents audiovisuels présentés aux 22 enseignants de FLE. Nous extrayons les paramètres nécessaires au modèle avant de calculer la prédiction objective du niveau de compréhensibilité. Dans le tableau 8.3, nous présentons les scores automatiques qui ont été obtenus, confrontés avec les classements médians des deux clusters de participants (pour rappel le système auomatique produit un score sur 100).

TABLE 8.4 – Corrélations de Spearman entre les classements médians et les scores automatiques (* indique que la *pvalue* est inférieure à 0,05)

	Cluster 1 (8 participants)	Cluster 2 (14 participants)
Corrélation	**0,68***	0,23

Nous comparons le comportement humain avec les scores automatiques en réalisant une corrélation de Spearman par rangs avec les classements médians. Ces résultats sont présentés dans la table 8.4.En étudiant la valeur du coefficient de corrélation, nous pouvons identifier de quel cluster de participants notre système automatique se rapproche le plus. Il apparaît que le classement médian du cluster 1 est corrélé significativement avec les scores automatiques de difficulté globale avec une valeur de 0,68, ce qui reflète une **corrélation forte** entre les classements médians des participants et les scores automatiques. Cela signifie que notre système automatique réalise un classement des extraits de manière similaire aux participants du cluster 1.

Ce qui est très encourageant car nous pouvons prétendre qu'une mesure objective du niveau de compréhensibilité peut être obtenue et qu'elle s'apparente à ce qu'une évaluation subjective pourrait produire.

8.5 Bilan

Dans ce dernier chapitre nous avons voulu valider notre proposition dédiée à la mesure objective du niveau de compréhensibilité. À partir d'une des approches proposées dans cette seconde partie et notamment du modèle ayant obtenu des performances les plus significatives, nous avons voulu comparer le comportement de notre système automatique avec des mesures proposées par des experts. Nous avons donc mis en place une seconde expérience auprès d'enseignants de FLE.

Nous avons fait évaluer dix nouveaux extraits, dont deux issus de films inconnus du système automatique, car non présents dans le corpus ESCAL. Au total, 22 participants ont contribué à cette expérience en réalisant un classement des dix extraits par difficulté globale croissante et en leur attribuant également un score de difficulté au niveau de la grammaire, du vocabulaire et de l'intelligibilité. Deux clusters de participants se sont distingués à partir des classements médians de chacun des extraits. Nous avons analysé les différences entre les deux clusters mettant en avant une possible influence de la durée des extraits et de

possibles différences de notations en lien l'intelligibilité.

Le score de prédiction du niveau de compréhensibilité obtenu de manière objective pour chacun des extraits de ce nouveau jeu de données avec le modèle issu de l'approche *interprétable* a été comparé aux classements médians des extraits. La corrélation de Spearman avec un coefficient de corrélation de 0,68, nous permet de constater que notre système à un comportement proche d'un ensemble d'annotateurs humains, et ce, dans une tâche de classement des extraits les uns par rapport aux autres en termes de difficulté globale. L'expérience a donc permis dans un premier temps de valider la cohérence du modèle et de constater qu'il n'est pas biaisé par les films avec lesquels il avait été entraîné.

La mise en place de cette expérience nous permet également d'avoir un protocole qui sera utile dans le cas de recueil de nouvelles données : il permettra d'une part de récupérer de nouvelles annotations humaines par le biais de l'interface qui a été implémentée, et d'autre part de vérifier si le modèle est en accord avec le comportement humain. C'est en fonction des conclusions tirées qu'une stratégie d'amélioration du modèle pourra être proposée par la suite.

Conclusion générale et perspectives

Conclusion

Dans ce travail de thèse, nous avons exploré plusieurs approches permettant de produire une **mesure objective du niveau de compréhensibilité** de contenus audiovisuels, extraits de documents de fiction. L'objectif applicatif de ce travail réalisé dans le cadre d'une thèse CIFRE, est de fournir des outils d'aide à la didactisation de ce type de contenus. Les enseignants de langue ont des pratiques centrées autour de l'utilisation de ce type de documents, appelés aussi documents authentiques. Nous avons voulu contribuer aux développements d'outils d'aide adressé à ce public d'usagers pour favoriser l'apprentissage des langues étrangères. Ce travail de thèse s'est donc organisé autour de deux axes, l'un lié à la didactique des langues et plus particulièrement à l'étude de la compréhensibilité dans ce contexte, l'autre à la possibilité offerte par les traitements automatiques qu'il était possible d'appliquer pour atteindre notre objectif. Ceci se reflète dans les deux parties de ce manuscrit.

Niveau de compréhensibilité : quels éléments étudier ?

La première étape a été de comprendre quels étaient les différents phénomènes et facteurs influant sur le niveau de compréhensibilité des contenus et ce, en fonction des modalités à disposition (audio, texte et images). La compréhensibilité est un sujet vastement étudié dans le domaine de la didactique des langues étrangères : les professeurs de langue étrangère devant définir le niveau de compréhensibilité d'un contenu pour pouvoir le présenter à leurs apprenants. Une étude de la littérature dans les domaines de la didactique des langues a été réalisée. Ceci nous a permis de comprendre le rôle que jouent l'aspect linguistique, notamment les dimensions lexicales et grammaticales, ainsi que l'intelligibilité de la parole sur le niveau de compréhensibilité. Nous avons également constaté qu'en fonction les modalités considérées, le niveau de compréhensibilité varie, et que la complémentarité des modalités a généralement un effet positif.

Validation d'hypothèses avec le corpus ESCAL

Cette étude de la littérature nous a permis d'une part d'avoir une meilleure connaissance du domaine de la didactique des langues et d'aborder ce travail sous l'angle de la multidisciplinarité. Cela a permis d'autre part de structurer notre travail autour de l'étude de quatre dimensions : la complexité du vocabulaire, de la grammaire ainsi que l'intelligibilité de la parole, toutes trois en lien avec la difficulté globale, correspondant au niveau de compréhensibilité que nous nous proposons de mesurer automatiquement. Un autre angle d'analyse a consisté également à étudier l'influence des modalités sur le niveau de compréhensibilité de contenus audiovisuels. Pour aller plus loin dans notre étude, il était nécessaire de prouver cette influence et de la quantifier. En l'absence de corpus adéquat, nous avons constitué notre propre corpus, centré sur l'évaluation subjective du niveau de compréhensibilité. La phase de collecte d'annotation a été menée auprès de 15 enseignants de français langue étrangère (FLE) par le biais d'une interface développée spécifiquement.

Nous avons constitué un corpus de 55 extraits issus de 15 films français. Ce choix de l'origine des films était également motivé par des aspects socio-culturels et interactionnels qui sont importants dans les pratiques des enseignants de langue. Chacun des extraits, présentés sous différentes combinaisons de modalités (allant des modalités seules à l'ensemble des modalités) représente un ensemble de 275 documents qui ont été annotés par des experts. Le **corpus ESCAL** a ainsi été constitué est donc enrichi de différents niveaux d'annotations : il rassemble des évaluations subjectives et des commentaires formulés par les annotateurs et justifiant les évaluations faites au niveau de compréhensibilité, à la complexité lexicale, à la complexité grammaticale et à l'intelligibilité.

En combinant les évaluations subjectives de chacune des dimensions considérées dans un modèle de régression linéaire multiple, nous prouvons que **ces trois dimensions permettent d'expliquer 82% du niveau de compréhensibilité des contenus audiovisuels**. En étudiant l'évolution des évaluations en fonction des modalités disponibles, nous constatons qu'elles jouent également un rôle sur le niveau de compréhensibilité. **Cela confirme également que plus il y a de modalités, plus la compréhension des documents est facilitée** : les documents évalués comme les plus faciles à comprendre étant ceux où le signal audio, vidéo et la transcription de la parole étaient présentés simultanément.

À partir de la littérature et de l'étude du corpus, nous avons ensuite répertorié plus finement l'ensemble des facteurs qui ont un impact sur la complexité du vocabulaire, de la grammaticale, sur l'intelligibilité de la parole et pour finir, sur le niveau de compréhensibilité global.

Vers une mesure objective du niveau de compréhensibilité : approche interprétable ou approche neuronale ?

Notre objectif était d'obtenir une mesure objective du niveau de compréhensibilité des contenus audiovisuels en nous reposant sur le corpus ESCAL pour

les valeurs de référence. Nous avons alors exploré deux types d'approches : une approche « interprétable » et une approche « neuronale ».

Retour sur l'approche interprétable

La première approche repose sur l'extraction d'un ensemble de paramètres bien identifiés issus des modalités audio, vidéo et texte. Ces paramètres extraits en lien avec les facteurs identifiés dans la première partie, ont été exploités pour construire des modèles de régression linéaire multiple permettant de prédire les quatre dimensions centrales dans notre étude : complexité du vocabulaire, complexité de la grammaire, intelligibilité de la parole et le niveau de compréhensibilité.

Pour les trois premières dimensions, plusieurs manières de combiner les paramètres ont été testées : soit en se limitant aux paramètres en lien avec une dimension (par exemple : construire le modèle de régression pour la prédiction de la difficulté du vocabulaire avec les paramètres du niveau lexical), soit en intégrant en plus de ces paramètres d'autres paramètres qui ne sont *a priori* pas en lien avec la dimension considérée mais qui peuvent avoir affecté le jugement des annotateurs humains. Ainsi, nous avons constaté que des paramètres liés à la complexité lexicale sont entrés en jeu dans l'évaluation subjective de la complexité grammaticale.

Pour la prédiction du niveau de compréhensibilité à partir des 3 autres dimensions, trois stratégies de fusion ont été explorés :

— **la fusion tardive** qui applique des statistiques sur les prédictions des trois autres dimensions pour prédire le niveau de compréhensibilité,
— **la fusion intermédiaire** qui conserve les paramètres gardés après une phase de sélection pour construire les modèles de prédiction des trois autres dimensions et ajoute les paramètres en lien avec le niveau de compréhensibilité pour créer un modèle de régression linéaire multiple,
— **la fusion précoce** qui exploite la totalité des paramètres extraits pour construire un modèle global.

En comparant les erreurs et les corrélations entre valeurs prédites et valeurs de référence issues du corpus ESCAL, en fonction des stratégies de fusions, nous avons conclu que c'est la **fusion intermédiaire** qui propose le modèle le plus satisfaisant pour prédire le niveau de compréhensibilité.

Retour sur l'approche neuronale

L'approche «interprétable» repose sur le contrôle des paramètres qui alimentent les modèles de régression linéaire et sur l'obtention de modèles dont les décisions peuvent être facilement expliquées et donc comprises à termes par les utilisateurs des futurs outils. Nous voulions comparer cette approche avec une approche neuronale où aucun contrôle n'est réalisé sur des paramètres alimentant des modèles qui ne sont eux-mêmes pas interprétables. L'approche neuronale repose sur l'exploitation de trois réseaux de neurones profonds choisis pour leur pertinence par rapport à notre problématique de prédiction du niveau de compréhensibilité : CamemBERT pré-entraîné pour des tâches de prédiction de

mots spécialisé pour la langue française, PASE+ pré-entraîné pour la reconnaissance de la parole dans des environnements bruités et ResNet 3D, pré-entraîné pour la reconnaissance d'actions dans un flux vidéo.

Ces trois réseaux permettent d'extraire des représentations pour chacune des modalités, et ce sont ces représentations qui sont utilisées pour alimenter divers modèles de régression : les SVM adaptés à la régression, ou SVR, qui ont permis d'obtenir les meilleurs résultats, les k plus proches voisins, les forêts d'arbres décisionnels et les perceptrons multicouches.

Approche «interprétable» ou approche «neuronale» ?

En comparant les divers modèles issus de l'approche neuronale et de l'approche interprétable, nous avons fait le constat les modèles issus de la première approche permettent de minimiser les erreurs et de maximiser les corrélations entre valeurs prédites et les valeurs de référence issues du corpus ESCAL. Une approche où nous connaissons et contrôlons les paramètres candidats pour construire nos modèles se montre plus efficace que d'utiliser des représentations issues de réseaux de neurones. De plus elle permet à terme de fournir un feedback pertinent vers les usagers en leur permettant ainsi de choisir de manière plus éclairée les supports pour leur enseignement de langue étrangère.

L'approche *interprétable* comporte en plus de nombreux avantages, notamment la mise en place d'une **relation mathématique** compréhensible entre les paramètres et les prédictions qui permet notamment d'**expliciter quels sont les paramètres qui ont le plus de poids** sur la dimension étudiée, la **reproductibilité** et l'**interprétabilité** des prédictions de part la connaissance des paramètres entrant en jeu dans les modèles de régression linéaire multiple.

Système automatique VS comportement humain

Nous avons exploré et comparé les approches *inteprétable* et *neuronale*, nous souhaitions valider notre modèle de régression issu de l'approche interprétable dans un contexte d'utilisation. Nous avons conçu une nouvelle expérience basée sur une tâche de classement en fonction de la difficulté globale perçue et sur une tâche d'évaluation des trois autres dimensions. Un jeu de dix extraits autres que ceux présents dans ESCAL ont été évalués par 22 professeurs de FLE.

En appliquant une méthode de partionnement des données (ou *clustering*) sur les classements réalisés nous avons constaté que les participants se séparaient en deux groupes. L'analyse des classements médians de chaque extrait permet d'observer que certains classements diffèrent en fonction des groupes. Le retour vers les commentaires des annotateurs permet d'envisager des pistes d'analyses complémentaires en lien avec la durée des extraits et l'intelligibilité de la parole. Ceci resterait à confirmer par une expérimentation de plus grande ampleur.

Pour finaliser notre étude nous nous sommes interrogée sur le comportement que pouvait avoir notre système automatique de prédiction du niveau de compréhensibilité en comparaison aux annotateurs humains.

Pour chacun des dix nouveaux extraits, nous avons prédit le niveau de compréhensibilité à partir de notre meilleur modèle obtenu avec l'approche inter-

prétable en réalisant une corrélation de Spearman entre le classement médian des extraits réalisé par chacun des clusters. Nous constatons que le système automatique a un comportement proche de celui des participants du groupe 1 et ce, pour la tâche de classement des extraits les uns par rapport aux autres en fonction de leur niveau de compréhensibilité.

Plusieurs contributions ont été apportées durant ce travail de thèse :

— une étude poussée de la littérature en didactique des langues concernant la notion de compréhensibilité et les différentes façons de l'évaluer ;
— un bilan des différents facteurs qui influent sur la compréhensibilité et de là une liste de 57 paramètres à extraire des contenus décomposés en différentes modalités ;
— l'étude et la comparaison de deux approches de traitement automatique, l'une centrée sur les paramètres et motivée par son caractère interprétable et l'autre basée sur des approches récentes autour des réseaux de neurones profonds ;
— la validation de l'approche interprétable sur un nouveau jeu de données. Si les métriques obtenues pour le modèle testé (r=*0,38* et *RMSE*=*17,88*) ne sont pas optimales, cette dernière expérience montre que la solution que nous proposons est capable de reproduire le comportement d'un groupe d'annotateurs humains quand il s'agit de classer des documents audiovisuels les uns par rapport aux autres en termes de niveau de compréhensibilité. Ceci valide l'ensemble de notre démarche.

Perspectives

Vers une mesure objective interprétable

À l'issue de nos travaux sur la recherche d'une mesure objective du niveau de compréhensibilité, nous avons abouti à un modèle de régression linéaire multiple. Si le choix de ce type de modèle permet de mettre en avant les paramètres qui sont entrés en jeu dans la prédiction, se pose la question de son interprétation. La mesure fournie (comme on peut le voir dans la table 8.3) est une valeur numérique comprise entre 0 et 100. Il n'est pas certain qu'un utilisateur soit capable de donner un sens à la mesure si elle est présentée sous cette forme. Quand nous regardons les diverses annotations existantes dans le domaine de la didactique des langues étrangères, en termes de niveau de difficulté de différents types de documents, elles reposent généralement sur l'association d'une catégorie à un niveau et non pas une valeur numérique précise. C'est le cas du référentiel commun [Conseil de l'Europe, 2003] qui utilise les niveaux A1, A2, B1, B2, C1 et C2 pour différencier les débutants (A), des apprenants intermédiaires (B) ou experts (C). Pour ces raisons, s'orienter vers un système de notation du niveau de compréhensibilité des documents ou extraits audiovisuels **sur la base de catégories** semble plus pertinent. Cela fait écho à une manière de noter la difficulté **connue** des enseignants de FLE et sera plus simple à comprendre et

à exploiter qu'une valeur numérique qui ne serait pas suffisamment porteuse de sens. Aux vues de la marge d'erreur de notre modèle, un découpage aussi fin que le niveau CECRL n'est pas envisageable. Nous pouvons cependant proposer un découpage en trois niveaux : facile, moyen et difficile. Les potentielles erreurs de classement de nos documents audiovisuels se situeront sur des documents pour lesquels le niveau de compréhensibilité prédit sera à la frontière de deux catégories (par exemple : facile-moyen ou moyen-difficile).

Amélioration de nos modèles

La création des modèles présentés dans ce manuscrit reposait sur le corpus ESCAL présenté dans le chapitre 3. Il s'agit d'un corpus de 1h20, basé sur 55 extraits de films. Si les résultats obtenus nous ont amené à conclure que notre approche ainsi que nos résultats étaient pertinents, la quantité de données sur laquelle nous nous sommes basé reste faible.

L'enrichissement du corpus à l'aide de nouveaux extraits pourrait donner lieu d'une part à un renforcement de nos modèles déjà obtenus pour l'approche interprétable et d'autre part à des améliorations des modèles de l'approche neuronale.

L'utilisation de réseaux de neurones plus récents tels que DistilCamemBERT [Delestre and Amar, 2022] ou HuBERT [Hsu et al., 2021] pour la modalité texte, et Wav2Vec 2.0 [Baevski et al., 2020] pour la modalité audio pourrait également amener un changement dans les résultats obtenus pour l'approche neuronale.

Intégration dans une plateforme dédiée aux enseignants

Transcription manuelle, transcription automatique ou sous-titres ?

Pour nous placer dans le cas idéal, nous avons fait le choix de travailler avec la transcription exacte de la parole tout le long de cette thèse. Mais, dans les faits, obtenir la transcription exacte de la parole peut s'avérer compliqué. Il est possible de réaliser la transcription manuellement ou d'utiliser un outil de transcription automatique (par exemple celui de Google ou de l'équipe SAMoVA[2]). Une autre alternative serait d'utiliser directement les sous-titres fournis avec un contenu audiovisuel. Les sous-titres peuvent être professionnels et suivre des normes précises en fonction de l'éditeur, réalisés par des amateurs ou encore générés automatiquement, comme c'est parfois le cas avec Youtube. Cependant, la qualité de la transcription automatique va être fonction de l'environnement sonore bruité ou réverbéré, et une transcription manuelle, en plus de demander du temps, peut aussi contenir des erreurs. En ce qui concerne les sous-titres, il s'agit rarement de la transcription exacte de la parole et il est également possible de trouver un nombre d'erreurs plus ou moins important, soit à cause d'une erreur humaine (mauvaise orthographe, syntaxe...) surtout s'il s'agit de sous-titres amateurs, soit à cause d'une erreur du système automatique.

Néanmoins, même si les transcriptions ou les sous-titres contiennent des erreurs, nous pouvons nous demander si les utiliser pour construire un modèle de prédiction du niveau de compréhensibilité dégraderait fortement la qualité des prédictions. Une expérience intéressante serait de construire de nouveaux modèles en exploitant non plus les transcriptions exactes obtenues manuellement, mais des transcriptions automatiques et des sous-titres réels. Ensuite, nous pourrions comparer les prédictions de ces modèles avec notre modèle actuel et la vérité terrain du corpus ESCAL. Dans le cas où les performances seraient similaires, nous pourrions envisager de substituer la transcription exacte par les sous-titres ou une transcription automatique dans notre solution de prédiction du niveau de compréhensibilité. Cela permettra ainsi de s'affranchir d'un travail en amont de transcription manuelle.

Prédiction du niveau de compréhensibilité pour d'autres langues : vers une démarche généralisable ?

Pour atteindre notre objectif de construction de mesure objective du niveau de compréhensibilité, nous sommes passée par plusieurs étapes : la compréhension des phénomènes et facteurs qui affectent la compréhensibilité, l'association de paramètres calculables à ces phénomènes et facteurs, la mise en place d'un corpus de référence (corpus ESCAL), l'identification d'outils automatiques et de ressources disponibles pour calculer ces paramètres, la construction de modèles avec une approche *interprétable* et *neuronale* et la confrontation des modèles avec le comportement humain. Si nous souhaitions construire une mesure objec-

2. https://paty.irit.fr/demo

tive du niveau de compréhensibilité pour un autre langage, nous pouvons nous demander s'il nous serait possible d'appliquer la même démarche.

La phase d'identification des phénomènes, facteurs et paramètres associés pourrait sembler redondante si l'on part de l'hypothèse que pour tous les langages le vocabulaire, la grammaire et l'intelligibilité seront systématiquement les principaux phénomènes qui impacteront la compréhensibilité. Chaque langue ayant néanmoins ses spécificités (par exemple l'absence d'accord des adjectifs au féminin ou au masculin en anglais), il est tout à fait possible que des études antérieures sur un langage donné aient permis de dégager des phénomènes que nous n'avons pas considéré pour le français. Il est également possible qu'en fonction du langage un des phénomènes soit une source moindre de difficulté et puisse être négligé.

En ce qui concerne la mise en place d'un corpus, s'il est inexistant dans la langue donnée, il est évident que la sélection de contenus et le recrutement d'experts pour la collecte de nouvelles annotations représentera un coût important. Il y a cependant des points forts dans notre démarche. Le premier est que la même interface de collecte d'annotation peut aisément être utilisée dans sa structure globale. Le second est que le corpus de sous-titres OPUS qui a servi de base pour réaliser la transcription exacte des documents audiovisuels est disponible dans d'autres langages. Cependant, il faudra faire appel à un natif de la langue pour réaliser la transcription. Il est néanmoins possible que dans d'autres langages, les outils de transcription automatique de la parole soient suffisamment performants pour que la transcription exacte soit obtenue automatiquement avec des corrections mineures à réaliser par un natif.

Il est possible que la phase d'identification d'outils automatiques et de ressources disponibles pour calculer ces paramètres mette en avant l'absence d'outils et de ressources. Par exemple, pour la connaissance des mots rares mots et fréquents, existe-t-il un équivalent à la liste de Brunet ou à Lexique 3 ? Il faudra réfléchir alors à l'aspect bloquant que cela représente et aux alternatives possibles : constituer nous-mêmes un outil ad-hoc, se dispenser de l'évaluation d'un phénomène en évaluant les conséquences que cela pourrait avoir sur la qualité des modèles...

Au premier abord, il semble que le passage d'un langage à un autre amène des contraintes inhérentes au langage considéré et aux avancées des études linguistiques menées et outils développés pour le langage. Le fait de reprendre l'ensemble de la démarche utilisée dans cette thèse sur plusieurs langages permettrait cependant, à terme, de répondre à des nouvelles questions telles que :

— les phénomènes qui influencent la compréhensibilité sont-ils identiques, peu importe le langage ?
— est-il possible de regrouper les langages en plusieurs grandes catégories, en fonction des phénomènes qui influencent leur niveau de compréhensibilité ?

— les modèles interprétables sont-ils toujours plus « performants » que les modèles neuronaux pour la prédiction du niveau de compréhensibilité ?

Démonstrateur actuel

Plus loin dans l'étude de la compréhensibilité

En 2019, un laboratoire commun entre l'entreprise Archean Technologies et l'IRIT a démarré. Ce LabCom appelé **ALAIA**[3] pour *Apprentissage des Langues Assisté par Intelligence Artificielle*[4] est financé par l'ANR (Agence Nationale de la Recherche). Cette collaboration vise à développer des services destinés aux apprenants pour un apprentissage en autonomie comme aux enseignants pour un apprentissage guidé (via des exercices proposés par les enseignants). ALAIA fait particulièrement le focus sur le travail des compétences orales des apprenants et se base sur le lien entre la langue maternelle des apprenants (L1) et la langue apprise ou langue cible (L2). La méthodologie établie doit permettre de passer ensuite à d'autres paires de langues L1/L2. Il s'agit d'analyser les productions des apprenants pour évaluer leurs compétences à différents niveaux (prononciation, choix du bon vocabulaire en structure de la phrase produite, discours) et mesurer la capacité de leur production à être comprise.

L'ensemble du travail entrepris pendant cette thèse pour prédire le niveau de compréhensibilité de contenus audiovisuels authentiques peut être réinvesti dans ces nouvelles études. Bien que le point de vue diffère, en s'intéressant directement aux apprenants, les paramètres identifiés au cours de l'étude menée dans le cadre de cette thèse peuvent être réexploités dans ALAIA pour estimer la production orale des apprenants.

3. https://www.irit.fr/SAMOVA/site/projects/current/labcom-alaia/
4. https://anr.fr/Projet-ANR-18-LCV3-0001

Bibliographie

[Abeillé et al., 2003] Abeillé, A., Clément, L., and Toussenel, F. (2003). Building a treebank for french. In *Treebanks*, pages 165–187.

[Adank et al., 2009] Adank, P., Evans, B. G., Stuart-Smith, J., and al (2009). Comprehension of familiar and unfamiliar native accents under adverse listening conditions. *Journal of Experimental Psychology : Human Perception and Performance*, 35(2) :520.

[Aiken et al., 2003] Aiken, L. S., West, S. G., and Pitts, S. C. (2003). Multiple linear regression. *Handbook of psychology*, pages 481–507.

[Akinci, 2005] Akinci, M.-A. (2005). La complexité syntaxique dans les textes écrits en français : étude chez des bilingues et monolingues. In *Papier présenté au colloque «Typologie et modélisation de la coordination et de la subordination»(26-28 mai 2005)[http ://www. cavi. univ-paris3. fr/ilpga/colloque-coord-subord-2005/pre-textes/Akinci. pdf].*

[Albawi et al., 2017] Albawi, S., Mohammed, T. A., and Al-Zawi, S. (2017). Understanding of a convolutional neural network. In *2017 International Conference on Engineering and Technology (ICET)*, pages 1–6. Ieee.

[Ali Batel, 2014] Ali Batel, E. (2014). The Effectiveness of Video vs. Written Text in English Comprehension and Acquisition of ESL Students. *Arab World English Journal*, 5(4).

[Anderson, 2005] Anderson, N. (2005). L2 strategy research. *Handbook of research in second language teaching and learning.*

[André-Obrecht and Jacob, 1997] André-Obrecht, R. and Jacob, B. (1997). Direct identification vs. correlated models to process acoustic and articulatory informations in automatic speech recognition. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 999–1002. IEEE.

[Aneiro, 1990] Aneiro, S. M. (1990). The influence of receiver apprehension in foreign language learners on listening comprehension among puerto rican college students. Unpublished doctoral dissertation.

[Artusi et al., 2002] Artusi, R., Verderio, P., and Marubini, E. (2002). Bravais-pearson and spearman correlation coefficients : meaning, test of hypothesis and confidence interval. *The International journal of biological markers*, 17(2) :148–151.

[Aslim-Yetis, 2010] Aslim-Yetis, V. (2010). Le document authentique : un exemple d'exploitation en classe de FLE. *Synergies Canada*, (2).

[Assaad, 2005] Assaad, M. (2005). *Le rôle culturel de la publicité dans l'enseignement/apprentissage du français langue étrangère*. PhD thesis, Rennes 2.

[Ausloos, 2008] Ausloos, M. (2008). Equilibrium and dynamic methods when comparing an english text and its esperanto translation. *Physica A : Statistical Mechanics and its Applications*, 387(25) :6411–6420.

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

[Beacco et al., 2004] Beacco, J.-C., Bouquet, S., and Porquier, R. (2004). *Niveau B2 pour le français, un référentiel*. Didier.

[Beacco and Porquier, 2007] Beacco, J.-C. and Porquier, R. (2007). *Niveau A1 pour le français, un référentiel*. Didier.

[Berry et al., 1985] Berry, W. D., Feldman, S., and Stanley Feldman, D. (1985). *Multiple regression in practice (No. 50)*. Sage.

[Blache, 2010] Blache, P. (2010). Un modèle de caractérisation de la complexité syntaxique. In *Traitement Automatique des Langues Naturelles*, pages 1–10.

[Blau, 1991] Blau, E. K. (1991). More on comprehensible input : The effect of pauses and hesitation markers on listening comprehension. Unpublished paper presented at the Annual Meeting of the Puerto Rico Teachers of English to Speakers of Other Languages.

[Bloomfield et al., 2011] Bloomfield, A., Wayland, S., Blodgett, A., and et al (2011). Factors related to passage length : Implications for second language listening comprehension. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33.

[Bloomfield et al., 2010] Bloomfield, A., Wayland, S. C., Rhoades, E., and et al (2010). *What makes listening difficult ? Factors affecting second language listening comprehension*. MARYLAND UNIV COLLEGE PARK.

[Bloomfield, 1942] Bloomfield, L. (1942). *Outline guide for the practical study of foreign languages*. Linguistic Society of America.

[Bourque, 1989] Bourque, G. (1989). Des mesures de lisibilité. Communication présentée au 57e Congrès de l'ACFAS.

[Boyle, 1984] Boyle, J. P. (1984). . factors affecting listening comprehension. *ELT journal*, 38(1) :34–38.

[Bradski and Kaehler, 2000] Bradski, G. and Kaehler, A. (2000). Opencv. *Dr. Dobb's journal of software tools*, 3.

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine learning*, 45(1) :5–32.

[Brindley and Wigglesworth, 1997] Brindley, G. and Wigglesworth, G., editors (1997). *Access : Issues in language test design and delivery*. National Centre for English Language Teaching and Research, Macquarie University.

[Buck, 2001] Buck, G. (2001). *Assessing listening*. Cambridge University Press.

[Bulté and Housen, 2012] Bulté, B. and Housen, A. (2012). Defining and operationalising l2 complexity. *Dimensions of L2 performance and proficiency : Complexity, accuracy and fluency in SLA*, pages 23–46.

[Byrnes and Sinicrope, 2008] Byrnes, H. and Sinicrope, C. (2008). 7 advancedness and the development of relativization in L2 German. *The longitudinal study of advanced L2 capacities*, pages 109–138.

[Camiciottoli, 2004] Camiciottoli, B. C. (2004). Interactive discourse structuring in L2 guest lectures : Some insights from a comparative corpus-based study. *Journal of English for Academic Purposes*, 3(1) :39–54.

[Camus, 1999] Camus, O. (1999). *Les interactions langagières*.

[Carlisle and Goodwin, 2014] Carlisle, J. F. and Goodwin, A. P. (2014). Handbook of language and literacy : Development and disorders (2e éd.). pages 265–282.

[Carrow-Woolfolk, 1999] Carrow-Woolfolk, E. (1999). *CASL : Comprehensive assessment of spoken language*. American Guidance Services.

[Chabert, 1989] Chabert, J.-L. (1989). Gauss et la méthode des moindres carrés. *Revue d'histoire des sciences*, pages 5–26.

[Chandrashekar and Sahin, 2014] Chandrashekar, G. and Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1) :16–28.

[Chang and Read, 2008] Chang, A. C.-S. and Read, J. (2008). Reducing listening test anxiety through various forms of listening support. *TESL-EJ*, 12(1) :n1.

[Chaudron, 1983] Chaudron, C. (1983). Simplification of input : Topic reinstatements and their effects on l2 learners' recognition and recall. *TESOL quarterly*, 17(3) :437–458.

[Chaudron and Richard, 1986] Chaudron, C. and Richard, J. C. (1986). "the effects of Discourse Markers on the Comprehension of Lectures. *Applied Linguistics*, 7 :113–27.

[Chauvet, 2008] Chauvet, A. (2008). *Référentiel pour le Cadre européen commun A1-A2-B1-B2-C1-C2*. Alliance Française/CLE international.

[Chok, 2010] Chok, N. S. (2010). *Pearson's versus Spearman's and Kendall's correlation coefficients for continuous data*. PhD thesis, University of Pittsburgh.

[Collins-Thompson and Callan, 2005] Collins-Thompson, K. and Callan, J. (2005). Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology*, 56(13) :1448–1462.

[Colé et al., 2004] Colé, P., Royer, C., Leuwers, C., and et al (2004). Les connaissances morphologiques dérivationnelles et l'apprentissage de la lecture chez l'apprenti-lecteur français du cp au ce2. *L'année psychologique*, 104(4) :701–750.

[Conseil de l'Europe, 2003] Conseil de l'Europe (2003). *Cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer*. Council of Europe.

[Corley and Hartsuiker, 2003] Corley, M. and Hartsuiker, R. J. (2003). Hesitation in speech can... um... help a listener understand. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

[Coste, 1970] Coste, D. (1970). Le renouvellement méthodologique dans l'enseignement du français langue étrangère : remarques sur les années 1955-1970. *Langue française*, (8) :7–23.

[Coste and et al, 1976] Coste, D. and et al (1976). Un niveau-seuil (A Threshold Level).

[Cunningham, 2008] Cunningham, P. (2008). Dimension reduction. In *Machine learning techniques for multimedia*, pages 91–112. Springer.

[Cuq et al., 2003] Cuq, J.-P. et al. (2003). Dictionnaire de didactique du français. *Paris : CLE international*, pages 214–216.

[Dahl and Ludvigsen, 2014] Dahl, T. I. and Ludvigsen, S. (2014). How i see what you're saying : The role of gestures in native and foreign language listening comprehension. *The Modern Language Journal*, 98(3) :813–833.

[Dale and Chall, 1948] Dale, E. and Chall, J. S. (1948). A formula for predicting readability : Instructions. *Educational research bulletin*, pages 37–54.

[Dale and Chall, 1949] Dale, E. and Chall, J. S. (1949). The concept of readability. *Elementary English*, 26(1) :19–26.

[Daneman and Merikle, 1996] Daneman, M. and Merikle, P. M. (1996). Working memory and language comprehension : A meta-analysis. *Psychonomic bulletin & review*, 3(4) :422–433.

[De Clercq, 2016] De Clercq, B. (2016). Le développement de la complexité syntaxique en français langue seconde : complexité structurelle et diversité. In *SHS Web of Conferences*, volume 27, page 07006. EDP Sciences.

[de Clercq, 2016] de Clercq, B. (2016). Le développement de la complexité syntaxique en français langue seconde : complexité structurelle et diversité. In *SHS Web of Conferences*, volume 27, page 07006.

[De Jong and Wempe, 2009] De Jong, N. H. and Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior research methods*, 41(2) :385–390.

[Delestre and Amar, 2022] Delestre, C. and Amar, A. (2022). Distilcamembert : une distillation du modèle français camembert. In *CAp (Conférence sur l'Apprentissage automatique)*.

[Denis and Varenne, 2019] Denis, C. and Varenne, F. (2019). Interprétabilité et explicabilité pour l'apprentissage machine : entre modèles descriptifs, modèles prédictifs et modèles causaux. une nécessaire clarification épistémologique. In *National (French) Conference on Artificial Intelligence (CNIA)-Artificial Intelligence Platform (PFIA)*, pages 60–68.

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

[Diedenhofen and Musch, 2015] Diedenhofen, B. and Musch, J. (2015). cocor : A comprehensive solution for the statistical comparison of correlations. *PloS one*, 10(4) :e0121945.

[Diependaele et al., 2012] Diependaele, K., Brysbaert, M., and Neri, P. (2012). How noisy is lexical decision ? *Frontiers in psychology*, 3 :348.

[DJEFFAL, 2012] DJEFFAL, A. (2012). *Utilisation des méthodes Support Vector Machine (SVM) dans l'analyse des bases de données*. PhD thesis, Université Mohamed Khider-Biskra.

[Donath et al., 2012] Donath, C., Gräßel, E., Baier, D., and et al (2012). Predictors of binge drinking in adolescents : ultimate and distal factors-a representative study. *BMC public health*, 12(1) :1–15.

[Dormann et al., 2013] Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J., et al. (2013). Collinearity : a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36(1) :27–46.

[Doukhan et al., 2018] Doukhan, D., Lechapt, E., Evrard, M., and et al (2018). Ina's mirex 2018 music and speech detection system. *Music Information Retrieval Evaluation eXchange (MIREX 2018)*.

[Dubey et al., 2019] Dubey, S., Boragule, A., and Jeon, M. (2019). 3d resnet with ranking loss function for abnormal activity detection in videos. In *2019 International Conference on Control, Automation and Information Sciences (ICCAIS)*, pages 1–6. IEEE.

[Dufour and Frauenfelder, 2007] Dufour, S. and Frauenfelder, U. H. (2007). L'activation et la sélection lexicales lors de la reconnaissance des mots parlés : modèles théoriques et données expérimentales. *L'Année psychologique*, 107(1) :87–111.

[Dufour et al., 2002] Dufour, S., Peereman, R., Pallier, C., and Radeau, M. (2002). Vocolex : Une base de données lexicales sur les similarités phonologiques entre les mots français. *Année psychologique*, 102(4) :725–745.

[Efroymson, 1960] Efroymson, M. (1960). Multiple regression analysis. *Mathematical methods for digital computers*, pages 191–203.

[Elkhafaifi, 2005] Elkhafaifi, H. (2005). Listening comprehension and anxiety in the arabic language classroom. *The modern language journal*, 89 :206–220.

[Ephraim and Malah, 1985] Ephraim, Y. and Malah, D. (1985). Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE transactions on acoustics, speech, and signal processing*, 33(2) :443–445.

[Falk et al., 2014] Falk, I., Bernhard, D., Gérard, C., and et al. (2014). Étiquetage morpho-syntaxique pour des mots nouveaux. In *21ème conférence sur le Traitement Automatique des Langues Naturelles*, page 431.

[Falk et al., 2010] Falk, T. H., Zheng, C., and Chan, W.-Y. (2010). A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(7) :1766–1774.

[Farinas et al., 2005] Farinas, J., Rouas, J.-L., Pellegrino, F., and André-Obrecht, R. (2005). 2-extraction automatique de paramètres prosodiques pour l'identification automatique des langues. *traitement du signal*.

[Ferreira, 1991] Ferreira, F. (1991). Effects of length and syntactic complexity on initiation times for prepared utterances. *Journal of Memory and Language*, 30(2) :210–233.

[Flesch, 1948] Flesch, R. (1948). A new readability yardstick. *Journal of applied psychology*, 32(3) :221.

[Florit et al., 2013] Florit, E., Roch, M., and Levorato, M. C. (2013). The relationship between listening comprehension of text and sentences in preschoolers : Specific or mediated by lower and higher level components. *Applied Psycholinguistics*, 34(2) :395–415.

[Française and Chauvet, 2008] Française, A. and Chauvet, A. (2008). *Référentiel de l'Alliance française pour le Cadre européen commun*.

[François, 2009] François, T. (2009). Modèles statistiques pour l'estimation automatique de la difficulté de textes de fle. In *Actes de RECITAL*.

[François and Fairon, 2012] François, T. and Fairon, C. (2012). An « AI readability » formula for french as a foreign language. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 466–477.

[François et al., 2014] François, T., Gala, N., Watrin, P., and et al. (2014). FLElex : a graded lexical resource for French foreign learners. In *International conference on Language Resources and Evaluation (LREC 2014)*.

[Gala et al., 2014] Gala, N., François, T., Bernhard, D., and et al (2014). Un modèle pour prédire la complexité lexicale et graduer les mots. In *TALN'2014*, pages 91–102.

[Gelin et al., 2021] Gelin, L., Daniel, M., Pinquier, J., and Pellegrini, T. (2021). End-to-end acoustic modelling for phone recognition of young readers. *arXiv preprint arXiv :2103.02899*.

[Ghio et al., 2016] Ghio, A., Giusti, L., Blanc, E., and et al (2016). Quels tests d'intelligibilité pour évaluer les troubles de production de la parole ? In *Journées d'Etude sur la Parole*, pages 589–596.

[Goel et al., 2017] Goel, E., Abhilasha, E., Goel, E., and Abhilasha, E. (2017). Random forest : A review. *International Journal of Advanced Research in Computer Science and Software Engineering*, 7(1).

[Goh, 1999] Goh, C. (1999). How much do learners know about the factors that influence their listening comprehension ? *Hong Kong Journal of Applied Linguistics*, 4(1) :17–42.

[Goldberg, 2017] Goldberg, Y. (2017). Neural network methods for natural language processing. *Synthesis lectures on human language technologies*, 10(1) :1–309.

[Goldin-Meadow and Alibali, 2013] Goldin-Meadow, S. and Alibali, M. W. (2013). Gesture's role in speaking, learning, and creating language. *Annual review of psychology*, 64 :257–283.

[Gougenheim et al., 1964] Gougenheim, G., Michea, R., Rivenc, P., and Sauvageot, A. (1964). *L'élaboration du français fondamental : étude sur l'établissement d'un vocabulaire et d'une grammaire de base*. Didier.

[Granfeldt, 2006] Granfeldt, J. (2006). Evaluation du niveau lexical et grammatical à l'écrit en français langue étrangère : l'apport des analyses automatiques. *Revue française de linguistique appliquée*, 11(1) :103–117.

[Guiraud, 1954] Guiraud, P. (1954). Stylistiques. *Neophilologus*, 38(1) :1–11.

[Guyon et al., 2002] Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1) :389–422.

[Hall, 1999] Hall, M. A. (1999). Correlation-based feature selection for machine learning.

[Hara et al., 2017] Hara, K., Kataoka, H., and Satoh, Y. (2017). Learning spatio-temporal features with 3d residual networks for action recognition. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 3154–3160.

[Harmer, 2007] Harmer, J. (2007). *The practice of English language teaching*. Pearson Longman.

[Harrell Jr et al., 1984] Harrell Jr, F. E., Lee, K. L., Califf, R. M., Pryor, D. B., and Rosati, R. A. (1984). Regression modelling strategies for improved prognostic prediction. *Statistics in medicine*, 3(2) :143–152.

[Hayati, 2010] Hayati, A. (2010). The effect of speech rate on listening comprehension of EFL learners. *Creative Education*, 1(2) :107.

[Hayati and Mohmedi, 2011] Hayati, A. and Mohmedi, F. (2011). The effect of films with and without subtitles on listening comprehension of EFL learners. *British Journal of Educational Technology*, 42(1) :181–192.

[Hearst et al., 1998a] Hearst, M. A., Dumais, S. T., Edgar, O., and et al (1998a). Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4) :18–28.

[Hearst et al., 1998b] Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., and Scholkopf, B. (1998b). Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4) :18–28.

[Hedaywa and Sourak, 2013] Hedaywa, J. and Sourak, S. (2013). Le rôle des documents authentiques dans l'enseignement /apprentissage du français langue étrangère. *Arts and Humanities Series*, 35(2).

[Henry, 1975] Henry, G. (1975). *Comment Mesurer La Lisibilite (How to Measure Readability)*.

[Henry, 1980] Henry, G. (1980). Lisibilité et compréhension. *Communication & langages*, 45(1) :7–16.

[Hocking, 1976] Hocking, R. R. (1976). A Biometrics invited paper. The analysis and selection of variables in linear regression. *Biometrics*, pages 1–49.

[Hollard, 2010] Hollard, S. (2010). Interprétation de textes polysémiques : une étude expérimentale appuyée sur l'oculométrie. *Glossa*, (109) :16–41.

[Hsu et al., 2021] Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. (2021). Hubert : Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29 :3451–3460.

[Hunt, 1965] Hunt, K. W. (1965). A synopsis of clause-to-sentence length factors. *The English Journal*, 54(4) :300–309.

[Ikeno and Hansen, 2006] Ikeno, A. and Hansen, J. H. (2006). Perceptual recognition cues in native English accent variation : "listener accent, perceived accent, and comprehension". In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 1, pages I–I.

[Ioannidou et al., 2019] Ioannidou, A., Chatzilari, E., Nikolopoulos, S., and Kompatsiaris, I. (2019). 3d resnets for 3d object classification. In *International Conference on Multimedia Modeling*, pages 495–506. Springer.

[Ivan et al., 2006] Ivan, M. et al. (2006). La méthode structuro-globale audio-visuelle (sgav). *Dialogos*, 7(14) :16–21.

[Ji et al., 2012] Ji, S., Xu, W., Yang, M., and Yu, K. (2012). 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1) :221–231.

[Johansson, 2008] Johansson, V. (2008). Lexical diversity and lexical density in speech and writing : A developmental perspective. *Working papers/Lund University, Department of Linguistics and Phonetics*, 53 :61–79.

[Johnson, 2006] Johnson, D. H. (2006). Signal-to-noise ratio. *Scholarpedia*, 1(12) :2088.

[Jones and Plass, 2002] Jones, L. C. and Plass, J. L. (2002). Supporting listening comprehension and vocabulary acquisition in french with multimedia annotations. *The modern language journal*, 86(4) :546–561.

[Kahane, 2001] Kahane, S. (2001). Grammaires de dépendance formelles et théorie sens-texte. *TALN 2001*.

[Kalyuga et al., 1999] Kalyuga, S., Chandler, P., and John, S. (1999). Managing split-attention and redundancy in multimedia instruction. *Applied Cognitive Psychology : The Official Journal of the Society for Applied Research in Memory and Cognition*, 13(4) :351–371.

[Kashioka, 1990] Kashioka, T. (1990). Les systèmes des temps verbaux français et japonais. *L'information grammaticale*, 47(1) :30–33.

[Kellerman, 1992] Kellerman, S. (1992). 'I see what you mean' : The role of kinesic behaviour in listening and implications for foreign and second language learning. *Applied linguistics*, 13(3) :239–258.

[Kennedy and Trofimovich, 2008] Kennedy, S. and 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.

[Kennedy-Shaffer, 2019] Kennedy-Shaffer, L. (2019). Before p< 0.05 to beyond p< 0.05 : using history to contextualize p-values and significance testing. *The American Statistician*, 73(sup1) :82–90.

[Kim and Stern, 2008] Kim, C. and Stern, R. M. (2008). Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis. In *Ninth Annual Conference of the International Speech Communication Association*.

[Kim and Phillips, 2014] Kim, Y.-S. and Phillips, B. (2014). Cognitive correlates of listening comprehension. *Reading Research Quarterly*, 49(3) :269–281.

[Klare, 1974] Klare, G. R. (1974). Assessing readability. *Reading Research quarterly*, pages 62–102.

[Kotula, 2014] Kotula, K. (2014). Les pratiques innovatrices en classe de langues. l'édition et la production des films dans l'enseignement de fle. *Studia Romanica Posnaniensia*, 41(3) :47–61.

[Kruger et al., 2013] Kruger, J.-L., Hefer, E., and Matthew, G. (2013). Measuring the impact of subtitles on cognitive load : Eye tracking and dynamic audiovisual texts. In *Proceedings of the 2013 Conference on Eye Tracking South Africa*, pages 62–66.

[Kumari, 2008] Kumari, S. (2008). Multicollinearity : Estimation and elimination. *Journal of Contemporary research in Management*, 3(1) :87–95.

[Kurita, 2012] Kurita, T. (2012). Issues in second language listening comprehension and the pedagogical implications. *Accents Asia*, 5(1) :30–44.

[Larsby et al., 2005] Larsby, B., Hällgren, M., Lyxell, B., and al (2005). Cognitive performance and perceived effort in speech processing tasks : effects of different noise backgrounds in normal-hearing and hearing-impaired subjects desempeño cognitivo y percepción del esfuerzo en tareas de procesamiento del lenguaje : Efectos de las diferentes condiciones de fondo en sujetos normales e hipoacúsicos. *International Journal of Audiology*, 44(3) :131–143.

[Lecumberri et al., 2010] Lecumberri, M. L. G., Cooke, M., and Cutler, A. (2010). Non-native speech perception in adverse conditions : A review. *Speech communication*, 52(11-12) :864–886.

[Legrand, 2004] Legrand, G. (2004). *Approche méthodologique de sélection et construction de variables pour l'amélioration du processus d'extraction des connaissances à partir de grandes bases de données*. PhD thesis, Lyon 2.

[Levenshtein, 1965] Levenshtein, V. (1965). Levenshtein distance.

[Likas et al., 2003] Likas, A., Vlassis, N., and Verbeek, J. J. (2003). The global k-means clustering algorithm. *Pattern recognition*, 36(2) :451–461.

[Liu et al., 2019] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta : A robustly optimized bert pretraining approach. *arXiv preprint arXiv :1907.11692*.

[Lively and Pressey, 1923] Lively, B. A. and Pressey, S. L. (1923). A method of measuring the 'vocabulary burden' of textbooks. *Educational Administration and Supervision*, 9 :389–398.

[Long, 2014] Long, M. (2014). Design of Studios and Listening Rooms. pages 829–871. Academic Press.

[Manning et al., 2014] Manning, C. D., Surdeanu, M., Bauer, J., and et al (2014). The stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics : system demonstrations*, pages 55–60.

[Markham et al., 2001] Markham, P. L., Peter, L. A., and McCarthy, T. J. (2001). The effects of native language vs. target language captions on foreign language students' DVD video comprehension. *Foreign language annals*, 34(5) :439–445.

[Marslen-Wilson and Welsh, 1978] Marslen-Wilson, W. D. and Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive psychology*, 10(1) :29–63.

[Martin et al., 2020] Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., de la Clergerie, E. V., Sagot, B., and Seddah, D. (2020). Les modèles de langue contextuels camembert pour le français : impact de la taille et de l'hétérogénéité des données d'entrainement. In *JEP-TALN-RECITAL 2020-33ème Journées d'Études sur la Parole, 27ème Conférence sur le Traitement Automatique des Langues Naturelles, 22ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, pages 54–65. ATALA ; AFCP.

[Martin et al., 2019] Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., de la Clergerie, É. V., Seddah, D., and Sagot, B. (2019). Camembert : a tasty french language model. *arXiv preprint arXiv :1911.03894*.

[Mayer, 2001] Mayer, R. E. (2001). *Multimedia learning*. New York : Cambridge University Press.

[Miles, 2014] Miles, J. (2014). R squared, adjusted r squared. *Wiley StatsRef : Statistics Reference Online*.

[Miller, 1956] Miller, J. G. (1956). General behavior systems theory and summary. *Journal of Counseling Psychology*, 3(2) :120.

[Mitterer and McQueen, 2009] Mitterer, H. and McQueen, J. M. (2009). Foreign subtitles help but native-language subtitles harm foreign speech perception. *PloS one*, 4(11) :e7785.

[Moddemeijer, 1989] Moddemeijer, R. (1989). On estimation of entropy and mutual information of continuous distributions. *Signal processing*, 16(3) :233–248.

[Moore, 2001] Moore, A. W. (2001). Cross-validation for detecting and preventing overfitting. School of Computer Science Carneigie Mellon University.

[Nachar et al., 2008] Nachar, N. et al. (2008). The mann-whitney u : A test for assessing whether two independent samples come from the same distribution. *Tutorials in quantitative Methods for Psychology*, 4(1) :13–20.

[Nagelkerke et al., 1991] Nagelkerke, N. J. et al. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, 78(3) :691–692.

[Nation, 2006] Nation, I. (2006). How large a vocabulary is needed for reading and listening ? *Canadian modern language review*, 63(1) :59–82.

[New et al., 2004] New, B., Pallier, C., Brysbaert, M., and et al. (2004). Lexique 2 : A new French lexical database. *Behavior Research Methods, Instruments, & Computers*, 36(3) :516–524.

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

[New et al., 2005b] New, B., Pallier, C., and Ferrand, L. (2005b). Manuel de lexique 3. *Behavior Research Methods, Instruments, & Computers*, 36(3) :516–524.

[Nissan et al., 1995] Nissan, S., DeVincenzi, F., and Tang, K. L. (1995). An analysis of factors affecting the difficulty of dialogue items in TOEFL listening comprehension. *ETS Research Report*, 1995(2) :i–42.

[Noriega, 2005] Noriega, L. (2005). Multilayer perceptron tutorial. School of Computing. Staffordshire University.

[Noro, 2006] Noro, T. (2006). Developing a construct model of “Listening Stress" : A qualitative study of the affective domain of the listening process. *ARELE : Annual Review of English Language Education in Japan*, 17 :61–70.

[Norris and Ortega, 2009] Norris, J. M. and Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA : The case of complexity. *Applied linguistics*, 30(4) :555–578.

[Ortiz Suárez et al., 2020] Ortiz Suárez, P. J., Romary, L., and Sagot, B. (2020). A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

[Oshiro et al., 2012] Oshiro, T. M., Perez, P. S., and Baranauskas, J. A. (2012). How many trees in a random forest ? In *International workshop on machine learning and data mining in pattern recognition*, pages 154–168. Springer.

[Pascual et al., 2019] Pascual, S., Ravanelli, M., Serrà, J., Bonafonte, A., and Bengio, Y. (2019). Learning Problem-Agnostic Speech Representations from Multiple Self-Supervised Tasks. In *Proc. of the Conf. of the Int. Speech Communication Association (INTERSPEECH)*, pages 161–165.

[Pearson, 1895] Pearson, K. (1895). VII. Note on regression and inheritance in the case of two parents. In *proceedings of the royal society of London*, volume 58, pages 240–242.

[Perez et al., 2013] Perez, M. M., Noortgate, W. V. D., and Desmet, P. (2013). Captioned video for L2 listening and vocabulary learning : A meta-analysis. *System*, 41(3) :720–739.

[Perfetti and Stafura, 2014] Perfetti, C. and Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific studies of Reading*, 18(1) :22–37.

[Peterson, 1883] Peterson, L. E. (1883). K-nearest neighbor. *Scholarpedia*, 4(2).

[Pinquier, 2004] Pinquier, J. (2004). *Indexation sonore : recherche de composantes primaires pour une structuration audiovisuelle*. PhD thesis, Université Toulouse III. Thèse de doctorat.

[Pinquier et al., 2002] Pinquier, J., Rouas, J.-L., and André-Obrecht, R. (2002). Robust speech/music classification in audio documents. In *Seventh International Conference on Spoken Language Processing*.

[Pinquier et al., 2003] Pinquier, J., Rouas, J.-L., and André-Obrecht, R. (2003). A fusion study in speech/music classification. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03)*, volume 2, pages II–17.

[Povey et al., 2011] Povey, D., Ghoshal, A., Boulianne, G., and et al (2011). The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*.

[Puren, 1988] Puren, C. (1988). *Histoire des méthodologies de l'enseignement des langues*. CLE International.

[Puren, 1989] Puren, C. (1989). L'enseignement scolaire des langues vivantes étrangères en france au xixe siècle ou la naissance d'une didactique. *Langue française*, (82) :8–19.

[Randria et al., 2020a] Randria, E., Fontan, L., Le Coz, M., Ferrané, I., and Pinquier, J. (2020a). Étude des facteurs affectant la compréhensibilité de documents multimodaux : une étude expérimentale. In *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 1 : Journées d'Études sur la Parole*, volume 1, pages 534–542. ATALA ; AFCP.

[Randria et al., 2020b] Randria, E., Fontan, L., Le Coz, M., Ferrané, I., and Pinquier, J. (2020b). Subjective evaluation of comprehensibility in movie interactions. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2348–2357.

[Ravanelli et al., 2020] Ravanelli, M., Zhong, J., Pascual, S., Swietojanski, P., Monteiro, J., Trmal, J., and Bengio, Y. (2020). Multi-task self-supervised learning for Robust Speech Recognition. *ArXiv :2001.09239*.

[Redmon et al., 2016] Redmon, J., Divvala, S., Girshick, R., and et al (2016). You only look once : Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788.

[Ringnér, 2008] Ringnér, M. (2008). What is principal component analysis ? *Nature biotechnology*, 26(3) :303–304.

[Riquois, 2010] Riquois, E. (2010). Évolutions méthodologiques des manuels et matériels didactiques complémentaires en FLE. *Education & Formation*, (e-292) :129–142.

[Robinson et al., 2003] Robinson, B. F., Mervis, C. B., and Robinson, B. W. (2003). The roles of verbal short-term memory and working memory in the acquisition of grammar by children with williams syndrome. *Developmental Neuropsychology*, 23(1-2) :13–31.

[Rousseeuw, 1987] Rousseeuw, P. J. (1987). Silhouettes : a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20 :53–65.

[Roze et al., 2012] Roze, C., Danlos, L., and Muller, P. (2012). LEXCONN : a French lexicon of discourse connectives. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (10).

[Rupp et al., 2001] Rupp, A. A., Garcia, P., and Jamieson, J. (2001). Combining multiple regression and cart to understand difficulty in second language reading and listening comprehension test items. *International Journal of Testing*, 1(3 & 4) :185–216.

[Saunders, 1996] Saunders, J. (1996). Real-time discrimination of broadcast speech/music. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, volume 2, pages 993–996. IEEE.

[Seara, 2001] Seara, A. R. (2001). L'évolution des méthodologies dans l'enseignement du français langue étrangère depuis la méthodologie traditionnelle jusqu'à nos jours. *Cuadernos del Marqués de San Adrián : revista de humanidades*, (1) :139–161.

[Seong et al., 2019] Seong, T. W., Ibrahim, M., and Mulvaney, D. (2019). Wadaw : A modified wada snr estimator for audio-visual speech recognition.

[Serban et al., 2017] Serban, I. V., Sankar, C., Germain, M., Zhang, S., Lin, Z., Subramanian, S., Kim, T., Pieper, M., Chandar, S., Ke, N. R., et al. (2017). A deep reinforcement learning chatbot. *arXiv preprint arXiv :1709.02349*.

[Sherman, 1893] Sherman, L. A. (1893). *Analytics of literature : A manual for the objective study of English prose and poetry*. Ginn.

[Stokes and Klee, 2009] Stokes, S. F. and Klee, T. (2009). Factors that influence vocabulary development in two-year-old children. *Journal of Child Psychology and Psychiatry*, 50(4) :498–505.

[Styler, 2013] Styler, W. (2013). *Using Praat for linguistic research*. University of Colorado at Boulder Phonetics Lab.

[Sueyoshi and Hardison, 2005] Sueyoshi, A. and Hardison, D. M. (2005). The role of gestures and facial cues in second language listening comprehension. *Language Learning*, 55(4) :661–699.

[Sweller et al., 1988] Sweller, J., Van Merrienboer, J. J., and Paas, F. G. (1988). Cognitive architecture and instructional design. *Educational psychology review*, 10(3) :251–296.

[Tabbers et al., 2004] Tabbers, H. K., Martens, R. L., and Van Merrienboer, J. J. (2004). Multimedia instructions and cognitive load theory : Effects of modality and cueing. *British journal of educational psychology*, 74(1) :71–81.

[Templin, 1957] Templin, M. C. (1957). *Certain language skills in children ; their development and interrelationships*.

[Thoiron and Arndaud, 1992] Thoiron, P. and Arndaud, P. J. (1992). Quelques aspects de la perception de la richesse lexicale. *Apparences textuelles et réalité linguistique, CYCNOS*, (8) :33–47.

[Thompson and Rubin, 1996] Thompson, I. and Rubin, J. (1996). Can strategy instruction improve listening comprehension ? *Foreign Language Annals*, 29(3) :331–342.

[Thorndike, 1921] Thorndike, E. L. (1921). *The teacher's word book*.

[Traverso, 1996] Traverso, V. (1996). *La conversation familière : analyse pragmatique des interactions*. Presses Universitaires Lyon.

[Urieli and Tanguy, 2013] Urieli, A. and Tanguy, L. (2013). L'apport du faisceau dans l'analyse syntaxique en dépendances par transitions : études de cas avec l'analyseur Talismane. In *20e conférence du Traitement Automatique du Langage Naturel (TALN)*, page (publication en ligne).

[Van Ek, 1975] Van Ek, J. A. (1975). The threshold-level. *Education and Culture*.

[Vandergrift, 2007] Vandergrift, L. (2007). Recent developments in second and foreign language listening comprehension research. *Language teaching*, 40(3) :191.

[Vargas, 2006] Vargas, C. (2006). Sociolinguistique et didactique de la langue première. *Skholê, Hors-série (1)*, pages 1–6.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv :1706.03762*.

[Voss, 1979] Voss, B. (1979). Hesitation phenomena as sources of perceptual errors for non-native speakers. *Language and Speech*, 22(2) :129–144.

[Voss, 1984] Voss, B. (1984). Perception of first language and second language texts : A comparative study. *Bielefelder Beitrage zur Spachlehrforschung*, 13 :131–53.

[Watson, 1916] Watson, J. B. (1916). The place of the conditioned-reflex in psychology. *Psychological review*, 23(2) :89.

[Watzlawick et al., 1972] Watzlawick, P., Beavin, J. H., Jackson, D. D., and Morche, J. (1972). Une logique de la communication.

[Weil, 2001] Weil, S. A. (2001). Foreign accented speech : Adaptation and generalization. Master's thesis, Ohio State University.

[Welch et al., 1994] Welch, B. L., Cole, D. N., McArthur, E. D., Booth, G. D., Geier-Hayes, K., and Sloan, J. P. (1994). *Identifying proxy sets in multiple linear regression : an aid to better coefficient interpretation.* Number 470-476. US Department of Agriculture, Forest Service, Intermountain Research Station.

[Wierzynski and Fiscus, 2000] Wierzynski, C. and Fiscus, J. (2000). "stnr.doc". Included with Speech File Manipulation Software (SPHERE) Package Version 2.7.

[Woisard et al., 2013] Woisard, V., Espesser, R., Ghio, A., and 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.

[Wong and Waring, 2010] Wong, J. and Waring, H. Z. (2010). *Conversation Analysis and Second Language Pedagogy.* NY : Taylor & Francis.

[Xu et al., 2017] Xu, M., Fralick, D., Zheng, J. Z., Wang, B., Tu, X. M., and Feng, C. (2017). The differences and similarities between two-sample t-test and paired t-test. *Shanghai archives of psychiatry*, 29(3) :184.

[Yang et al., 2016] Yang, S., Luo, P., Loy, C.-C., and et al. (2016). Wider face : A face detection benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5525–5533.

[Yasin et al., 2017] Yasin, B., Mustafa, F., and Permatasari, R. (2017). How Much Videos Win over Audios in Listening Instruction for EFL Learners. *Turkish Online Journal of Educational Technology-TOJET*, 17(1) :92–100.

[Yu et al., 2012] Yu, D., Seide, F., and Li, G. (2012). Conversational speech transcription using context-dependent deep neural networks. In *ICML.*

[Zbib et al., 2019] Zbib, R., Zhao, L., Karakos, D., Hartmann, W., DeYoung, J., Huang, Z., Jiang, Z., Rivkin, N., Zhang, L., Schwartz, R., et al. (2019). Neural-network lexical translation for cross-lingual ir from text and speech. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 645–654.

[Zhang et al., 2013] Zhang, L., Liu, Z., and Ni, J. (2013). Feature-based assessment of text readability. In *2013 Seventh International Conference on Internet Computing for Engineering and Science*, pages 51–54.

[Zhang et al., 2020] Zhang, Y., Tiňo, P., Leonardis, A., and Tang, K. (2020). A survey on neural network interpretability. *arXiv preprint arXiv :2012.14261.*

[Zhao, 1997] Zhao, Y. (1997). The effects of listeners' control of speech rate on second language comprehension. *Applied linguistics*, 18(1) :49–68.

[Zipf, 1935] Zipf, G. K. (1935). *The psycho-biology of language : An introduction to dynamic philology.* MA : MIT Press.