Towards multimodal assessment of L2 level : speech and eye tracking features in a cross-cultural setting

Sofiya Kobylyanskaya

université PARIS-SACLAY

Towards multimodal assessment of L2 level: speech and eye tracking features in a cross-cultural setting

Vers l'évaluation multimodale du niveau de L2: caractéristiques de la parole et du mouvement des yeux dans un cadre multimodal

Thèse de doctorat de l'université Paris-Saclay

École doctorale n°580 Sciences et Technologies de l'Information et de la Communication (STIC)
Spécialité de doctorat : Informatique
Graduate School : Informatique et Sciences du Numérique.
Référent Faculté des Sciences d'Orsay:

Thèse préparée dans Laboratoire Interdisciplinaire des Sciences du Numérique (Université Paris-Saclay, CNRS), sous la direction de **Ioana VASILESCU,** directrice de recherche, le co-encadrement de **Laurence DEVILLERS,** Professeure des universités

Thèse soutenue à Paris-Saclay, le 19 décembre 2024, par

Sofiya KOBYLYANSKAYA

Composition du Jury

Membres du jury avec voix délibérative

Cédric GENDROT Professeur, Université Sorbonne Nouvelle	Président
Véronique DELVAUX Professeure, Université de Mons	Rapporteure & Examinatrice
Sophie HERMENT Professeure, Université d'Aix-Marseille	Rapporteure & Examinatrice
Ralph ROSE Professeur, Université Waseda	Examinateur
Sébastien LALLÉ Maître de Conférences, Sorbonne Université	Examinateur
Nicolas AUDIBERT Maître de Conférences, Université Sorbonne Nouvelle	Examinateur

THÈSE DE DOCTORAT

NNT : 2024UPASG111

Acknowledgements

These three years have been among the most challenging of my life, but also among the most rewarding ones. During this journey, I got to know people who helped me to grow professionally and personally.

First of all, I am expressing my deepest gratitude to my research directors, Ioana Vasilescu and Laurence Devillers for their high-quality mentorship, their support and guidance throughout these three years. Thanks to you, I learned how to conduct rigorous research and how to think critically. I am deeply grateful for everything you have done.

I am also thankful to Olivier Augereau for his guidance for eye-tracking research.

I am thankful to our partners from Osaka Metropolitan University: Koichi Kise, Motoi Iwata and Andrew Vargo for their investment in the LeCycl project, for their nicest reception during my stay in Japan which was the most rewarding part of my thesis.

I am thankful to people who helped me discover life in Japan, its cuisine and culture : Taro, Peter, Kubra, and Christian.

I am thankful to all the participants of the experiments without whom this thesis would be impossible to realize. Thank you for putting your time and energy into devotion to science!

I am thankful to all the lab members, who made my work in the lab much more productive and colorful: Natalia, Tom, Yajing, Hugues, Théo, Paul, Elise, Alban, François, Emmett, Simon, Lucas, Shu, Rémi, Léa-Marie, Camille, Nicolas, Nesrine, Marco, Armand, Mathieu, Paritosh, Atilla, Pierre, Maxime, David, Isabelle.

I am thankful to Anisia for her help with data modeling and her support at conferences.

I am thankful to Mathilde, by best "colloque de bureau" for her support and our Parisian “goûters”.

I am thankful to my closest friends for their support and sense of humor, who make me feel at home everywhere I go: Natalia, Gulnara, Victoria, Anna, Kateryna, one more Anna, Kristina, Panagiota, Irving, and Nadine. I have known you all since my arrival to France (or even before!) and you became my dearest friends who make my life cozier and the days sunnier.

I am thankful to Juan, my partner in crime, who has become an important part of my life and who has always been there for me even when the ocean (the Atlantic or the Pacific) set us apart.

I am thankful to my family, especially to my mum, Natalia, my best friend, my soulmate, who gave me life, has always encouraged and supported me. I love you to the moon and back.

I consider myself a lucky person, because I am surrounded by wonderful people.

I am thankful to each and all of you!

Résumé

Ces dernières années, le monde de l'éducation a subi des changements importants, notamment avec la digitalisation massive du système en 2020 et l'avancement des technologies d'IA générative.

Le projet ANR trilatéral LeCycl "Learning Cyclotron" (Vargo et al., 2023) entre la France (LISN, IMT Brest), le Japon (Université Métropolitaine d'Osaka) et l'Allemagne (DFKI), s'inscrit dans cette tendance scientifique et vise à accélérer la circulation des connaissances. Il se focalise sur trois processus principaux d'apprentissage : la perception, la maîtrise et le transfert des connaissances. Cette thèse, faisant partie du projet LeCycl, exploite des stratégies de lecture à voix haute en langue étrangère (L2) en analysant les difficultés rencontrées par des représentants de différentes langues maternelles et leurs stratégies pour les surmonter. À cette fin, nous utilisons des indices multimodaux (la parole et le mouvement des yeux), ainsi qu'un protocole original introduisant des nudges (représentés par des manga, des bandes dessinées japonaises) pour l'adaptation culturelle (Hutin et al., 2022). Le protocole implique la collecte de données de lecture à voix haute des locuteurs Français et Japonais de l'anglais L2 (enregistrés dans leur pays d'origine en France ou au Japon) et des locuteurs natifs de l'anglais dont la production nous a servi de référence (Kobylyanskaya, 2022).

La performance des locuteurs a été évaluée à travers des mesures multimodales, que sont la parole et le mouvement des yeux (El Baha et al., 2022; Kobylyanskaya et al., 2023). Notamment, nous avons modélisé des mesures acoustiques et linguistiques, par exemple, la réalisation des phonèmes, la prosodie et l'utilisation des disfluences, telles que les pauses, les hésitations, et les troncations qui sont des indicateurs de la fluence en L2. L'autre modalité considérée est le mouvement oculaire qui a été associé dans la littérature à des informations métacognitives, telles que la compréhension du texte (Augereau et al., 2016a), la charge cognitive (Bax, 2013) ou des émotions (Lallé et al., 2021). La modélisation statistique des caractéristiques multimodales a mis en valeur leur corrélation avec le niveau de la maîtrise de l'anglais L2 et la langue d'origine des participants.

Notre deuxième contribution consiste en l'implémentation des méthodes d'apprentissage automatique pour la prédiction du niveau de la maîtrise de la L2. Pour cela, nous nous somme basés sur les caractéristiques de la parole, du mouvement des yeux ou leur combinaison, et avons exploré les caractéristiques qui contribuent le plus à la décision du modèle. Les résultats soulignent que le modèle se base sur différentes caractéristiques vocales et occulaires pour détecter le niveau L2 des représentants de différentes langues maternelles.

Notre dernière contribution consiste en l'élaboration d'un protocole préliminaire permettant d'évaluer comment la combinaison des informations textuelles et visuelles dans les mangas contribuent à la performance des locuteurs lors de la lecture à voix haute. Ce protocole ouvre des perspectives vers l'implémentation des nudges adaptés à la culture et à la langue maternelle des apprenants de la L2. Ce travail a également mis en lumière des difficultés de transfert d'un support de lecture spécifique à une culture vers une autre.

En conclusion, ce travail de thèse a mis en évidence que les représentants de langues maternelles différentes, tels que les Français et les Japonais, sont confrontés à des défis spécifiques lorsqu'ils lisent des textes en langue étrangère ; ils adoptent des stratégies variées pour les surmonter, qui se traduisent à la fois aux niveaux verbal et oculaire. Les résultats de notre travail soulignent l'intérêt de l'utilisation des caractéristiques multimodales pour l'évaluation du niveau L2, ainsi que la nécessité de développer des outils d'apprentissage adaptés aux cultures et les défis qui sont associés à leur conception.

Contents

1 Introduction **10**

1.1 General framework . . . 10

1.2 Thesis framework . . . 11

1.3 Research Questions . . . 14

1.4 Work Hypotheses . . . 14

1.5 Thesis plan . . . 15

1.6 List of contributions . . . 16

2 State of the art **17**

2.1 Automatic L2 pronunciation and fluency assessment measures . . . 18

2.1.1 Segmental level . . . 19

2.1.2 Suprasegmental level . . . 22

2.1.3 Features affecting speech fluency . . . 23

2.2 Eye-tracking in L2 studies . . . 27

2.3 L2 learning modeling with machine learning approaches . . . 29

2.4 Nudges in education and L2 learning . . . 32

2.4.1 Summary . . . 33

3 Data collection and processing **35**

3.1 Existing corpora and the need of a new data collection . . . 35

3.1.1 Eye tracking corpora . . . 35

3.1.2 Speech corpora . . . 36

3.1.3 Multimodal data: eye tracking and reading aloud . . . 36

3.1.4 Summary . . . 37

3.2 Data collection . . . 37

3.2.1 Participants . . . 39

3.2.2 Reading support . . . 39
3.2.3 Equipment . . . 42
3.2.4 Procedure . . . 45
3.2.5 Summary . . . 46
3.3 Data processing . . . 47
3.3.1 Speech data processing . . . 47
3.3.2 Eye tracking data processing . . . 49

4 Speech and eye-tracking features extraction and modeling 53
4.1 Text level . . . 54
4.1.1 Speech features . . . 54
4.1.2 Eye Tracking . . . 66
4.1.3 Summary . . . 70
4.2 Word Level . . . 70
4.2.1 Frequent and infrequent lexical words vs function words . . . 70
4.2.1.1 Speech features . . . 70
4.2.1.2 Eye tracking features . . . 76
4.2.2 Difficult words . . . 77
4.3 Phone level . . . 82

5 L2 level prediction based on speech and eye tracking features 91
5.1 Data pre-processing . . . 91
5.2 Model selection . . . 93
5.3 Results: all features . . . 94
5.4 Results: selected features . . . 98
5.4.1 Speech features . . . 98
5.4.2 Eye-tracking features . . . 100
5.4.3 Speech and eye tracking features . . . 102
5.5 Summary . . . 106

6 Towards using culturally adapted nudging strategies in L2 learning: example of manga 109
6.1 Data acquisition and processing . . . 110
6.2 Preliminary results . . . 112
6.3 Limitations and future work . . . 116

7 Conclusions and perspectives **119**

7.1 Conclusions 119

7.2 Future work 122

A Texts for reading aloud experiments **124**

A.1 Beginner text 124

A.2 Intermediate text 124

A.3 Advanced text 125

B Manga extract in textual form **127**

List of Figures

1.1 Learning Cyclotron 12

2.1 Vocal trapezius English(black) vs French (blue) 20
2.2 Vocal trapezius English(black) vs Japanese (red) 20
2.3 English diphthongs 21

3.1 Manga example (from "Delicious in Dungeon") 43
3.2 EyeGotIt system algorithm (El Baha et al., 2022) 44
3.3 Eye movement recording using Tobii eye tracker 45
3.4 Eye movement visualization during a reading aloud task 45
3.5 Alignment correction in Praat 48
3.6 Speech processing algorithm 49
3.7 Example of well calibrated eye-tracking data 49
3.8 Example of poorly calibrated eye-tracking data 50
3.9 Definition of word region and calculation of gazed words 51

4.1 Speech rate per text level, L2 level and L1 55
4.2 Pitch variation (semitone) per text level, L2 level, and L1 58
4.3 Mean intensity per text level, L2 level and L1 59
4.4 Intensity variation per text level, L2 level, and L1 60
4.5 Percent of disfluencies per text level, L2 level, and L1 62
4.6 Number of pauses by text level, L2 level, and L1 64
4.7 Percentage of pauses per text level, L2 level, and L1 64
4.8 Mean pauses duration by text level, L2 level, and L1 65
4.9 Pauses duration variation by text level, L2 level, and L1 65
4.10 Mean fixation duration per text level, L2 level, and L1 68
4.11 Number of fixations per text level, L2 level, and L1 68

4.12 Total fixation duration per text level, L2 level, and L1 69
4.13 Pitch variation (semitone) per word type, L2 level, and L1 71
4.14 Intensity variation per word type, L2 level, and L1 72
4.15 Word duration by type, L2 level, and L1 73
4.16 Number of syllables per word type 73
4.17 Total fixation duration per word type, L2 level, and L1 77
4.18 Number of difficult words reported by French and Japanese speakers 78
4.19 Difficult words analysis: sliding window 79
4.20 Percentage of cases with pauses before (non-)difficult words 80
4.21 Percentage of cases with truncations before (non-)difficult words 80
4.22 Percentage of cases with hesitations before (non-)difficult words 81
4.23 Pauses duration before (non-)difficult words 81
4.24 Mean fixation duration before (non-)difficult words 82
4.25 Number of fixations before (non-)difficult words 83
4.26 [u] location in the vocal space of English (black), Japanese (red) and French (blue) languages 84
4.27 Vocal space of /u/ and /ʊ/ of French speakers 85
4.28 Vocal space of /u/ and /ʊ/ of Japanese speakers 85
4.29 Vocal space of /u/ and /ʊ/ of English speakers 86
4.30 F1 of /u/ and /ʊ/ realized by French, Japanese and English speakers 87
4.31 F2 of /u/ and /ʊ/ realized by French, Japanese and English speakers 87
4.32 F3 of /u/ and /ʊ/ realized by French, Japanese and English speakers 88
4.33 Vowel Duration of /u/ and /ʊ/ realized by French, Japanese and English speakers 89

5.1 L2 level prediction per page and per speaker 94
5.2 Important audio features for L2 level prediction (French speakers) 96
5.3 Important audio features for L2 level prediction (Japanese speakers) 97
5.4 Important eye tracking features for L2 level prediction (French speakers) 97
5.5 Important eye tracking features for L2 level prediction (Japanese speakers) 98
5.6 Important updated speech features for L2 level prediction (French speakers) 99
5.7 Important updated speech features for L2 level prediction (Japanese speakers) 100
5.8 Important updated eye-tracking features for L2 level prediction (French speakers) 101
5.9 Important updated eye-tracking features for L2 level prediction (Japanese speakers) 101
5.10 Important updated combined speech and eye-tracking features for L2 level prediction (French speakers)104

5.11 Important updated combined speech and eye-tracking features for L2 level prediction (Japanese speakers) 104

6.1 Percentages of pauses of Japanese speakers reading manga in original vs textual form 113
6.2 Phonation duration (s) of Japanese speakers reading manga in original vs textual form 114
6.3 Articulation rate of Japanese speakers reading manga in original vs textual form 114
6.4 Percentage of disfluencies used by French speakers reading manga vs texts 115
6.5 Percentages of pauses used by French speakers reading manga vs texts 116
6.6 Articulation rate used by French speakers reading manga vs texts 116

List of Tables

2.1 English vowel substitutions by French and Japanese speakers as highlighted by state of the art. 21
2.2 English consonants (in blue - consonants absent in French, in red - consonants absent in Japanese, violet - consonants absent both in French and Japanese) 22
2.3 Japanese consonants (in red - consonants absent in English) 22
2.4 French consonants (in blue - consonant present in French and absent in English) 22

3.1 Research protocols and collected data as function of experimental set up and L1. 38
3.2 Number of French and Japanese participants: Protocols 1 and 3 39
3.3 Lexical characteristics of the selected texts 41
3.4 Syntactic characteristics of selected texts 41
3.5 Number of speakers and files before and after filtering eye-tracking data 50

4.1 Number of speakers using disfluencies 61
4.2 Sample count using sliding window technique 79
4.3 Number of vowels before/after filtering 85
4.4 Pilai score results: vowel space overlap 86

5.1 Feature sets for L2 level prediction (in red deleted redundant features) 92
5.2 Number of speakers and vectors used for L2 level prediction 93
5.3 Results of L2 level prediction without feature selection 95
5.4 L2 prediction results with updated speech features for French (6 features) and Japanese speakers (9 features) 99
5.5 L2 prediction results with updated eye-tracking features for French (9 features) and Japanese speakers (8 features) 101
5.6 Updated and reduced speech and eye-tracking feature sets for French and Japanese speakers 102
5.7 L2 prediction results with updated combined speech and eye-tracking features (French and Japanese speakers) 103

5.8 Per class performance for the feature set (speech + eye-tracking) 105
5.9 Per class performance with undersampling the majority class (French speakers) 105
5.10 Features detected with Linear Mixed Effect model (LME) vs features detected with Random Forest classifier (RF) 106

6.1 Number of words in manga read by French and Japanese speakers 112

Chapter 1

Introduction

1.1 General framework

In recent years, the world of education has undergone critical changes. Since 2020, there has been a massive digitalization of the education system, affecting the ways in which teachers and students interact and how students consume information. The entry of generative AI technologies such as ChatGPT in 2022 has further contributed to drastic transformation of the education sector. These changes have highlighted the need to adapt learning methods, knowledge assessment, teacher-student and student-student communication, as well as to improve computer-assisted learning tools and regulate their use. This need is recognized by international institutions. For example, the OECD's annual report highlights the importance of adapting to new learning contexts in order to reinvent learning environments by contributing to digital development without replacing teacher-learner and learner-learner relationships (Schleicher, 2020). At the European level, we can note the European Commission's Digital Education Action Plan, a policy initiative that aims to support the sustainable and effective adaptation of EU member states' education and training systems to the digital age. It covers digital technologies, support for the digitization of teaching methods and pedagogies, and the establishment of the infrastructures needed for inclusive and resilient distance learning. Finally, renowned language-learning institutions such as the British Council are placing a strong emphasis on artificial intelligence in digital learning, which they describe as the future of education.

OECD highlights the opportunities and provides guidelines for an efficient and ethical use of Artificial Intelligence in education (OCDE, 2023). Indeed, the use of AI helps educators in a variety of tasks, such as the creation of educational content, student performance evaluation, and the completion of administrative tasks. Furthermore, these technologies can make education more inclusive by personalizing learning for students facing difficulties and by making educational content more accessible to students with impairments (OCDE, 2023). Despite numerous benefits, the integration of AI into the educational process raises ethical concerns, such as the collection and sharing of personal data, particularly for minors, the reinforcement of global inequalities, algorithmic biases, and lack of ex-

plainability. This latter issue has been amplified with the emergence of generative AI technologies such as ChatGPT, which, despite their undeniable advantages, carry significant informational and cultural biases. These models are mostly trained on data from the United States, making their responses shaped to American culture and English language (Cao et al., 2023). This leads to a potential standardization of thought, social concepts, and language use. This cultural unification is a crucial problem for education, because representatives of different cultures require different learning approaches depending on their educational and communication habits as well as learning difficulties. The thesis investigates cross-culturality in education through the lens of L2 acquisition. We investigate oral L2 production by adopting a framework inspired by contemporary L2 level assessment tests and providing an in-depth, multimodal assessment utilizing spoken language variation and eye tracking features. These features provide valuable insights into the pronunciation difficulties faced by the speakers having different linguistic and cultural backgrounds and strategies they employ to cope with them. A cultural paradigm is a key aspect of this thesis. Specifically, we rely on Japanese manga, an educational medium which is known for being efficient for knowledge acquisition in Japanese culture and has gained popularity all over the world. In this study, we apply this cultural learning tool to two distinct cultures in order to discover how this approach can affect L2 learning across various cultural, educational and linguistic backgrounds. Furthermore, we demonstrate a use of natural language processing (NLP) and explicable machine learning (ML) techniques for L2 performance assessment and learning personalization based on cultural and individual patterns. Our results highlight the necessity and the challenges of developing culturally adapted learning tools, a relevant issue given the emergence of new technologies that may lead to blurring cultural boundaries. In the next section, we will present the ANR LeCycl (Learning Cyclotron) project (Vargo et al., 2023) that the thesis is part of, and the thesis overview.

1.2 Thesis framework

ANR LeCycl (Vargo et al., 2023) is an interdisciplinary trilateral project involving researchers from 3 countries (France, Japan, Germany) different fields such as computer science, artificial intelligence and data analysis, the humanities and social sciences (e.g. linguistics), economics, ethics, etc. As for the French side the project is funded by the ANR, Laurence Devillers being the French principal investigator. This project is linked to the HUMAAINE AI chair investigating on nudging strategies.

The project focuses on three major learning stages: Perception (via text, audio and eye tracking), Mastery (via training and exercises) and Transfer (via discussions and presentations) 1.1. At the Perception stage, the learner identifies gaps in their knowledge. Then, the system highlights the learner's weak points by tracking their behavior using the information captured by a variety of sensors such as microphone, eye tracking, video and so on. These tools measure factors such as attention, hesitations, and voice timbre to assess students' emotional and cognitive states as well as the knowledge acquisition process. Next, in the Master phase, the system suggests personalized

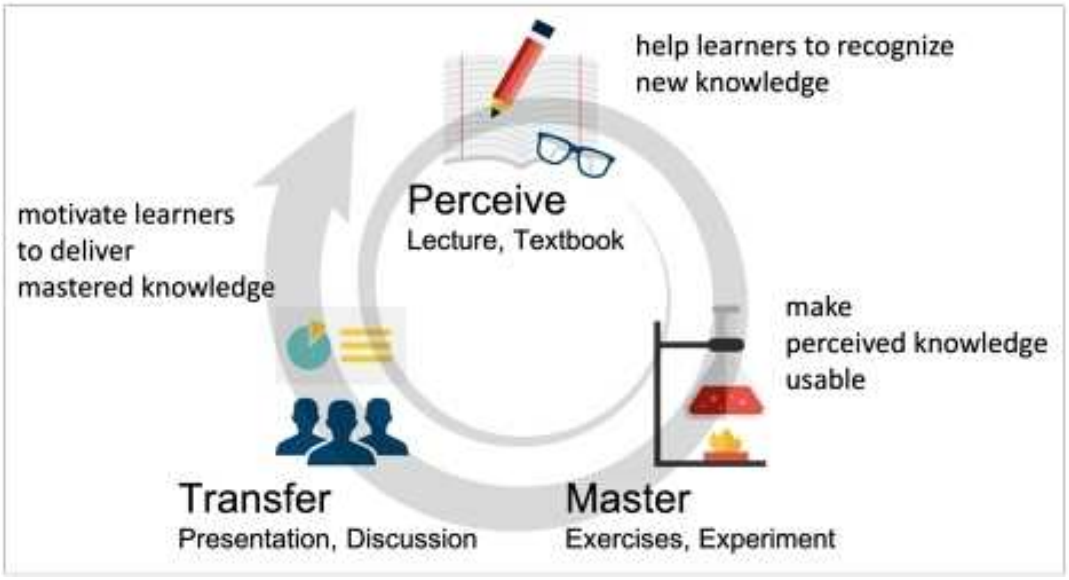

Figure 1.1: Learning Cyclotron

exercises in order to improve the area of knowledge that appeared to be challenging in the first stage. Finally, when the knowledge is efficiently mastered, the learner can transfer the acquired knowledge to other peers, facilitating knowledge circulation (Vargo et al., 2023). This last transfer stage is realized through video conferencing platforms such as Zoom, recording multiple modalities including speech, video and biological features derived from wearable devices such as E4 wristband. Additionally, the project employs nudging strategies (Thaler and Sunstein, 2008) - indirect influence strategies affecting student behavior by appealing to students' cognitive and emotional biases. Nuding techniques have been shown to be effective in various learning contexts (Hutin et al., 2022; Azmat and Iriberri, 2010; Goulas and Megalokonomou, 2015), and we hypothesize that they will enable us to guide the learning process and ensure effective knowledge acquisition and transfer.

The thesis is a part of the LeCycl project and focuses on learning English as a foreign language (L2) and concerns mostly the Perception phrase of the learning cycle.

The thesis is a continuation of work conducted by the project partners. In particular, our partners from Osaka Metropolitan University and ENIB lab (Brest) are specialists in pattern recognition from text and images, as well as in the use of different biological captors, such as eye-trackers. They conducted multiple studies in which eye tracking was employed with different reading supports, such as text and manga in order to assessed comprehension among other aspects. For example, they developed a content-based information retrieval method from manga images (Iwata et al., 2014), which are commonly used in Japan, both for leisure and education. They also showed how eye-tracking can be used for information retrieval from manga (Rigaud et al., 2016). Other studies focus on textual support and reading comprehension assessment during silent reading in a foreign language (Augereau et al., 2016a). As for French partner in the project, researchers from LISN have been working for a long time on speech variation, to a large extent in first language (L1) but also in a second language (L2). For example, they participated

in the creation and linguistic exploration of a corpus including the production of non-native English speakers TED (Translanguage English Database) (Kipp et al., 2002), they investigated the influence of L1 on the use of disfluencies (Vasilescu and Adda-Decker, 2006), the characteristics of a foreign accent (Boula de Mareüil and Vieru-Dimulescu, 2006). In addition, at LISN a particular emphasis is put on nudging strategies in the context of the ANR Chair AI HUMAAINE: HUman-MAchine Affective INteraction & Ethics directed by Laurence Devillers. For example, some studies of the Chair investigated the effect of nudges in the context of human-robot interaction (Ali Mehenni et al., 2021a,b; Kalashnikova et al., 2024).

In this thesis we aimed both at developing an original protocol in a multimodal context and in taking advantage of the unique framework of the chair HUMAAINE and of the trilateral research project.

For this purpose, we imagined a research protocol that conveys both speech and eye movement through a reading aloud set up, while eliciting multimodal nudging strategies (visual and textual) by variating the reading medium, from text to imaged books (manga). In particular, we aim at modeling cultural and linguistic specificities of learning English, assessing the proficiency level based on speakers' self-reported level, predicting challenges in order to improve the learning process and furthermore, create the conditions for knowledge transfer.

We propose and describe in the following chapters a complex protocol involving speech and eye movement allowing to elicit multimodal cues including acoustic, prosodic, paralinguistic and eye tracking and to introduce the notion of nudges related to cultural adaptation in foreign language learning.

Specifically, we examine the performance of French and Japanese speakers of English during a reading aloud task. The choice of these two categories of the participants is explained by their distinct linguistic and cultural backgrounds. These distinctions likely lead to different challenges when learning English and varying strategies for addressing these challenges. For example, we hypothesize that representatives of different cultures would show distinct phoneme realization, pausing patterns, disfluencies such as hesitations or truncations, or eye movement behavior depending on their reading habits. Additionally, French and Japanese speakers have different educational cultures, including the types of learning materials and approaches they are used to. For example, in Japanese culture, manga is a commonly used learning aid for various subjects, including foreign language learning. As for French culture, manga is a popular option for leisure reading, but is rarely used in educational contexts. Therefore, our study includes both text and manga as reading support that allows us to compare them in terms of contribution to the participants' oral production and reading comprehension.

Tu sum up in this thesis we aim at exploring the contribution of multimodal cues for estimating and automatically predicting L2 pronunciation difficulties by subjects of different mother tongues. In this purpose, we selected mother tongue (L1) languages, French and Japanese, that are genetically and typologically distant, in order to estimate the

impact of cultural differences in a broad sense. We also strive to estimate the contribution of different modalities: we focus on speech and eye movement, as well as the relationship between the variations observed and the English language selected as second (L2) language level as commonly estimated in language certifications worldwide. Finally, we strive to estimate the impact of nudges through multimodal learning settings that involve both text and image medium and answer, at least in a preliminary way, the question: is such reading and learning medium, commonly used in Japanese culture, easily transposable to another (Western) culture?

To meet these goals, we propose a protocol that includes such tasks as the comparison of French and Japanese speakers' performance during reading aloud English texts, L2 English proficiency assessment with selected multi-modal features, the use of machine learning approaches to model and predict pronunciation challenges in L2, and finally the evaluation of the contribution of the manga images to the ease the oral production.

1.3 Research Questions

In this thesis we build on the above issues and on the related work presented in the state-of-the art section as well as on the challenges of Educational Technologies as highligted by LeCycl project framework, and we aim at answering the following research questions :

1. **Q1:** Are the multimodal measure compared to speech (segmental and supra-segmental) measures better indicators of the level of L2 oral production (pronunciation, accent) and L2 mastery?
2. **Q2:** Can we generalize the above observations to speakers' cultural patterns? How the mother tongue and, more broadly, the cultural profile of subjects influence these patterns?
3. **Q3:** How machine learning techniques can help us model and predict these patterns?
4. **Q4:** How can we effectively implement nudges in education? Specifically, can we subtly influence reading performance using nudges? Here, the nudges will be manga images, which we assume might enhance text comprehension and thus are supposed to influence in the sense of efficiency and improvement, pronunciation and intonation.

1.4 Work Hypotheses

Based on these questions, we formulate the following **working hypotheses**:

1. **H1: The impact of multimodality:** Eye tracking features supports the assessment of the level of expression in L2 and can complement linguistic cues (including acoustic, prosodic and paralinguistic cues, on which we particularly focus in this work) (Augereau et al., 2016a; Kobylyanskaya, 2022).

2. **H2: Speech features (partially) predict L2 level:** The L2 level that the participants declare during the experiment will only partially correlate with their oral expression features translating in various segmental and supra-segmental parameters. Indeed, language proficiency as measured in education, is reflected in a variety of skills, such as written and oral comprehension and expression, and includes mastery of vocabulary and syntactic structures. Thus, this thesis will estimate to what extent pronunciation is related to L2 level through statistical correlation and machine learning techniques in order to predict the level. This approach has long term potential applications as is intended to help detecting the features that best correlate with L2 and can succesfully been implemented in real-life applications.

3. **H3: Manga as nudges, a preliminary experiment:** The reading support plays a role in L2 expression, which is why one of our experimental protocols involves reading texts with different support (text vs. manga).

4. **H4: The relevance of cultural patterns:** The cross-cultural dimension is important and is due to the biases of the educational system and the linguistic environment (Grainger, 2012), which is why we collected data from a French-speaking and a Japanese-speaking audience, as well as control data from native English speakers, and that in several experimental conditions.

1.5 Thesis plan

In order to answer the research questions and test the hypotheses, we organized our manuscript in the following way.

In Chapter 2, we provide a literature review with respect to the main aspects of the thesis. We thus address the challenges of automatic L2 fluency assessment and present an overview of speech and eye tracking features commonly used in recent studies involving L2 proficiency. We also present the specificities of French and Japanese languages compared to English in order to highlight the challenges that the speakers might face due to their L1. Finally we propose a state of the art of NLP (natural language processing) approaches with the aim of highlighting speech and eye tracking features generally modeled and the most efficient techniques currently being used. In the last part of this chapter, we discuss the role of nudging strategies in Education and L2 acquisition.

Chapter 3 is dedicated to the description of the experimental protocols, the methodology of data collection and processing. We also address the challenges of fully-automatic non-native speech processing.

In Chapter 4, we propose a range of speech and eye tracking characteristics allowing us to assess the L2 level and to underpin cultural differences. We provide statistical modeling to capture the relationship between the extracted features and the participants' L2 level, L1 influence, text comprehension score, text type etc.

Then, in Chapter 5, we utilize speech and eye tracking features in a machine learning algorithm in order to automatically assess L2 level and evaluate the most important features influencing the model's decision-making.

Chapter 6 discusses the preliminary results of a prococol that relies manga as culturally adapted nudging strategy and propose future improvements.
In Chapter 7, we discuss our contribution and propose future steps.

1.6 List of contributions

- Kobylyanskaya S, Popescu A.. The use of eye movement and disfluencies for difficult word decoding by L2 speakers, Gazing into Language, 2024
- Kobylyankaya S., Vasilescu I., Devillers L. Vers la compréhension des difficultés de lecture en L2 à travers des paramètres acoustiques et de mouvement des yeux, Conférence sur les Environnements Informatiques pour l'Apprentissage Humain (EIAH), 2023
- Kobylyanskaya S. Speech and eye tracking features for L2 acquisition: a multimodal experiment, Artificial Intelligence for Education (AIED), 2022
- El Baha M., Augereau O., Kobylyanskaya S., Vasilescu I., Devillers L. Eye Got It : a System for Automatic Calculation of the Eye-Voice Span. International Workshop on Document Analysis System (DAS), 2022
- Hutin M., Kobylyanskaya S., and Devillers L Nudges in Technology-Mediated Knowledge Transfer: Two Experimental Designs UbiComp/ISWC Adjunct, 2023
- Vargo A., Iwata M., Hutin M., Kobylyanskaya S., Vasilescu I., Augereau O., Watanabe K., Ishimaru S., Tag B., Dingler T., Kise K., Devillers L., and Dengel A. Learning Cyclotron: An Ecosystem of Knowledge Circulation. UbiComp/ISWC Adjunct, 2023

Chapter 2

State of the art

In this chapter we will explore in detail the research background relevant to the thesis. In particular, we start by defining the concept of L2 proficiency and addressing the challenges of its automatic assessment and prediction. Next, we present a range of speech features that are described in the litterature as relevant to evaluate L2 fluency, accentedness, comprehensibility and intelligibility both at segmental and suprasegmental levels. Then, we describe French and Japanese phonetic and phonological systems, compare them, and discuss how L1 specificities influence English L2 pronunciation. In a multimodal perspective, we also provide an overview of the use of eye tracking in L2 acquisition research, in particular L2 proficiency assessment. Finally, we review relevant studies demonstrating an efficient use of nudging technologies in education and L2 learning.

The Common European Framework of Reference for Languages (CEFR) (of Europe, 2001) defines 5 levels of foreign language skills: reading comprehension, listening comprehension, written and oral production, and interaction. For our study, we rely on the skills concerning oral production and reading comprehension, because these skills are the more relevant ones given our research protocol based on a reading aloud setup, as it will be presented later.

At each level of L2 proficiency, the learner must demonstrate the acquisition of a set of linguistic and conversational skills defined by the CEFR. These skills include the mastery of grammatical and syntactic structure, the phonological system, and the lexical inventory enabling the learner to express themselves on topics corresponding to their level of proficiency (of Europe, 2001). For example, at beginner level, the learner's aim is to be able to express themselves and be understood in everyday situations (ibid), while at advanced levels (C1-C2) they should be able to express and understand subtle nuances of meaning.

Pronunciation quality can also indicate the speaker's level of language mastery. At more advanced levels, the learner's speech is more comprehensible to native speakers, more fluent (with fewer disfluencies such as pauses, hesitations, repetitions), and the speaker can demonstrate their mastery of intonation to express the detailed meaning.

Language level can be assessed by standardized international exams, such as TOEFL, TOEIC, CAE, and IELTS for English. These exams are often expensive and usually require travel. Nowadays, in the post-COVID era, the question of online learning and certification, particularly of foreign languages, has become increasingly relevant. This is why, both industrials and researchers are trying to develop fast, effective and affordable tools for assessing L2 skills. These tools can be used for a variety of purposes:

- an assessment of pronunciation intelligibility, e.g. solutions proposed by Rosetta Stone[1] and Sanako Pronounce[2];
- as an assessment of language level for professional communication, e.g. IXL Learning[3];
- as an official certification, e.g. Educational Testing Center (ETS)[4], which offers L2 English tests worldwide. The center has implemented the SpeechRater (Hsieh et al., 2019) system to automate the level verification process, based on fluency measurements.
- as an estimate of level before taking the official exam. For example, Augereau et al. (2016a) proposed a text comprehension level assessment system based on eye movement measurements.

In the next parts of this Chapter, we will define the fluency based on existing research and highlight common practices that researchers use for automatic assessment. We will explore the characteristics that are the most relevant to fluency according to the scientific community. This analysis will guide us to select our own features in order to model fluency and to understand the challenges that automatic assessment represents.

2.1 Automatic L2 pronunciation and fluency assessment measures

Oral proficiency in L2 can be assessed through a variety of metrics from both the segmental and suprasegmental perspectives. Today, a rich literature highlights that L1 influences L2 at both the segmental and suprasegmental levels. However, speech fluency is a more global concept that includes, for instance, strategy in the use of pauses, a given speech rate and rhythmic patterns which influence accentedness, as well as measures of comprehensibility and intelligibility. Pronunciation is reflected at segmental level only, but can also affect accentedness, comprehensibility and intelligibility at different degrees (for instance through high/low functional load).

In line with such empirical observations, we dedicate the following paragraph to identify through the description of the three target languages, that is English as L2 and French and Japanese as L2 the segmental and supra-segmental features that may play a role in fluency as vehiculated by the pronunciation specificities.

First, we will consider the segmental or phoneme level, observe the phonological differences between English, French

[1] https://fr.rosettastone.com/
[2] https://sanako.com/automated-pronunciation-grading
[3] https://www.ixl.com/company
[4] https://www.ets.org/speechrater.html

and Japanese languages, recall and/or formulate hypotheses about the production of English sounds under the influence of L1 based on the reviewed literature. Then, we will focus on suprasegmental or prosodic levels, with the same objective that is to underline the influence of L1 French and Japanese. Following this section, we will determine how the L2 speech characteristics reflected on both segmental and suprasegmental levels interact with listener's judgment and automatic assessment of speech comprehensibility, intelligibility and accentedness. Finally, we will consider speech fluency, which, as seen in the literature, is reflected mostly at suprasegmental level, as it is closely related to the use of speech rate and disfluencies.

It is commonly accepted that pronunciation inconsistencies can be caused by L1 that influences the production and the perception of the L2 sounds Flege (1995). Indeed, if we place ourseleves in the position of a L2 speaker, due to the phonological filter of our mother tongue and our articulatory habits, we tend to substitute the L2 sounds by those existing in our mother tongue and close to the target ones Flege (1995).

This work focuses on the effect of L1 Japanese and French on L2 English: in order to propose relevant hypotheses about how L1 will affect the L2 production in our corpora, we will start by describing the phonological systems i.e. consonantal and vowel inventories of French and Japanese languages and compare them to English ones. Then, we will examine the differences between these languages at suprasegmental level i.e. prosody, rhythm etc. Finally, we will explore how the segmental and suprasegmental differences are translated into L2 and influence the fluency, as well as speech intelligibility, comprehensibility and accentedness.

2.1.1 Segmental level

The segmental level of speech concerns phoneme production. In this section we thus examine vowel and consonant inventories of French, Japanese and English languages. We will specifically recall the most commonly discussed influences of L1 English on L2 French and Japanese in order to both build an efficient data acquisition protocol and to successfully detect relevant and implement implement with machine learning techniques.

Vowels

In this paragraph, we compare the vowel systems of the English, French, and Japanese languages. The figure **??** represents vowel trapezoids of the three languages. A vowel triangle or trapezoid is a visualization of a vowel space according to their acoustic characteristics, concretely the first formants F1 and F2. The vertical axis represents the degree of vowel highness or openness as well as the F1: the higher the F1, the lower/closer the vowel. The horizontal axis represents the vowel backness and the F2: the higher the F2, the more front the vowel.[5]

As shown in Figure 2.2, Japanese language is characterized by 5 vowels, French has 15 vowels (Figure 2.1), 4 of which are nasal. As for English, it has 11 monophthongs (represented by unique sound) including tense and lax vowels, as well as 8 diphthongs (containing two sounds) (Figure 2.3). The distinctions into monophthongs vs diphthongs and tense vs lax vowels does not uniformly characterize the three languages. One can hypothesize

[5] https://corpus.eduhk.hk/english_pronunciation/index.php/2-2-formants-of-vowels/

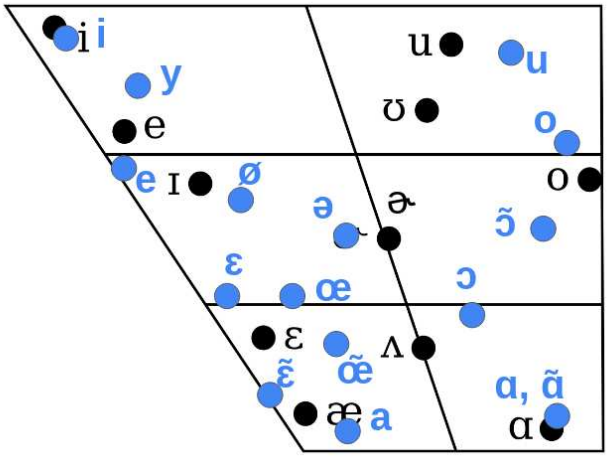

Figure 2.1: Vocal trapezius English(black) vs French (blue)

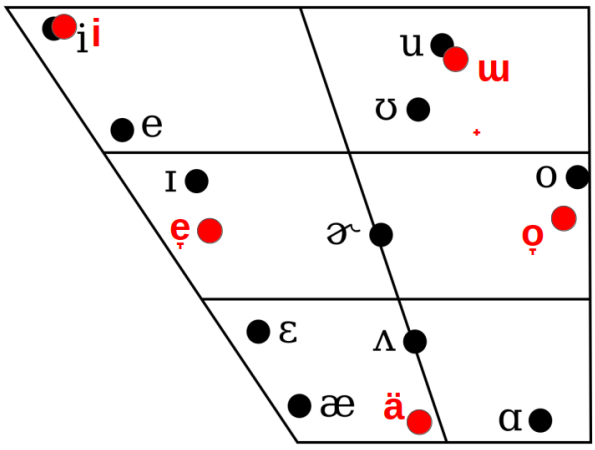

Figure 2.2: Vocal trapezius English(black) vs Japanese (red)

Table 4 – English diphthongs (according to Cruttenden, pp. 119–128)

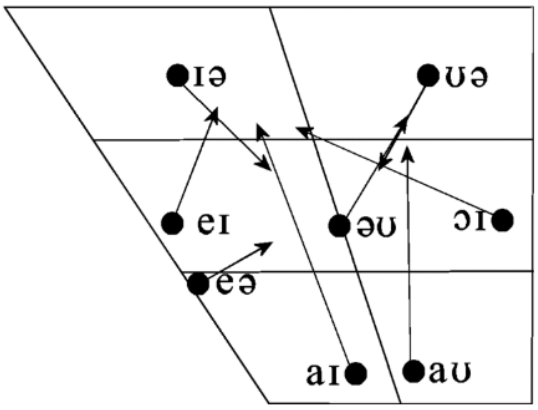

Figure 2.3: English diphthongs

English	i	ɪ	ɛ	e	æ	ɑ	ʌ	ə	ɚ	ɔ	o	ʊ	u
French	/i/		/e/, /ɛ/		/a/, /e/	/a/	/a/, /œ/, /u/	/ø/		/o/		/u/, /y/	
Japanese	/i/		/e/		/a/, /e/	/a/		/a/, /o/		/o/		/u/	

Table 2.1: English vowel substitutions by French and Japanese speakers as highlighted by state of the art.

that both French and Japanese speakers may tend to substitute and/or approximate English vowels by those from their L1 and in particular make no or little difference between the tense and lax vowels, as this distinction is absent in both French and Japanese. In Table 2.1 we propose a potential substitution of English vowels by French and Japanese speakers based on the vocalic specificities of their mother tongue and mentioned in the literature (Riney and Anderson-Hsieh, 1993; Capliez, 2011, 2016) .

We hypothesize that with L2 level advancement, speakers will approach their vowel realization to those of the target target language. This hypothesis will be tested in Chapter 4.

Consonants

The consonantal systems of three languages are presented in the Tables **??**. The consonants present in one language and absent in another are color coded. For instance, in the Table 2.2 representing English consonantal system, the consonants are colored in red if they are absent in Japanese language, in blue if they are absent in French system and in violet if neither French nor Japanese systems do not have these consonants i.e. /θ/ and /ð/.

One of the most common challenges in English pronunciation by Japanese speakers is the pronunciation of the liquids /r/ and /l/ that both are realized as /ɾ/, which is a common cause of intelligibility drop (Pérez-Ramón et al., 2023). Moreover, the consonants /p^h/, /t^h/, /k^h/ are unaspirated. The consonant /f/ is pronounced as /ɸ/ and /v/ as /b/ (Pérez-Ramón et al., 2023).

Japanese does not have /ʃ/, instead it has /ɕ/ which is an allophone of /s/ before high vowels such as /i/ (the sequence /si/ does not exist) (Riney and Anderson-Hsieh, 1993; Kondo et al., 2015). Additional common pronunciation specificities of Japanese speakers were described by (Riney and Anderson-Hsieh, 1993): “A mora nasal convention-

	labial	dental	alveolar	post-alveolar	palatal	velar	glottal
nasal	m		n			ŋ	
plosive	p, b		t, d			k, g	
affricate				tʃ, dʒ			
fricative	f, v	θ, ð	s, z	ʃ, ʒ			h
approximant	w		l	r	j		

Table 2.2: English consonants (in blue - consonants absent in French, in red - consonants absent in Japanese, violet - consonants absent both in French and Japanese)

	labial	dental	alveolar	alveolo-palatal	palatal	velar	uvular	glottal
nasal	m		n	ɲ		ŋ	N	
plosive	p, b		t, d			k, g		
affricate			ts, dz	tɕ, dʑ				
fricative	f		s, z	ɕ, ʑ	ç			h
tap or flap			ɾ					
semivowel					j	w		

Table 2.3: Japanese consonants (in red - consonants absent in English)

ally represented as /N/ becomes /m/ before /p, b, m/, /n/ before /t, d, n/, and /ɳ/ before /k, g, ɳ/. Japanese also has a mora obstruent represented as /θ/, which is always realized as the same obstruent that follows it, creating a geminate (or "double") consonant. Only /ɳ/ and /θ/ can close syllables" .

Both French and Japanese languages have no interdental sounds such as /θ/, /ð/, so both speakers will have difficulties while pronouncing these sounds and will replace them by /s/ or /z/ respectively. In some cases, French speakers can also replace these consonants by /f/, /v/ respectively (Capliez, 2011). In addition, both French and Japanese speakers have troubles with the pronunciation of aspirated consonants /p^h/, /t^h/, /k^h/ that are realized as unaspirated ones. Another problematic consonant for French speakers is /h/, which can be omitted or inserted (Capliez, 2011).

2.1.2 Suprasegmental level

The suprasegmental or prosodic level represents high-level characteristics, such as stress, intonation, rhythm. These features differ among languages depending on their isochrony, meaning rhythmic patterns that allow to describe the languages as **syllable-timed, mora-timed** or **stress-timed** (Pike, 1945).

For instance, English is considered as a stress-timed language that is characterized by the equal length of stressed syllables, the length reduction of the other syllables and an approximately equal duration between each stressed

	labial	dental/alveolar	post-alveolar/palatal	velar/uvular
nasal	m	n	ŋ	**(ŋ)**
plosive	p, b	t, d		k, g
fricative	f, v	s, z	ʃ, ʒ	ʁ
approximant		l	j, ɥ	w

Table 2.4: French consonants (in blue - consonant present in French and absent in English)

syllable. The stressed syllables are highlighted by both intensity and duration. French is a syllable-timed language whose syllables are all of the same length, and the stressed syllables are highlighted by an increase in intensity (Capliez, 2011). Japanese is a mora-based language where the duration of every mora is equal, and the stress is expressed by pitch rather than intensity (Watanabe, 1988; Riney and Anderson-Hsieh, 1993). Despite the fact that the language isochrony has been challenged by some authors (for a review see (Capliez, 2011)), it allows to and to highlight the difference between L1 and L2 rhythmic patterns (Capliez, 2011).

Thus, one may hypothesize that as all the three languages have different rhythmic patterns, both Japanese and French speakers will have difficulties when placing an accent in English. French speakers tend to place the accent on the last syllable of the words both lexical and grammatical and attribute the same weight to all the syllables (Capliez, 2011). As for Japanese speakers, they also have difficulty identifying stressed syllables, because unlike English speakers, they use pitch to mark stress instead of vowel intensity or duration (Watanabe, 1988; Riney and Anderson-Hsieh, 1993; Kondo et al., 2015). In addition, the distinction between vowel duration before voiced and voiceless consonants can also be problematic for them (Riney and Anderson-Hsieh, 1993). At the level of phrasal accent, the Japanese system is based on monotonic accentuation: the accent is placed on each word via the pitch change, whereas in English the accent is placed on the last lexical word of the segment (of a rhythmic group, for example) (Avery and Ehrlich, 1992).

The Japanese syllabic system, unlike English, is marked by the absence of consonant clusters and the absence of occlusives and fricatives in the coda. Whereas the English system contains "47 consonantal clusters in initial position and 169 clusters in final position" (Prator and Robinett, 1986; Riney and Anderson-Hsieh, 1993). Thus, the simplification strategy used by Japanese consists in inserting epenthetic vowels between consonants (Riney and Anderson-Hsieh, 1993). However, the epenthetic vowel changes depending on the context: in consonantal clusters after /t/ or /d/ a Japanese speaker would tend to insert an /o/; after /ʧ/ or /ʤ/ they would use an /i/; in other cases after non nasal consonants they would insert an /u/ (Kondo et al., 2015). In some cases, "unvoiced clusters containing sibilants and occlusives in the final position of words, the phonological simplification strategy is different: the Japanese resort to reduction instead of insertion" (Saunders, 1987) cited (Riney and Anderson-Hsieh, 1993).

2.1.3 Features affecting speech fluency

In the previous section, we observed the differences at segmental and suprasegmental levels between French, Japanese and English languages and formulate hypotheses about direction of L1 Japanese and French on L2 language. In this section, based on previous studies we will estimate how such specificities affect **fluency**, as well as **accentedness, comprehensibility** and **intelligibility**.

According to the reviewed literature, spoken fluency can be defined as the speaker's ability to formulate utterances characterized by an appropriate use of speech rate, rhythm, pausing patterns and disfluencies, whitout limit-

ing the listener's comprehension. As for oral reading fluency, besides the mentioned characteristics, it also includes accuracy, automaticity and reading comprehension (Kuhn et al., 2010).

As for accentedness, intelligibility and comprehensibility, they are important parts of overall spoken proficiency assessment and are reflected at both segmental and suprasegmental level. In particular, "**intelligibility** refers to listeners' actual understanding of an utterance, while **comprehensibility** refers to their perception of how easy or difficult the utterance is for them to understand" (Kennedy and Trofimovich, 2008). As for **accentedness,** it refers to how far a speaker's pronunciation is from native production. Even though these aspects are correlated, a strong foreign accent does not necessarily decrease the intelligibility or comprehensibility scores given by raters (Munro and Derwing, 2006).

In the following paragraphs, we will provide an overview of features at segmental and suprasegmental levels that contribute to speech intelligibility, comprehensibility, accentedness as well as fluency.

Accentedness, comprehensibility and intelligibility

Segmental level

As discussed earlier, the mother tongue affects the realization of L2 segments and can lead to a substitution of L2 sounds by segments of L1. This substitution may cause intelligibility and comprehensibility issues in connected speech. However, literature shows that all segmental mispronunciations does not influence equally the intelligibility (Hsieh et al., 2019; Isaacs, 2013). In order to understand the influence of the substitution of L2 phonemes by those of L1, we appeal to the concept of **functional load**.

Functional load is a measure used to compare the functioning of a pair of contrasting sounds in a language. For example, the pairs [d] / [t] and [ð] / [θ] in English carry very different functional loads, because there are many words that can change meaning if one replaces [d] with [t] (e.g. 'bad' vs 'bat'), but there are fewer cases where swapping [ð] with [θ] may engender semantic ambiguity. In terms of functional load, this relation of segments vs semantic impact translated in the fact that [d] / [t] have a high functional load, while the functional load of [f] / [θ] is rather low.

According to (Munro and Derwing, 2006), confusions with a high functional load have a greater influence on the accentedness and comprehensibility of the utterance than those with a low functional load. The process can be related to the fact that mispronunciations involving high functional load sounds are more salient and prevent the access to the meaning by native speakers (Munro and Derwing, 2006).

The influence of phoneme substitutions on speech intelligibility is also dependent on the speakers' and listeners' L1. For example, (Pérez-Ramón et al., 2023)] conducted a study in which French native speakers were asked to rate the nativeness and the intelligibility of English words pronounced by Japanese native speakers. The words included consonants that were subject to the most frequent mispronunciation by Japanese speakers. The authors concluded that the remplacement of [f] by [ɸ] and [ɹ] by [ɾ] caused a significant loss in Japanese speakers' intelligibility perceived by French speakers. This is due to the fact that [f]/[ɸ] and [ɹ]/[ɾ] phoneme substitutions do not occur in the non native speech of French, as they have different challenges when pronouncing English sounds. In contrast,

such substitutions as [θ] vs [s] and [t^h] vs [t] did not affect speech intelligibility perception, as they are common among French speakers as well. However, when speakers share the same L1 as the listeners, their pronunciation is perceived to be as intelligible as that of native speakers (Bent and Bradlow, 2003).

The differences in functional load can influence both human and automatic processing of speech. For instance, (Renwick et al., 2016) analyzed the functional load of vowels in Romanian and its relation to the performance of automatic speech transcription systems. A large-scale corpus study in which pairs of vowels with high and low functional loads were merged in order to test the influence on the automatic transcription performance showed that merging vowels such as [ʌ] and [i] having a low functional load into a single phoneme does not cause a drop in automatic performance: "ASR experiments show that a 6-vowel system where the /ʌ/-/i/ contrast is removed by merging the two vowels perform as well as a classical 7-vowel system".

The lessons to be learned from this section pointed out that some phoneme substitutions worsen the performance of automatic as well as human speech processing. It concerns, for example, the substitution of /l/-/r/ (Popescu et al., 2024) which is a common characteristic of Japanese pronunciation of English and has a high functional load. With respect to the research questions and the objectives of this thesis, we will take advantage of such observations to model the impact of native phonetic systems on L2 English for both automatic pre-processing of the corpora we will describe in the next chapter as well for the modeling of pronunciation issues of French and Japanese speakers of English recorded for the purpose of this work.

Suprasegmental level

Several studies report the contribution of prosody to listener's judgment of foreign accent and intelligibility. For example, (Boula de Mareüil and Vieru-Dimulescu, 2006) examined the contribution of segmental and suprasegmental characteristics of speech to determine wheter a foreign accent can be rather defined as Spanish vs. Italian and found out that prosody has more impact on listener's judgment than phoneme characteristics. (Anderson-Hsieh et al., 1992; Hsieh et al., 2019) found similar results, saying that prosody was more correlated with pronunciation scores than segmental factors. The study of (Kang, 2010) points out the importance of prosodic factors such as pitch height, pauses, stress that account for speaker's comprehensibility and oral proficiency. Other studies highlight the role of lexical stress in speech intelligibility and comprehensibility (Field, 2005; Hahn, 2004; Hsieh et al., 2019). The research conducted by (Polyanskaya et al., 2017) points out the role of both speech rate and rhythm on the perception of the foreign accent with a greater emphasis on speech rhythm.

A study by (Saito and Akiyama, 2017) conducted specifically on Japanese speakers of English, also highlights higher influence of the suprasegmental factors on listeners' ranking than that of segmental ones.

Numerous studies point out the correlation between pitch variation and native language or second language proficiency (Busà et al., 2011; Zechner et al., 2009; Kang, 2012; Zimmerer et al., 2014). According to these studies, L2 speakers use less pitch variation than L1 or more proficient L2 speakers. (Zimmerer et al., 2014) suggests that

narrower pitch variation among less proficient speakers can be due to a lack of confidence and to a higher concentration on the production of the segments than overall intonation. (Peters et al., 2023) who explored speech features variation among bilinguals suggest that "decreased jitter and shimmer, as well as increased HNR, CPP(s), and mean f0, [...] can serve as indicators of increased cognitive load" when speaking a less-dominant language.

Speech fluency

Speech fluency is often considered along with speech intelligibility, accentedness and comprehensibility when providing assessment to L2 proficiency. Indeed, according to (de Jong, 2016), fluency related features characterize effectively the L2 mastering and are strongly correlated with human scores.

What exactly does speech fluency represent? According to (Tavakoli et al., 2005) there are three main types of fluency, such as **speed fluency,** that includes such features as articulation and phonation rate, **breakdown fluency** that includes disfluencies such as silent and filled pauses, **repair fluency**, the frequency of troncations and corrections.

In this chapter, we will provide an overview of speech fluency features from the point of view of these three categories.

Speed fluency

(de Jong, 2016) discusses the most frequently used features for fluency assessment, such as speech rate (number of syllables / total speech duration), articulation rate (number of syllables / phonation duration (without silent pauses))

We would naturally assume that speakers who speak faster are more advanced in L2 than those whose speech rate is slower. Indeed, the study by (Iwashita et al., 2008) confirms this hypothesis and claims that the speech rate is one of the most salient predictors of the speech fluency. However, the study conducted by (Munro and Derwing, 2001) points out that despite a correlation between proficiency in L2 and speech rate, both the speakers whose speech rate was very fast or very slow received lower evaluation by human raters.

There are also some peculiarities when calculating speech rate. (de Jong, 2016) underlines that more advanced L2 speakers tend to use less frequent words that are often longer and points out the necessity to take into account word frequency when calculating the number of syllables and speech rate. The author also highlights the difficulty of cross-linguistic analysis of speech fluency, for example when analyzing languages having different rhythmic patterns.

Breakdown fluency

As for the filled pauses, they can also be L1 and L2 level dependent, can vary in terms of the number of syllables, their duration, the place in which they occur as well as their role in the sentence.

For example, the study conducted by (Rose, 2017) highlights the difference of the use of **filled pauses** by native Japanese and English speakers depending on the L1 and the L2 proficiency (for Japanese). The findings concern especially the difference in acoustic characteristics (F1) of filled pauses among L1 vs L2 speakers, the use of mono-

phonemic or polyphonemic filled pauses as function of L2 proficiency, the difference of duration of filled pauses as function of the L1/L2 and independent on the general articulation rate. The studies by (Kang et al., 2010; Vasilescu and Adda-Decker, 2006) however suggest that the filled pausing patterns relate more to the individual speaking style rather than L2 proficiency. Moreover, (de Jong et al., 2015) highlights that in order to distinguish speaking style from L2 proficiency level, it is needed to take into account the performance in the speaker's L1. According to the author, adjusting the features with respect to the L1 production improved the results of L2 level estimation.

As for the **silent pauses**, differences between the occurrence within L1 vs L2 speech have been also noticed. According to (Davies, 2003), such differences refer not only to their frequency or their duration but also to the site of occurrence within a sentence or text: native speakers tend to pause more often between the clauses while non native speakers can pause inside the clause (Davies, 2003; de Jong, 2016; Tavakoli, 2011). Such findings illustrate the cognitive effort related to disfluency phenomena and is related to speech planning, reformulation (Tavakoli, 2011).

Repair fluency

Among the characteristics of the repair fluency, we can notice the frequency and the place of truncations, repetitions and repairs.

These features, however, contributed little to the L2 proficiency modeling compared to the features related to speed and breakdown frequency according to (Bosker et al., 2014; Hsieh et al., 2019). In Chapter 5, we will examine the role of repair fluency related features in L2 automatic assessment.

2.2 Eye-tracking in L2 studies

Eye tracking is a technique used to measure eye movements. By tracking and analyzing eye movements, researchers can obtain valuable insights into human behavior, physiology, psychology, perception and visual attention. Furthermore, eye movements can serve as an alternative method to interact with the environment and digital interfaces. Recently, eye tracking has been used in the field of applied linguistics, psycholinguistics and L2 assessment that we will discuss later in this chapter.

Eye tracking provides a wide range of metrics. In this section, we will focus on the most commonly used ones.

Fixations refers to the moments when a person's gaze stops at a region of interest (ROI) which can include objects, words or groups of words. Fixations are typically measured in milliseconds or seconds and can be divided into three categories:

- *First fixation duration*: the duration of the first fixation that a ROI receives when the participant sees it for the first time.
- *First pass:* the sum of the durations of fixations on a ROI before the gaze moves to the next word.
- *Total pass*: the sum of all the fixations on a ROI.

These metrics are useful for measuring the cognitive load associated with a ROI. The more frequent and longer the fixations on a ROI, the greater cognitive effort it requires (Conklin et al., 2018).

Saccades are rapid eye movements that occur between two fixations and can move forward or backward. The **forward saccades** correspond to a typical movement from the left to the right seen in horizontal reading. While **backward saccades**, or **regressions**, are related to re-reading or moving to the next line. They can be associated with cognitive effort due to comprehension difficulties or text ambiguity (Conklin and Pellicer-Sánchez, 2016).

Saccades are usually measured in terms of their length, duration, and acceleration.

Skipping rate refers to the percentage of words that do not receive a fixation. According to (Rayner, 1998) about one third of the words are skipped during silent reading, and the skipping rate is influenced by the frequency (Roberts and Siyanova-Chanturia, 2013), contextual predictability, and word length (Rayner et al., 2011). The more predictable the word is, the more likely it is to be skipped. For example, grammar words are frequently skipped due to their length and high predictability.

The skipping rate is also influenced by the speaker's reading experience Esteve et al. (2024). Experienced readers can make predictions concerning the upcoming words relying on their practice. In contrast, less experienced readers need to focus on the words individually in order to access their meaning and phonological representation. We suppose that in our corpora, more proficient L2 readers will tend to skip grammar and other predictable words more often than beginners.

Pupil dilation is a process in which the pupil enlarges due to some stimuli. For example, it can reflect emotional states or cognitive load (Perkhofer and Lehner, 2019). However, it can depend on external factors, such as light exposure, so the experimental conditions require a closer attention (Hu and Aryadoust, 2024; Holmqvist et al., 2011).

There are multiple other eye movement features that can be measured and used in L2 acquisition studies, but we presented the most basic ones. These characteristics can be measured either overall (mean and standard deviation per task), or locally (by line, word, object). Overall measures provide insights into general reading behavior, while local metrics offer more detailed information, such as which words are skipped or fixated on, and for how long.

Eye tracking has been increasingly used in L2 research, in particular in reading that represents 15.3% of L2 eye-tracking studies (Hu and Aryadoust, 2024). It can provide important information about speech planning (Huettig et al., 2011), grammatical processing Frenck-Mestre (2005); Cunnings (2017), emotions Lallé et al. (2021), vocabulary learning Pellicer-Sánchez et al. (2022) and phonology Ito et al. (2018).

It allows to differentiate reading strategies of native from those of non-native readers (Kang, 2014), to detect involvement in the reading task (Kunze et al., 2015) as well as text comprehension level (Augereau et al., 2016a).

These metrics can illustrate the text's difficulty for the reader. Indeed, when the text is challenging to understand to the reader, their eye movement behavior will be characterized by longer and more frequent fixations (as they will spend more time at words) and shorter saccades (as they will move more frequently from one fixation to another)

(Conklin et al., 2018; Aryadoust et al., 2022; Révész et al., 2022). The metrics can also help identify the skills of the reader, such as L2 proficiency level (Augereau et al., 2016a; Berzak et al., 2018). Additionally, eye tracking features can be useful to anticipate reading challenges, for example to predict unknown words (Garain et al., 2017; Takaike et al., 2023).

The eye movement behavior may change depending on the support type: for example, when reading the text horizontally from left to right, when looking at an image, when reading comic books combining both images and texts and requires reading from right to left from top to bottom. It is also dependent on the language's writing system. For example, in Chinese the sense is quite densely represented compared to English. Although both readers of Chinese and English have similar fixations duration and frequency, the saccades of Chinese speakers are shorter than of English speakers (Feng et al., 2009; Conklin et al., 2018). (Siegelman et al., 2022) collected a multilingual eye tracking corpora (13 languages) and provided a cross-linguistic analysis of eye tracking behavior. The most salient language distinguishing factor concerns the skipping rate which depends on the average word length used in the language. For example, the participants reading in languages with longer words (such as Turkish or Finnish), tend to skip less words while reading. On the contrary, the participants reading in languages with shorter words (such as Korean or Hebrew) tend to skip words more frequently.

Eye movement behavior also changes depending on the reading modes: silent or oral reading. In oral reading compared to silent reading, participants tend to read more slowly and require more effort for comprehension (Takahashi and Kiyokawa, 2011; Vorstius et al., 2014). Consequently, they use longer fixations. This is particularly true for less skilled readers, who read orally at a considerably slower pace (Vorstius et al., 2014). In oral reading, participants also use regressions for rereading less frequently due to the need to coordinate their voice with eye movements (Vorstius et al., 2014). Indeed, oral reading can be more challenging than silent reading because of the number of cognitive processes involved: meaning decoding, phonological decoding, prosody, and eye-voice coordination.

The literature review reveals that eye-movement can bring significant information to the analysis of learners' level and text comprehension, and we believe that it can enrich spoken characteristics. Combining these metrics (speech and eye-tracking) during the read-aloud task is a promising direction, as it enables both conscious and unconscious processes to be captured (Godfroid and Hui, 2020). In the thesis, eye tracking is considered in correlation with speech in order to understand the read-aloud strategies of English learners with different language levels and provide them with specific help for their reading difficulties.

2.3 L2 learning modeling with machine learning approaches

In recent years, predictive machine learning techniques such as Random Forest (Kobayashi and Abe, 2016), Support vector Machine (SVM) (Yoon et al., 2009), Neural Network (Gretter et al., 2019) and more recently transformers for

audio (Baevski et al., 2020) and text (Devlin et al., 2019) have been increasingly used in the field of L2 assessment. Machine learning algorithms have been used for automatic L2 proficiency assessment, for example to facilitate the task of evaluators when a large number of test takers provide a significant amount of work to be evaluated. In this section, we provide an overview of studies using machine learning techniques and specifically pronunciation features with respect to segmental and supra-segmental levels, as well as eye tracking and multimodal features, allowing to account and model the L2 level.

For example, English Testing Service (ETS) uses SpeechRater including a variety of hand-crafted features for automatic fluency and pronunciation assessment (Hsieh et al., 2019). The implemented feature set includes features related to fluency and pronunciation at segmental and suprasegmental levels. Fluency features are also considered and they correspond to three types of fluency aspects such as breakdown fluency (filled and silent pauses), speed fluency (speaking and articulation rate) and repair fluency (repetitions and interruptions). As for segmental features, they measure the distance between speakers' phoneme realization with the reference native speakers' production using ASR. In addition, they measure suprasegmental characteristics, such as stress (percent of stressed syllables), rhythmic patterns (distance between stressed syllables) and other features.

With the same idea of modeling the segmental and supra-segmental features allowing to account for L2 level (Kyriakopoulos et al., 2019) investigated spoken proficiency assessment using hand-crafted rhythmic patterns (such as proportion of vocalic/consonantal segments, standard deviation of the duration of vocalic/consonantal segments, variability between successive measures) and those captured by a deep learning model. The authors highlight that rhythmic features extracted using a deep learning approach outperform the hand-crafted features for L1 automatic assessment. However, when rhythmic features are combined with pronunciation characteristics, specifically phone distance described in (Kyriakopoulos et al., 2018), both hand-crafted and deep learning approaches are efficient.

The study of (Gretter et al., 2019) demonstrates the use of ASR and feed-forward neural network for students' L2 proficiency assessment. Over 100 features were extracted based on the output provided by an ASR system, including the calculation of the sentence probability, out-of-vocabulary words, as well as acoustic features, such as length of utterance, silence duration, edit distance between the phonetic ASR outputs for native and non-native speakers, as well as other features. In a similar vein, the study by (Bannò et al., 2022) evaluates L2 proficiency assessment performance provided using three machine learning models: 1. wav2vec-based (Baevski et al., 2020) model which does not require transcription input provided by an ASR system to two other models: 2. a BERT-based (Devlin et al., 2019) model requiring ASR transcription and word embeddings 3. Deep Density Network trained on hand-crafted features including statistics representing phone realization, rhythmic patterns, fluency and intonation as in (Malinin et al., 2017). The models present comparable performance for L2 assessment, however neither BERT-based model nor the wav2vec-based one take into account all the aspects of L2 oral performance and are not interpretable for providing feedback. However, the authors suggest that the models are complementary and provide better performance when combined. A step forward is proposed in (Phan et al., 2024). The authors propose a new approach

including the combination of ASR with Large Language Models (LLM) such as GPT-4 (OpenAI et al., 2024) for spoken proficiency detection and feedback generation. To sum up, with respect to speech features aimed at automatically modeling of L2 proficiency, numerous studies have been published in recent years that encourage us to consider the segmental and supra-segmental features as reliable indicators of L2 proficiency level and provide valuable ideas for our own approach in terms of features selection and automatic design for prediction.

As for the eye tracking features, the litterature underline some recent progress. For instance reading comprehension can be assessed using eye-tracking features fed into a machine learning model, as it have been underlined in (Augereau et al., 2016a; Copeland et al., 2014). Eye tracking characteristics can include those related to blinks, fixations, saccades, their length and duration, etc. According to the study of (Augereau et al., 2016a), total reading duration, saccades number and velocity, as well as the number of blinks contribute the most to the model's decision-making. The study of (Berzak et al., 2018) aiming at predicting L2 proficiency based on reading behavior, relies mostly on fixation-related features and their correlation with the word type, its frequency and the context in which the word appears. (Garain et al., 2017) proposes a method for identifying difficult words while reading based on eye-tracking behavior correlated with word length and frequency. This approach shows good performance and paves the way into personalization of learning content based on individual difficulties. In addition to L2 proficiency level detection and reading content personalization, machine learning can be employed for L2 learners' skill improvement (see (Wu, 2024) for a review). For example, it can be used to the adapt listening tools and to improve oral comprehension by personalizing the captions (Mirzaei and Meshgi, 2023), writing skills by employing computer-mediated corrective feedback (Mohsen, 2022), to improve speaking skills and willingness to communicate (Fathi et al., 2024). Additionally, machine learning based tools can be used as an alternative learning method to decrease learning anxiety and cognitive load (Chen et al., 2022). Tu sum up the eye tracking automatic modeling, recent litterature underline the correlation between eye tracking and comprehension and the modeling of the levels of comprehension have been an objective for a number of studies using machine learning approaches.

The literature highlighted above underline methods that proved their effectiveness in L2 learning and assessment. However, one can notice that there is still a need for rule-base systems and hand-crafted feature extraction in order to gain interpretability and transparency into system decision-making, which can be beneficial for pedagogical goals (Lee and Lee, 2024). Besides, to the best of our knowledge joint approaches that combine connected speech as in real-life situations and eye tracking are lacking. The approach we will describe later in this thesis relies on hand-crafted speech and eye tracking features that showed their effectiveness in the detection of L2 oral fluency level and classical machine learning algorithms, such as Random Forests. Indeed, with respect to our work hypotheses and objectives, the explainability of extracted features and algorithm's decision allow us to to get insights into pronunciation difficulties and reading strategies implemented by the participants of our reading aloud experiment

described in Chapter 3. Moreover, it allows to capture measurable differences between the production of speakers having different linguistic and cultural backgrounds.

2.4 Nudges in education and L2 learning

Nudges are indirect, influential strategies derived from economic theory (Thaler and Sunstein, 2008) that aim to modify behavior without restricting choices. In recent years, nudges have become ubiquitous and are now applied beyond the field of economics, including in education, particularly in second language (L2) learning. (Damgaard and Nielsen, 2018; Hutin et al., 2022) provide extensive literature reviews on the types and examples of nudges used in education, while (Neuhaus, 2021) specifically focuses on nudges in L2 learning.

Various types of nudges have been shown to be effective in educational settings, as they help to improve motivation, set up learning goals, and facilitate knowledge acquisition. Common examples include setting deadlines (Ariely and Wertenbroch, 2002) sending reminders (Castleman and Page, 2015), providing personalized feedback (van Oldenbeek et al., 2019), and using social comparison in grading (Davis et al., 2017).

In the context of L2 learning, particular attention should be given to addressing errors and mistakes, as well as to providing appropriate feedback (McDonough, 2005; Neuhaus, 2021). Moreover, students self-identification with the learning material can enhance their sense of ownership, and consequently boost their motivation and increase the time they invest in learning (Neuhaus, 2021). This can be achieved by allowing students to personalize their learning materials, such as creating portfolios, texts, or videos. This personalization motivates students to improve their work and review it again, helping them connect more with the language and become more proficient (Neuhaus, 2021).

Some studies (Bovens, 2010; Selinger and Whyte, 2010) highlight the importance of cultural aspects in nudges effectiveness. Indeed, different cultures are subject to various perceptual and behavioral biases, which means that nudges can have different outcomes depending on the cultural context of people they are applied to.

In our study, we rely on multimodal and culturally adapted nudges including both linguistic and visual cues. This approach is realized by the use of manga, Japanese comic books which is a popular leisure mode in Japanese culture that has gained popularity worldwide, including France. The interest in Japanese manga and anime motivates people learn the Japanese language. Manga is frequently used in the classroom as a reading support not only for language learning (Unser-Schutz, 2011; Sarada, 2016), but also for other subjects such as history, economy, and others (Murakami and Bryce, 2009). Due to their use of images, speech balloons, modern vocabulary (including slang) and captivating stories, manga can enhance learners' motivation, attention and comprehension (Eneh and Eneh, 2008).

However, there are some limitations for the use of manga as a learning tool. For example, the study by (Arlin and Roth, 1978), suggests that comic reading improved reading time and comprehension only for skilled readers, while less proficient readers did not benefit as much, as they were distracted by the images and spent less time reading the

text, which led to poorer comprehension. Another study (Singer et al., 1973-1974; Arlin and Roth, 1978) challenges the effectiveness of comic books in vocabulary learning: students performed well when words were associated with images, but struggled to recognize the words in the text without images.

In our case, we put forward the hypothesis that the use of images and dialogues in manga could enhance readers' oral performance, for instance by improving prosody variation which tends to be flat among Japanese speakers as underlined in (Avery and Ehrlich, 1992) and by increasing engagement in the task.

2.4.1 Summary

To sum up, in this chapter, we provided a literature review on the differences between French, Japanese, and English languages, and discussed L1 influences on English L2 oral production. We also gave an overview of the segmental and suprasegmental features used by researchers to automatically assess L2 proficiency levels. Additionally, we explored the use of eye tracking in L2 assessment and the potential for correlating spoken and eye tracking features for this purpose. We have also reviewed recent studies that use machine learning to model L2 pronunciation and comprehension with respect to segmental, supra-segmental and eye tracking features. Finally, we addressed the use of nudges in education, particularly in L2 learning, and examined the combination of visual and linguistic cues for knowledge acquisition.

In the next chapter, we will present our data collection and processing protocol based on the reviewed literature. Later in this work, we will examine how both eye tracking and spoken features interact with L1 and L2 proficiency levels in our corpora, as well as the role of reading support in L2 oral production and eye movement behavior.

Chapter 3

Data collection and processing

In this chapter, we first focus on the existing corpora that partially meet the needs of this thesis and therefore the need to collect new data on L2 speech and eye movement, then provide a detailed description of our research protocol and data processing steps.

3.1 Existing corpora and the need of a new data collection

Available multimodal corpora that combine both connected speech and eye movement are still rare. Current datasets include eye-tracking data for silent reading in L1 or L2, speech data of spontaneous, guided and read speech, and a multimodal dataset that combines eye-tracking during reading aloud in L1. Therefore the need for new data collection may be questioned given the availability of existing speech and eye-tracking datasets used by the research community for L2 assessment and other purposes. In this section, we address this concern by providing an overview of existing datasets and discussing the purpose of a new data collection.

3.1.1 Eye tracking corpora

This section is dedicated to eye tracking corpora focusing on L1 and/or L2 reading and comprehension. We identified three corpora that partially meet our objectives:

1. **MECO (Multilingual Eye-tracking Corpus)** (Siegelman et al., 2022) is a result of a collaborative international project that collects eye-tracking data for silent reading in L1. MECO includes data from 13 languages representing diverse phonological, morphological, and syntactic systems, allowing for direct comparisons between different writing systems. The corpus also includes measures of comprehension, demographic information, and individual differences. The corpus however does not include reading in French and Japanese languages.

2. **CELER (Corpus of Eye Movements in L1 and L2 English Reading)** (Berzak et al., 2022) eye tracking dataset of silent reading in English. It includes data from participants having different English proficiency levels and linguistic backgrounds, such as Arabic, Chinese, Japanese, Portuguese, and Spanish. The dataset comprises eye tracking data from 365 participants: 69 are English native speakers and 296 are non-native speakers. The participants read sentences from Wall Street Journal articles.

3. **TECO** (Nahatame et al., 2024) is an eye tracking corpus that captures silent reading of English by Japanese speakers. The corpus includes data from 41 participants (26 women and 15 men) with an average age of 21, having intermediate and advanced L2 proficiency levels and reading texts of varying difficulty.

3.1.2 Speech corpora

As for spoken and multimodal (speech and eye tracking) corpora the current state of the art is at follows.

Spontaneous and guided speech

1. **CLIJAF** (Detey, 2011–2019; De Fino et al., 2022) is a speech corpus featuring Japanese speakers of French, including data from 38 participants engaged in spontaneous speech and semi-directive interviews resulting in 26h of orthographically transcribed speech. The corpus includes participants' proficiency level in French as a second language (L2) as assessed by professional teachers of French as a Foreign Language.

2. **NICT JLE** (Izumi et al., 2004) is a corpus of Japanese learners of English. It consists of transcripts from 1,281 audio-recorded English oral proficiency interview tests, totaling 1.2 million words and 300 hours of data. It includes proficiency levels as well as two annotation levels: basic tags and error tags. Basic tags, over 30 in total, cover aspects such as interview structure, interviewee profiles, speaker turns, and utterance phenomena like fillers, repetitions, and self-corrections. Error tags focus on grammatical and lexical errors and consist of 47 tags. The corpus also includes data from English native speakers in order to provide a baseline for a comparative analysis.

Reading aloud UME-ERJ [1]is an English Speech Database Read by Japanese Students. It includes a variety of sentences that are read aloud by 202 speakers (100 male and 102 female), in particular sentences designed for learning phonemic pronunciation with the focus on phoneme clusters difficult to pronounce for Japanese speakers, as well as sentences for prosody learning, including sentences with different intonational and rhythmic patterns.

3.1.3 Multimodal data: eye tracking and reading aloud

We have identified a unique corpus that comes close to our objectives.

ReadLet (Ferro et al., 2024) is a corpus featuring silent reading and reading aloud in Italian as a first language, pro-

[1]Speech Database Committee of the Priority Areas Project on "Advanced Utilization of Multimedia to Promote Higher Education Reform" during 2000-2002. https://doi.org/10.32130/src.UME-ERJ

duced by both children and adults. The corpus includes multimodal data such as speech recordings, eye tracking, and finger tracking. The children's group consists of 44 males and 50 females aged 7 to 12 years, while the adult group comprises 27 males and 28 females with an average age of 27 years.

3.1.4 Summary

To sum up we identified and presented 8 corpora that partially meet our objectives. All corpora are valuable for analyzing L1 and/or L2 speech and eye movements: they provide reliable information with respect to data acquisition and processing, annotation and useful cues to explore in line with our own objectives. More largely, they provide a foundation for understanding the challenges of L2 proficiency assessment, the influence of L1 and L2 level on eye movement during silent reading and speech production.

Among the reviewed corpora, only the ReadLet corpus combined spoken and eye-tracking data during a reading aloud setup. However, this corpus, released in 2024, involves L1 Italian speakers, which does not align with our specific case study. At the time our research began, there were no available datasets that combined both speech and eye movement modalities. Additionally, none of the existing datasets included different types of reading support that incorporate both visual and linguistic cues, as we do by including manga in our study.

In order to address this gap, we designed a research protocol specific to our needs, aiming to compare oral reading strategies between representatives of two distinct cultures: French and Japanese. Our goals included evaluating the contribution of both spoken (phoneme and prosody realization) and eye-tracking features to L2 fluency assessment, comparing the contribution of both linguistic and visual cues for speech production and eye movement behavior, and examining the effect culture-specific nudges on French participants when adapted to Japanese culture.

It is important to mention that the data collection process was the most rewarding part of my thesis, as it allowed me to meet many people with different backgrounds, to travel to Japan for three months, to gain a deeper understanding of Japanese culture, and to learn a new language. It also provided an opportunity to strengthen our relationships with our Japanese partners and paved the way for future collaborations.

3.2 Data collection

To meet the research questions and the goals of the LeCycl project a multilingual corpus consisting of simultaneous recordings of speech and eye tracking from three populations of speakers was designed and implemented. In this purpose multiple data collection sessions were required. These sessions included recording participants from two countries while utilizing different reading support types. Data collection was partially carried out during the COVID

	Protocol 1	Protocol 2	Protocol 3	Protocol 4
Team(s)	LISN-CNRS	OMU	LISN-CNRS + OMU	LISN-CNRS + OMU
Country	France	Japan	Japan	France
Nbr participants	50 French + 12 English native speakers	20	24	14
Reading support	texts	manga	text + manga + manga without images	manga
Duration	30 min	3h	1h: text + manga without images 2h: manga	1h
Total nbr hours of collected speech	French speakers: 4h English speakers: 1h	45h	text: 2h manga without images: 4,5h manga: 36h	10,5h

Table 3.1: Research protocols and collected data as function of experimental set up and L1.

pandemic and required a 3-months mobility at Osaka Metropolitan University, Japan in the team and under the supervision of the Professor Koichi Kise. The mobility allowed collecting data under the same conditions as in France. 12 native English speakers were recruited and recorded in France (Parisian region).

Finally, all data collection protocols that will be described in the following paragraphs have first been approved by the Paris Saclay University's ethics committee.

In this section, we will describe four experimental protocols, which are presented in Table 3.1, with particular focus on Protocols 1 and 3 as they form the foundation of this thesis.
First, we will briefly present the differences between these protocols. Then, in the following sections, we will provide a detailed description of the datasets obtained, the equipment used, and the tasks completed by participants during the experiments.

Protocol 1 was conducted in France and included French speakers of English and native English speakers (mostly American English), who were asked to read aloud texts in English. At the same time, Protocol 2 was conducted in Japan by our Japanese partners, where Japanese speakers were asked to read comic books aloud. The data from this experiment was later used for unknown word prediction (Takaike et al., 2023). Protocol 3 was carried out during my stay in Japan and represents a combination of the first two experiments. Here, Japanese speakers were asked to read aloud texts used in Protocol 1, as well as manga, similar to Protocol 2. Additionally, we selected an extract from a manga that participants read both in its textual form and in its original format, including images. This approach allowed us to examine the role of images in reading aloud production. The fourth Protocol was conducted in France, where French speakers read manga aloud. It represents an adaptation of Protocol 2 to the French readers. In total we recorded 120 subjects for a total of (more than) 100 hours of speech and eye tracking data.

We will now discuss in detail the data acquisition process, with a focus on Protocol 1 and 3.

Level	French		Japanese	
	male	female	male	female
A2	1	2	0	0
B1	2	8	9	3
B2	8	6	7	4
C1	13	8	0	1
C2	2	0	0	0
Total	**26**	**24**	**16**	**8**
	50		**24**	

Table 3.2: Number of French and Japanese participants: Protocols 1 and 3

3.2.1 Participants

The participants are 50 French (Protocol 1), 24 Japanese (Protocol 2) adults (> 18 y.o.) speaking English as a foreign language and 12 native English speakers (mostly American English) (Table 3.2). The data collected data from native English speakers allows us to provide a baseline for the data analysis.

French and Japanese participants had varying proficiency levels according to the CEFR scale. The majority of the French participants had B2 or C1 levels, only few of them had A2 and C2. As for Japanese participants, almost all the participants had B1 and B2 levels, except for one participant who had a C1 level. The participants were recorded in their natural setup: in France and Japan, respectively. They were mainly recruited among Paris Saclay and OMU university students and recorded in corresponding conditions in the laboratories (that is quiet office dedicated to the experiment). In France, the proportion of men and women is nearly equal, with 26 men and 24 women, while among the Japanese speakers, men represent the majority, with 16 men and 8 women. This gender imbalance can be explained by the fact that the majority of the students of the Department of Informatics, where the study was conducted, are male. Moreover, we obtained a dataset twice as large for the French speakers (4h) compared to the Japanese speakers (2h). This is due to time constraints faced during my visit in Japan, including the time required to update the protocol for the Japanese audience and to establish contact with the students and invite them to participate.

Despite the fact that the dataset is not perfectly balanced in terms of L2 proficiency levels and the number of French and Japanese speakers, the data analysis allowed us to highlight relevant features that will be discussed in the next chapter.

3.2.2 Reading support

In both experiments of Protocols 1 and 3, French and Japanese native speakers were asked to read aloud three texts in English. All the participants read the same texts. In Protocol 3, Japanese speakers read both manga and texts. In this section, we will motivate the choice of the reading material.

Texts

In order to conduct a cross-cultural and cross-linguistic analysis of oral reading among French and Japanese speakers, we asked the participants to read 3 texts (see Appendix A) of varying complexity levels. The texts were selected on the *English For Everyone*[2] website dedicated to learning English. We found this site to be a reliable source of data, as the texts are unique, written by professional teachers specifically for this resource, and adapted to different comprehension levels. The texts are accompanied by comprehension questions, which we used as a post-reading activity. The texts were validated by the two teams concerned by the protocol in LeCycl project, that is LISN and OMU team, the latter validating the consistency with previous silent reading protocols.

The website presents three basic levels of text complexity: beginner, intermediate, and advanced, each subdivided into low, mid, and high levels. The levels with the prefixes "low-" and "high-" are similar to the adjacent levels (below or above the declared level): for example, low-intermediate texts are similar in complexity to high-beginner. Therefore, we have selected the "mid-" levels , as they are the most representative of the declared level.

The texts are categorized by type (e.g. dialogue, narration, informative text, etc.), subject (technical subjects, stories), and level. The following criteria were utilized to select the texts:

- Levels: mid-beginner, mid-intermediate, mid-advanced
- Word count: <350 words for advanced, <300 words for intermediate, <200 words for beginner.
- Format: "Short stories", as texts in this format contain different types of discourse, such as narrativized and direct discourse. Presumably, more intonation variants could be observed due to the variation in discourse type.
- The variety of sentence types: declarative, interrogative, exclamatory, which could also contribute to intonation variation.
- Vocabulary: the absence of complex numbers and uncommon proper nouns, whose knowledge does not generally illustrate the level of mastery of L2 and which could increase the complexity of the read-aloud task and provoke more disfluencies (pauses, hesitations, etc.).
- Measures of text readability: must correspond to the level in question.
- Large phonetic inventory: the selected texts represent the majority of phonemes of English phonological inventory: except for the diphthongs /eə/, /iə/, /ʊə/ and the consonant /ʒ/

Readability measures

To ensure that the level of text complexity corresponds to the declared level, we calculated their readability measures. Following the approach used in the article (Novikova et al., 2019), we used the Lexical Complexity Analyzer

[2] https://englishforeveryone.org

Text	Level	Nbr word types	Nbr token types)	Nbr sophisticated token types	Type / token ratio
A happy visitor	Beginner	55	176	9	0.3
Time with Grandpa	Intermediate	133	242	31	0.5
Fried	Advanced	143	237	46	0.6

Table 3.3: Lexical characteristics of the selected texts

Text	Level	Nbr words	Nbr phrases	Nbr clauses	Average phrase len	Nbr subord clauses
A happy visitor	Beginner	177	41	39	4.3	1
Time with grandpa	Intermediate	241	20	30	12	9
Fried	Advanced	237	24	25	9.9	5

Table 3.4: Syntactic characteristics of selected texts

software (Ai and Lu, 2010; Lu, 2012) to calculate lexical and syntactic measures of the selected texts. This software extracts "25 measures illustrating lexical variability, lexical density, lexical sophistication" (Ai and Lu, 2010).

Lexical density (Ure, 1971 as cited in Ai and Lu, 2010) is a measure calculated as the ratio between lexically full words (as opposed to grammatical words) and the total number of words in the text. *Lexical sophistication* (or sparsity) is a set of measures used to identify the proportion of rare or advanced words in the text (Read, 2000 as cited in Ai and Lu, 2010). *Lexical variation,* also known as *lexical diversity* or *lexical range* (Malvern et al., 2004, as cited in Ai and Lu, 2010), identifies the range of words used/known by the speaker. (Ai and Lu, 2010) implements 19 measures for this calculation.

With regard to syntax, the software calculates "the number of production units, the number of coordination and subordination relations, the degree of sentence sophistication and the overall complexity of the sentence" (Ai and Lu, 2010).

The analysis demonstrates the correlation between text complexity at both syntactic and lexical level with the declared text levels. In the tables below (Table 3.3 and Table 3.4), we present examples of lexical (Table 3.3) and syntactic (Table 3.4) features for the selected texts. Generally speaking, the number of words, tokens, sophisticated words and the type/token ratio increase with the difficulty of the level. As far as syntax is concerned, the advanced text (“Fried”) is less in line with the overall trend: the number of words, phrases and subordinate clauses is lower than in the intermediate text (“Time with Grandpa”). On the other hand, the advanced text (“Fried”) stands out for its number of sophisticated words. These appear to be the most important parameters for identifying the text level as “advanced”.

Manga

Manga served as the reading material for Protocols 2 and 3. As discussed in Chapter 2, manga is a common reading support in Japanese culture, used both for leisure and learning of multiple subjects, including L2 acquisi-

tion. Manga often represents engaging stories and includes dialogues, contemporary vocabulary, short sentences of various types, such as declarative, interrogative, and exclamatory. Manga is designed to capture the reader's attention, contribute to reading comprehension, and engage learning motivation. In addition, it is adapted to Japanese speakers' culture and reading style: from right to left, from top to bottom.

The manga selected for this study are original Japanese manga translated into English by professional translators. The choice of manga books is explained by the goal of the study conducted by the Japanese team, which aimed to identify speech balloons in which unknown words occurred. For this reason, they empirically estimated the number of unknown words that Japanese participants might encounter while reading. The actual number of unknown words was reported by the participants after the reading part of the experiment.

The French team's objective differed and focused on evaluating the contribution of images to oral reading production. Therefore, a specific manga extract was selected and presented to participants in two formats: its original version with images (as in Figure 3.1 and as plain text (see Appendix B). The selected extract is part of the manga "Delicious in Dungeon" by Ryoko Kui, which is quite popular among Japanese audience. It is an adventure fantasy story about a group of explorers who want to save their friend from a red dragon and who cook monsters they encounter in the dungeon in order to survive. The story contains detailed descriptions and illustrations of cooking recipes and monsters. The selected extract describes the process of cooking a scorpion hot pot Figure 3.1, the text can be found in Appendix B).

The choice of a manga extract was based on linguistic criteria (phonology, lexicon, syntax). From a syntactic point of view, the manga contained mostly short sentences, without relative subordinates or complex verb tenses, as it represents interactive direct discourse. However, it included advanced level vocabulary, which made it comparable to the advanced text used in Protocol 1.

3.2.3 Equipment

For the experiments, the following equipment was used:

- laptop 14 inches
- microphone AKG Perception Wireless 45 Sports Set Band-A 500-865 MHz
- eye tracker Tobii Nano Pro conceived for research
- EyeGotIt software (El Baha et al., 2022)

EyeGotIt (El Baha et al., 2022) is an open-source software specifically designed for reading aloud experiments and the LeCycl project. It enables recording and processing of both audio and eye-tracking data. The medium connects the recording devices (microphone and stationary eye-tracker) and provides a comprehensive interface allowing to

Figure 3.1: Manga example (from "Delicious in Dungeon")

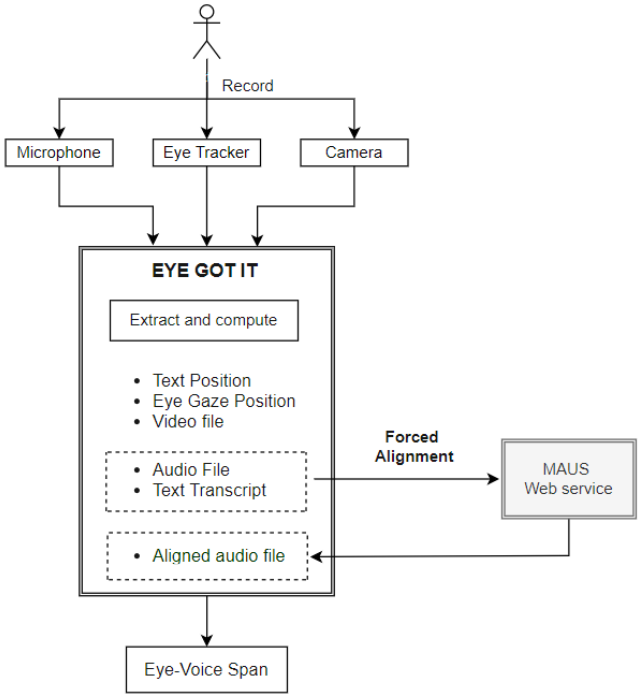

Figure 3.2: EyeGotIt system algorithm (El Baha et al., 2022)

display the texts to be read and to manage other reading related tasks, such as multiple choice questions, processes raw data and calculates eye-voice span (the distance between the pronounced and the fixated word) (Figure 3.2). The software is coded in Python and runs on the Windows operating system.

For eye movement recording, we used Tobii Nano Pro, an eye-tracking device that can be attached under the computer screen as shown on Image 3.3. It provides raw eye gaze data with their corresponding coordinates on the screen, calculated in pixels and the time at which the eye gaze occurred. The role of the EyeGotIt software is to calculate eye-movement features based on this raw data. First, it determines the word position on the screen and then, computes the fixations and saccades from the raw data. For this purpose it relies on two widely used algorithms developed by (Buscher et al., 2008) and (Nyström and Holmqvist, 2010). For our analysis, we relied on the algorithm developped by (Buscher et al., 2008), as the results were more precise. This algorithm calculates the fixations based on the proximity of neighboring eye gazes. For example, in the Figure 3.4, the cluster of red dots (eye gaze) forms a fixation, depicted by a circle. The circle diameter is proportional to the fixation duration: the longer the fixation, the larger the circle diameter. The lines between the circles are the saccades that represent rapid eye movements. The Figure 3.4 shows an ideal sample which is difficult to acquire in real experimental conditions. Indeed, the quality of the eye tracking data depends on multiple factors:

- the head movement: when the participant approaches the screen or the head exceeds the working zone of the eye tracker (the rectangular zone on the Figure 3.3.

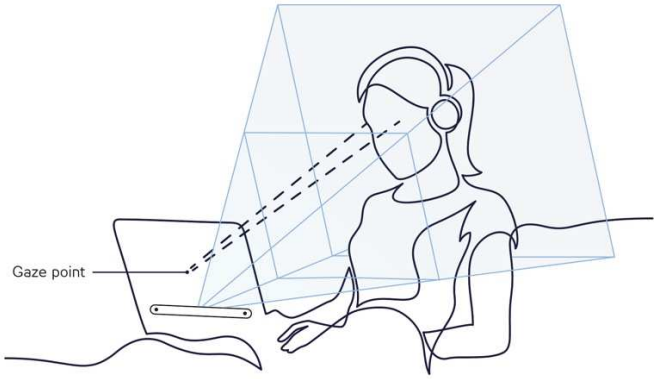

Figure 3.3: Eye movement recording using Tobii eye tracker

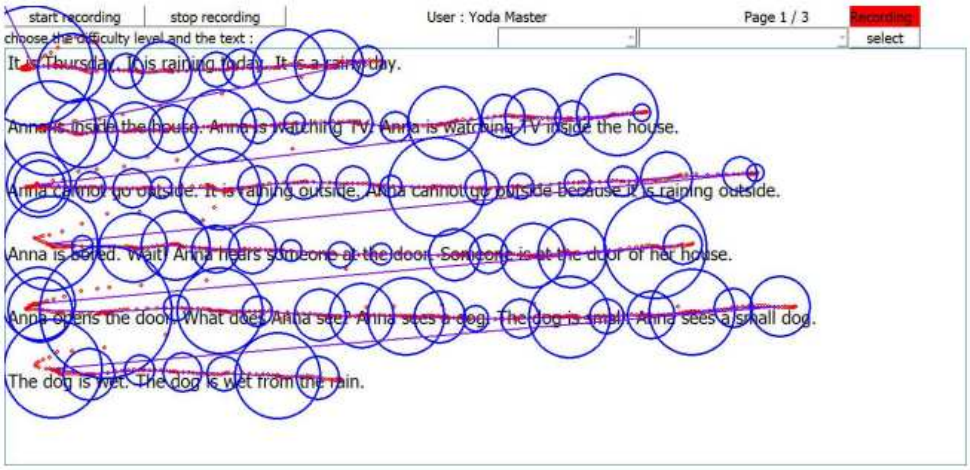

Figure 3.4: Eye movement visualization during a reading aloud task

- the light condition: sunlight or devices producing flickering light might be disturbing
- heavy eye makeup
- progressive glasses
- some eye conditions and surgery
- etc

3.2.4 Procedure

The experiment was carried out as follows. The recording took place in a quiet, but not soundproof room. The preparation phase takes around 10 minutes and includes signing data collection agreement, explaining the tasks, demonstrating the equipment and interface, and calibrating the eye-tracker.

Next, we proceed to the recording. The speakers read aloud three texts in English described in the section 3.2.2, starting with the beginner text and progressing to the more complex ones. After reading each text, they are asked to answer a comprehension questionnaire (taken from the same website as the texts) and to report difficult words, whether the difficulty was due to comprehension or pronunciation. We do not differentiate between the causes of word difficulty faced by the participants, but it is an improvement point that should be considered for future work.

Additionally, the participants are asked to complete a survey regarding their language background, for example, level according to their evaluation, an English exam score such as TOEFL, TOEIC etc., age at which they started learning English, duration of their stays in English-speaking countries, exposure to other languages, gender, mother tongue. This task can be completed independently on the experiment time and date.

The total time of the text reading experiment varied across participants and took 30-50 minutes depending on the time needed for equipment calibration and the participants' reading speed. All in all, 4h of speech data was collected for French speakers and 2h for Japanese speakers reading texts (see Table 3.1).

As for the experiment including reading manga, the participants reported unknown words and completed a survey about their L2 learning background, but were not asked to answer comprehension questions. The part of the experiment where Japanese speakers read an extract of manga in its textual form lasted between 10 and 15 min and resulted in 1h of total time of collected speech. The part of the experiment in which the participants read manga in its original form (including the extract of manga read in textual form) lasted about 2h and was divided into 1h sessions: 45 minutes of reading time and 15 min of difficult word reporting. The manga reading part resulted in 36h of recorded speech (see Table 3.1).

3.2.5 Summary

To sum up, four data collection protocols were designed and implemented for the LeCycl project, two of them being the foundation of this thesis. We collected reading aloud data from 50 French, 24 Japanese and 12 native English speakers, using different reading supports: texts of varying complexity levels, manga in its original and textual form. The data collection completes a lack of multimodal datasets including both eye-tracking and audio data, different text types and cross-cultural aspects. The collected dataset allows us to compare reading strategies of participants with different linguistic and cultural background, to evaluate the role of visual cues adapted to Japanese culture on oral reading performance of both Japanese and French participants. In the next chapters we will discuss the data processing techniques, as well as the results obtained on the presented data.

3.3 Data processing

In section3.2, we described the data collection procedure required for the implementation of our research protocol and addressing the research questions. We collected eye tracking and speech data, as well as metadata, such as participants' L2 proficiency and reading comprehension scores necessary for the analysis. This section is focused on the data processing procedure for both speech and eye tracking.

3.3.1 Speech data processing

As discussed in the section 3.2.3 we used the EyeGotIt (El Baha et al., 2022) software for the experiment, which ensures both data recording and processing. In particular, it uses WebMAUS service (Kisler et al., 2017) to align text with speech. However, the resulting alignments often lack precision due to non-native speech and the presence of disfluencies, such as truncations, hesitations, and repetitions. The result is in line with the state of the art dedicated to the speech recognition issues with respect to non-native speech. Indeed, it has been underlined that the performance of forced aligners and automatic speech recognition (ASR) systems worsens when dealing with non-native and/or disfluent speech (Ballier et al., 2023; Graham and Roll, 2024; Williams et al., 2024). Despite these challenges, ASR systems (text-to-speech conversion) bring useful insights into L2 pronunciation (Popescu et al., 2024; Ballier et al., 2023).

The Figure 3.5 illustrates the challenges faced during speech processing. It shows three examples of word boundaries obtained with and without the use of ASR or manual annotation. The middle annotation level "ORT-MAU" is the initial annotation output from WebMAUS (pipeline without ASR) without manual correction of text or boundaries. The bottom layer "ORT-MAU-ASR" represents the output of WebMAUS (pipeline with ASR) without manual correction. And the top tier "ORT-MAU-corr" is the result of the use of WebMAUS (without ASR) along with manual corrections, including the addition of disfluencies and the adjustment of word boundaries.

In this example, the original phrase is:

"Ben grabbed his backpack and (...)".

The speaker's production differed from the original phrase: the participant mispronounced the word "backpack" and instead, pronounced /baspak/, and then he self-corrected by using a truncation "back" (in red) and a repetition "backpack" (in blue). The pronounced phrase was:

"Ben grabbed his backpack (/baspak/) his back backpack and (...)".

As shown on the Figure 3.5, the layer "ORT-MAU" illustrates that the system is unable to output correct word boundaries if the text does not contain all the spoken words and disfluencies. The layer "ORT-MAU-ASR" shows that the system captures mispronunciations highlighted in a violet circle, such as "bathpack" instead of "backpack", however it fails to capture disfluencies, such as the truncation "back". In our study, it is essential to accurately identify the number and location of the disfluencies and the word boundaries. It is important for us to obtain word

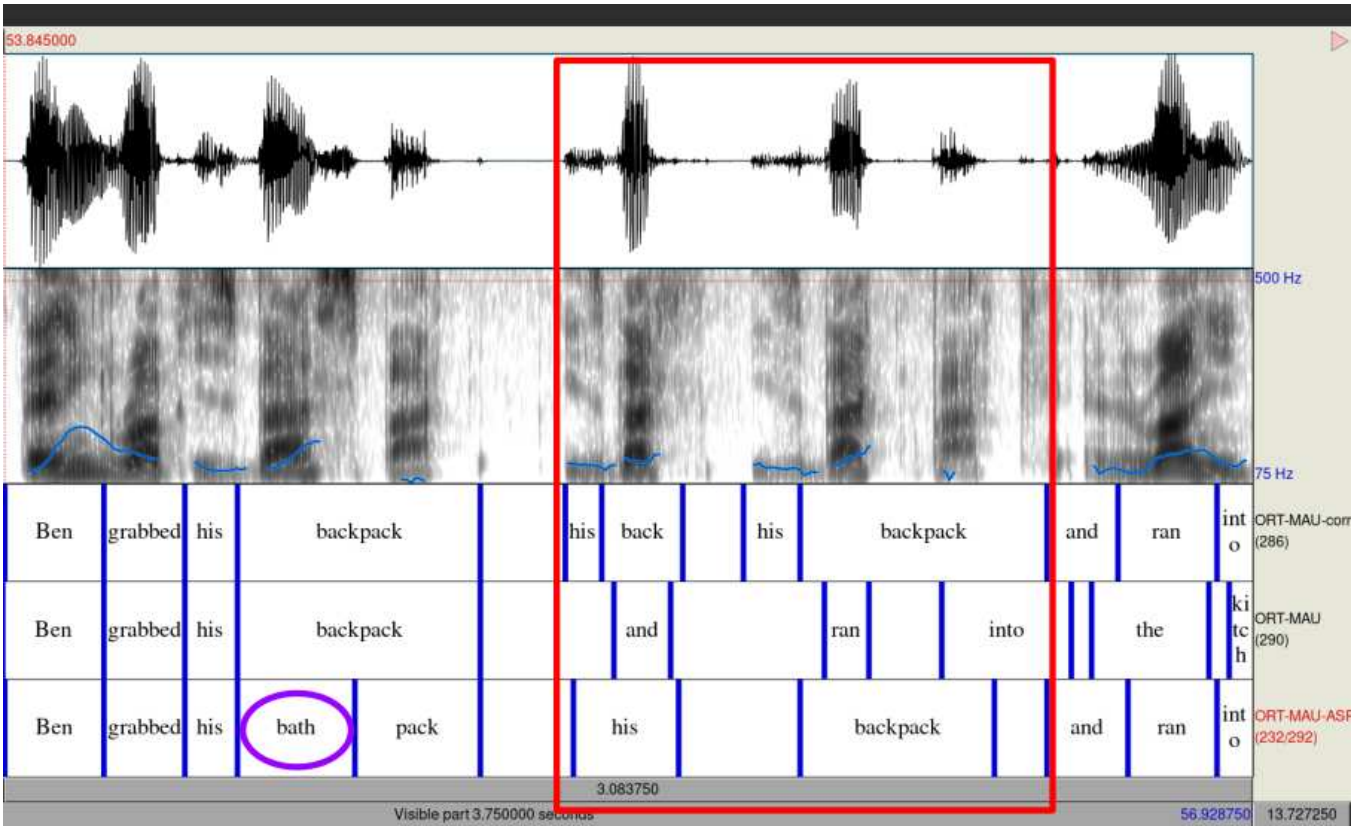

Figure 3.5: Alignment correction in Praat

boundaries based on the original text rather than words generated by ASR, as we need to align speech with eye movement. Since the combination of words “bath pack” generated by ASR does not appear in the original text, we would need to add a new annotation layer indicating that it corresponds to the original word “backpack”, which would be a time-consuming process.

Therefore, for our analysis, we prefer the output shown in the layer “ORT-MAU-corr”. This layer represents the disfluencies, the original words, and their boundaries. As for the mispronunciations, they can be handled by manually correcting the annotation layer containing the phones (this is also output by WebMAUS but not illustrated here). For example, we have manually corrected all the boundaries of vowels /u/ and /ʊ/ based on both the visual presence of F1 and F2, as well as the auditory criteria for further fine-grained analysis. It is worth mentioning that for the purposes of this thesis, only these two vowels have been analysed as the manual processing of /u/ and /ʊ/ out of all data (that is Protocols 1 and 3) requires about 3 months of manual work.

To sum up, for speech alignment, we first manually enrich texts in disfluencies, then we run WebMAUS tool (the pipeline without ASR) and after, we manually verify the resulted word boundaries. As for the phones boundaries we proceed to a manual correction of /u/ and /ʊ/ boundaries [Figure 3.6].

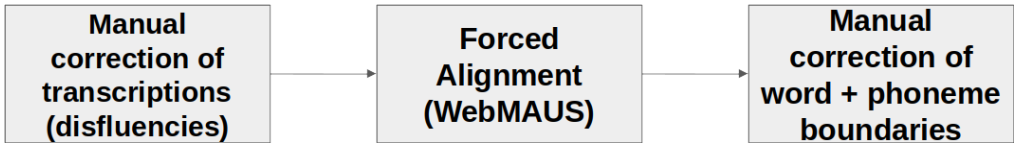

Figure 3.6: Speech processing algorithm

3.3.2 Eye tracking data processing

Data filtering

As shown on Figure 3.4 in the previous section and Figure 3.7 below, the EyeGotIt (El Baha et al., 2022) system provides illustrations of eye movements correlated with text. We used these images to assess the calibration quality. Below, we provide examples of well calibrated data (Figure 3.7) and of poorly calibrated data (Figure 3.8). To remind, the red dots represent the eye gaze and the circles represent the fixations, a region of agglomerated eye gaze. The diameter of the circles is proportional to the fixation length. The lines between the circles are the saccades, which are the movements between two fixations.

In the Figure 3.7, the fixations and the saccades follow the text lines, clearly indicating which words receive fixations and which ones are skipped. Although in some cases, the fixations appear above or below the text lines, it does not prevent from recognizing gazed words. In such cases, the gazed words can be manually checked and adjusted if necessary.

The Figure 3.8 shows poorly calibrated data. Here, the fixations and the saccades are right skewed and do not follow the text lines. The red dots appear randomly around the text and are not always combined to fixations. Therefore, the calculation of the fixations and the identification of gazed words are not sufficiently precise.

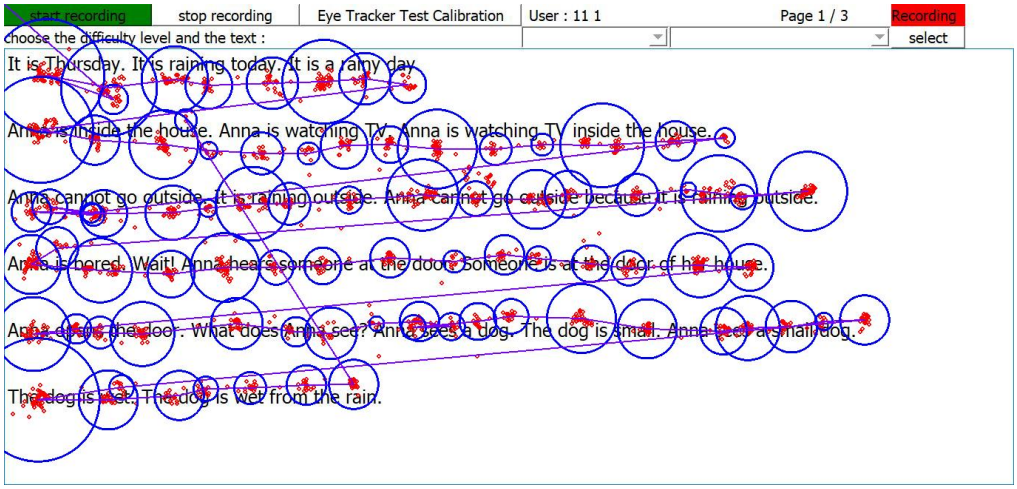

Figure 3.7: Example of well calibrated eye-tracking data

Based on our observations of images depicting eye movements, only well calibrated data was included in our analysis. As a result, the data of 1 native speaker, 3 French speakers and of 9 Japanese speakers were excluded from the analysis. The Table 3.5 presents the original number of speakers, the number of speakers remaining after the

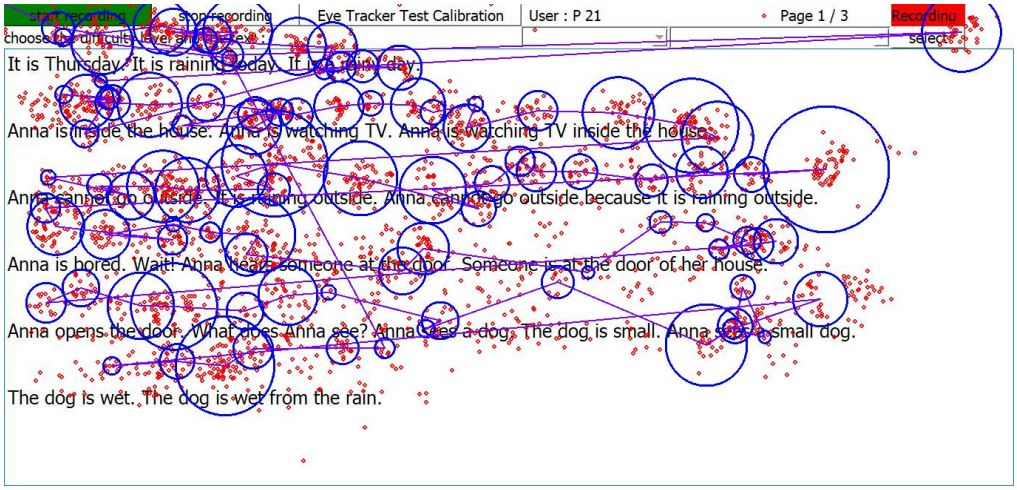

Figure 3.8: Example of poorly calibrated eye-tracking data

Native language	**Nbr speakers before filtering**	**Nbr speakers after filtering**	**Total nbr of files (3 texts * nbr of speakers after filtering)**
French	46	43	129
Japanese	24	15	45
Native speakers	12	11	33

Table 3.5: Number of speakers and files before and after filtering eye-tracking data

filtering, and the total number of remaining files per L1 group.

As shown in the Table 3.5, the quality of the data recorded during the experiment with Japanese students was poorer compared to that of French and native English speakers, even though the exact same experimental setup was used in all cases. We are aware of several factors that can influence the eye-tracking calibration quality, such as lightning conditions, excessive makeup, light or dark eye color, and body movements (Conklin et al., 2018). Additionally, a study (Blignaut and Wium, 2014) suggests a potential correlation between experimental conditions and ethnicity with eye tracking data quality, reporting lower accuracy for Asian participants. However, this finding requires further investigation through more experiments using different eye-tracking devices and experimental designs. In our study, the exact factors contributing to poorer calibration for the Japanese speakers remain unclear.

Despite the unbalanced dataset, we can still observe tendencies among the remaining speakers and formulate hypotheses about those whose data was excluded from the analysis. In future steps, acquiring more data and paying closer attention to calibration issues will be essential points to improve.

Defining gazed words

The EyeGotIt system provides word coordinates (in pixels), its length and width, as well as fixation coordinates (in pixels). Using this information, we can define a word region and correlate words with their corresponding fixations.

The Figure 3.9 represents a text example, where the red rectangle depicts the calculated word region based on the word's coordinates, length and width. An additional margin of error, or "epsilon," of 10 pixels was added to

this calculation. The blue dot indicates a fixation. If the coordinates of the fixation fall within the word region, we consider the word as being gazed. Moreover, we have the time information of both eye gaze and speech, which allows us to determine when the word was gazed at and pronounced.

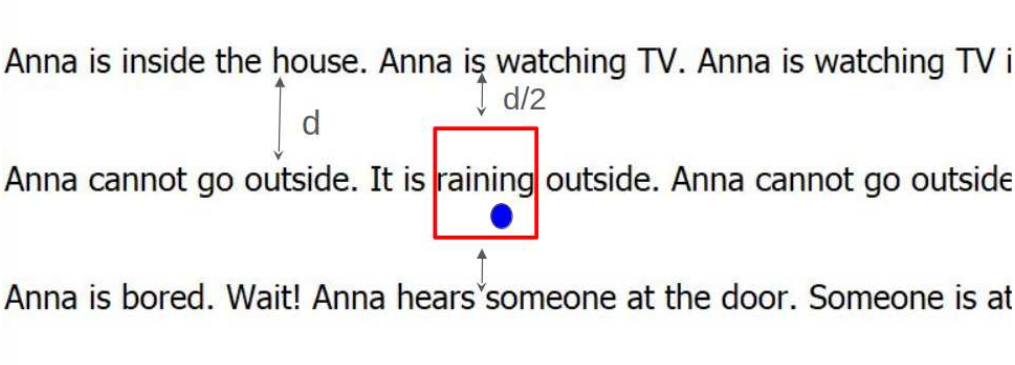

Figure 3.9: Definition of word region and calculation of gazed words

Summary

To sum up, data processing is a time-consuming task that requires manual intervention. Eye tracking data is particularly sensitive to a variety of external factors, which can lead to a reduction of the dataset size. Although the number of samples of Japanese speakers is significantly smaller than that of French speakers, we still obtained interesting results and identified tendencies that could be further verified with a larger and more balanced dataset.

Chapter 4

Speech and eye-tracking features extraction and modeling

This section is dedicated to the analysis of speech and eye-tracking features out of the data collected in the conditions described in the previous section. The selected features are commonly used for L2 automatic assessment and are highlighted by state of the art proposed in Chapter 2. They include both speech and eye-tracking characteristics, such as speech rate, pitch, intensity, speech/word/phone duration, phone formants, oral (hesitations, truncations, repetitions) and silent (silent pauses) disfluencies, fixations, their number and duration. We hypothesize that this analysis will give us a better understanding of the challenges that both French and Japanese speakers face when reading aloud text in English compared to native English speakers, as well as the strategies of coping with them from the perspective of both eye movement and speech. We investigate whether these characteristics align with the level reported by the participants when completing the questionnaire about their linguistic background (see Chapter 3 about data acquisition), whether they reveal cultural differences, and, finally, if they can be integrated in a machine learning model to identify L2 level (see Chapter 5).

The analysis will be presented at different levels in top-down perspective: globally, on the **text level** in order to get a clearer idea of overall speaker's fluency; at the **word level** to investigate the difference in processing frequent and infrequent words, as well as to get a better understanding of the unknown words decoding strategies implemented by the participants having different linguistic and cultural backgrounds; and finally, the analysis will be provided at the **phone level** to acquire a deeper insights into the pronunciation challenges faced by the speakers.

For each analysis level, the results are illustrated with boxplots and statistical modeling, such as Linear Mixed Effect Models (LME). LME models are a common practice among researchers especially in the linguistic community. They enable us to examine a correlation between the target variable with other parameters taking into consideration both fixed effects (representing variables of interest) and random factors, such as groups and individual variation.

The models are applied using the lmer package in R (Kuznetsova et al., 2017). In order to select the model that provides the best fit to the data, we relied on the Akaike's Information Criterion (AIC) to compare the models and to the residual plots. The results presented below correspond to the models chosen this way.

In this section, we focus on the textual data collected during 1 and 3 Protocols among French, Japanese and English native speakers. A separate Chapter 7 will be dedicated to manga reading.

4.1 Text level

4.1.1 Speech features

Speech rate

Speech rate is a parameter strongly correlated with L2 level: a rich literature reviewed in the Chapter 2 highlights among others that speech rate is an important factor often included in automatic fluency assessment and influencing L2 fluency perception by native speakers. In line with these findings, we hypothesize that our experimental configuration will highlight differences between representatives of different cultures and L2 proficiency that can be integrated in an automatic fluency assessment model, discussed in Chapter 5.

$$speechrate = nbrSyllables/phonationDuration$$

$$speechrate = nbrWords/phonationDuration$$

We used both methods, but as the results were quite similar, we report below only the results calculated using the first equation. The number of syllables taken into account corresponds to the number of syllables expected to be in the word, not the number of syllables that were actually pronounced. This decision is explained by the need of additional manual annotation for the phenomena of syllable deletion or insertion. The disfluencies and their corresponding number of syllables (manually calculated) are also included in the total number of syllables. Phonation duration is the duration of speech without pauses. It was calculated based on the word boundaries obtained after the manual alignment correction, as explained in the previous chapter 3.

Based on the reviewed literature in Chapter2, we formulate the following hypotheses:

1. **Hyp 1:** non-native speakers read aloud slower than native speakers
2. **Hyp 2:** non native speakers read faster as their L2 level improves

The boxplots (Figure 4.1) show the distribution of speech rate among speakers of different L1 and L2 levels reading texts of different complexity. To remind, the L2 levels were divided into "beginner" corresponding to speakers having A2 or B1 according to their self-evaluation, and "advanced" corresponding to B2, C1 and C2 level. The distribution shows that English native speakers read aloud faster than both French and Japanese speakers, and French

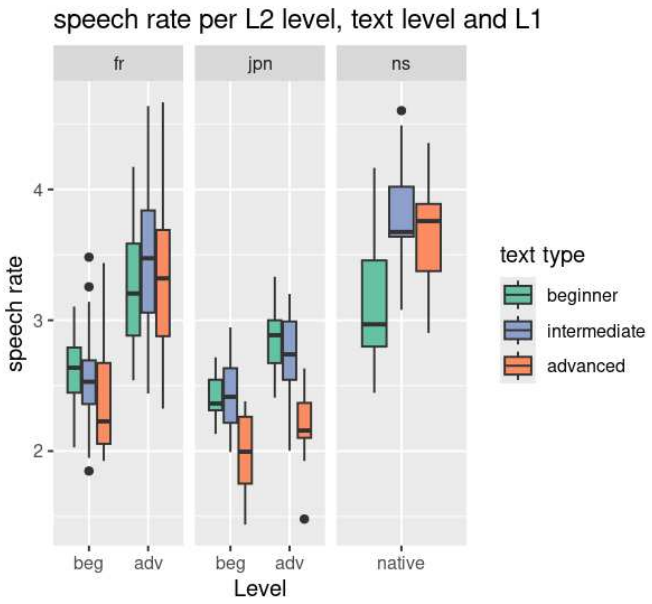

Figure 4.1: Speech rate per text level, L2 level and L1

speakers read faster than Japanese speakers. Both French and Japanese speakers improve their speech rate with L2 level progression, however the improvement is less salient for Japanese speakers. We observe lower speech rate with the advancement of text complexity among French and Japanese speakers. However, native English speakers read the advanced level text as fast as intermediate ones. Indeed, during discussions with native English speakers after the experiment, most of them mentioned an increase of texts' naturalness as the level increases. The advanced text was considered by them as the most natural sounding besides the presence of infrequent words. As for the beginner text, it sounded the most unnatural due to the presence of short sentences.

In order to examine whether the results observed in the boxplots are statistically significant, we fitted a Linear Mixed Effect Model (LME):

$$speechRate \sim textLevel + L2Level + nativeLang + (1 \mid speaker) \tag{4.1}$$

The model calculates the correlation between speech rate and fixed factors such as text level, L2 level, and native language, as well as accounts for individual differences. The results reported below are obtained by applying the model on the whole dataset including all the speakers (when comparing the results for all the L1) or on specific subsets of data depending on the L1 or L2 level. Applying the model on subsets of data helps to account for specific correlations, for example beginner and advanced speakers within the same L1 group.

The model shows similar tendencies to those illustrated with the boxplots. Native speakers read aloud faster than both French (estimate=-0.4, p=0.1) and Japanese speakers (estimate=-1, p<0.001). An increase in text complexity is associated with a slower speech rate: the speech rate decreases by 0.09 when reading an advanced text compared to beginner one (estimate=-0.09, p=0.03).

Advanced French speakers read faster than beginner ones (estimate=0.78, p<0.001). Beginner French speakers read advanced text slower than the beginner one (estimate=-0.2, p=0.02), but no significant difference was observed between the speech rate during beginner and intermediate texts reading. Advanced French speakers read the intermediate text faster than the beginner one (estimate=0.2, p<0.001), however the results for the advanced text are not significant.

Advanced Japanese speakers also read faster than beginner speakers (estimate=0.3, p=0.003). Beginner Japanese speakers read the advanced text slower compared to the beginner one (estimate=-0.4, p<0.001), but the results for the intermediate text are not significant. Advanced Japanese speakers also read the advanced texts slower than the beginner one (estimate=-0.68, p<0.001), and the results are not significant for the intermediate text.

In general, beginner Japanese speakers read slower than beginner French speakers (estimate=-0.29, p=0.04). The difference is more salient between advanced Japanese and French speakers (estimate=-0.7, p<0.001).

The results show that the native language and the L2 level used separately as fixed factors contribute to the estimation of the speaker's speech rate. Based on that, we investigated whether an interaction of native language and L2 level is significant. For this purpose, we updated the model and applied it on a subset of data for French and Japanese speakers only:

$$speechRate \sim textLevel + L2Level * nativeLang + (1 \mid speaker) \tag{4.2}$$

The updated model revealed significant contributions from the L2 level and its interaction with the native language. Specifically, the predicted speech rate for advanced French speakers was significantly higher than expected based on the main effect of L2 level and native language alone (estimate=0.45, p=0.03). The result suggests that the effect of proficiency level on the speech rate differs depending on the speaker's native language.

According to the reviewed literature in Chapter 2, Japanese speakers tend to insert epenthetic vowels to preserve consonant-vowel syllables structure that characterizes their phonological system. We hypothesize that their speech rate can be partly due to the vowel insertion phenomenon. This hypothesis needs further investigation, and in particular, manual annotation.

To sum up, as anticipated in our hypotheses the results confirm that native speakers are the fastest readers. The interesting add is the difference between French and Japanese speakers although both population provide comparable L2 level estimations. French speakers read faster than Japanese speakers, and their speech rate improves more noticeably as their L2 proficiency increases compared to that of Japanese speakers. The speech rate parameter will be included in the L2 level prediction model.

Pitch

The literature presented in Chapter 2 suggests that pitch variation is associated with L1 and L2 proficiency levels. Specifically, the literature highlights that non-native speakers produce less pitch variation than native speakers which

can be due to the lack of confidence in their speaking skills (Zimmerer et al., 2014). Additionally, Japanese speakers naturally have a more monotonous intonation (Avery and Ehrlich, 1992). In this section, we will investigate whether the results obtained on our dataset correspond to the state of the art and whether pitch variation can be a significant factor enabling us to predict L2 level and identify cultural differences.

Based on the literature, we hypothesize that:

Hyp 1: Japanese speakers have less pitch variation, as according to the literature, Japanese language is characterized by monotonous intonation

Hyp 2: more advanced speakers would have more pitch variation, as they would focus less on segmental pronunciation, and could use more intonation to express the meaning of the texts. To remind, the texts (see Appendix A) are short stories containing both direct and indirect speech and different phrase types such as exclamatory and interrogative, that are selected with a goal to elicit speakers to use a variety of intonations.

To extract pitch values, we relied on a Praat (Boersma and Weenink, 1992–2022) script[1] that extracts pitch in Hz every 10ms. Male and female speakers were processed separately, as they require different settings for pitch extraction: for male speakers we chose 75Hz as pitch floor and 300Hz for pitch ceiling, and for female speakers - 100Hz and 500Hz respectively. A filter was applied to the resulting values to exclude those equal to 0, as it corresponds to pauses or to unvoiced segments.

Then, the values were converted to a nonlinear scale, semitones, which enables data normalization. Indeed, pitch calculated in Hz is gender dependent: female speakers have higher pitch than male. In order to smooth these differences, we applied the following formula for Hz to semitone conversion:

$$\text{semitone} = 12 \cdot \log_2 \left(\frac{F_0}{F_{0\text{base}}} \right) \tag{4.3}$$

There are multiple ways to choose the F0base value, it can be equal to 1, 50, 100 Hz or mean/median per speaker. In our case, F0base is speaker dependent and represents the median F0 per speaker. The median value was chosen over mean, as it is less affected by outliers.

Then, we calculated the average, standard deviation, range and other statistics per speaker and per text. We report here the results of standard deviation, as they appear the most significant.

The Figure 4.2 illustrates that native speakers use more pitch variation than French and Japanese speakers, and French speakers more pitch variation than Japanese speakers. To verify if the observations are statistically significant, we applied a LME in R:

$$stdSemitonePitch \sim textLevel + L2Level + gender + nativeLanguage + (1 \mid speaker) \tag{4.4}$$

Here, the std semitone pitch is a dependent variable, text level, gender and native language as fixed factors and

[1] script by by Setsuko Shirai http://phonetics.linguistics.ucla.edu/facilities/acoustic/CreateTable.txt

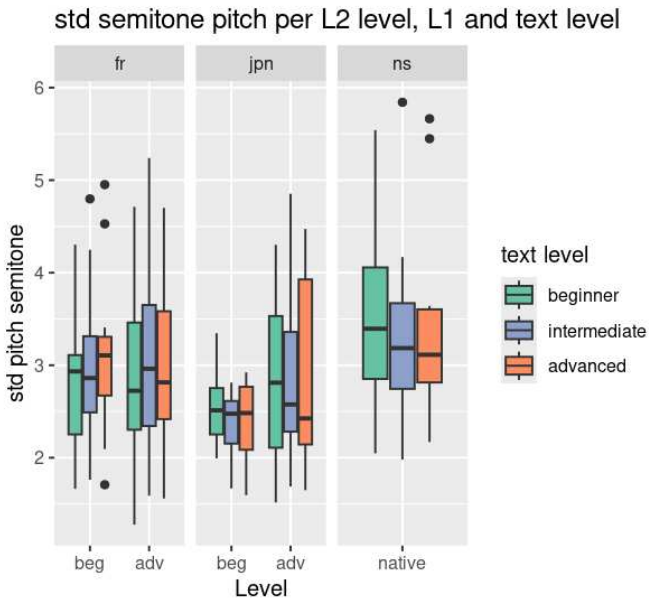

Figure 4.2: Pitch variation (semitone) per text level, L2 level, and L1

speaker is as a random factor.

The results show that gender and native language are significant factors influencing pitch variation. Being a man is associated with a decrease in std pitch by 0.54 ($p<0.01$). Native speakers have higher pitch standard deviation than French speakers (estimate=0.5, $p=0.03$) and higher than Japanese speakers (estimate= 0.78, $p<0.01$). The L2 level and text level are not significant.

The results correspond to those reported in the literature and are in line with our hypotheses, stating that native speakers have higher pitch variation, and that Japanese speakers have a more monotonous intonation (Avery and Ehrlich, 1992). However, the model shows no significant difference between L2 levels among French nor Japanese speakers. Nevertheless, the pitch parameter will be incorporated into our L2 level identification model in order to investigate whether it enhances the model prediction when combined with other features.

Intensity

The use of intensity varies across languages. As discussed in Chapter 2, in English (stress-timed language) intensity along with duration are used to highlight stressed syllables; in French (syllable-timed language) intensity is also used as a distinction between stressed and unstressed syllables; however, in Japanese (mora-timed language), the speakers rely on pitch to express the stress rather than intensity.

Moreover, intensity is one of the metrics allowing the listener's perception of confidence and doubt (Jiang and Pell, 2018).

Based on this information, we hypothesize that:

Hyp 1: Japanese speakers' intensity variation is lower

Hyp 2: More proficient speakers read louder, as they are more confident in their pronunciation.

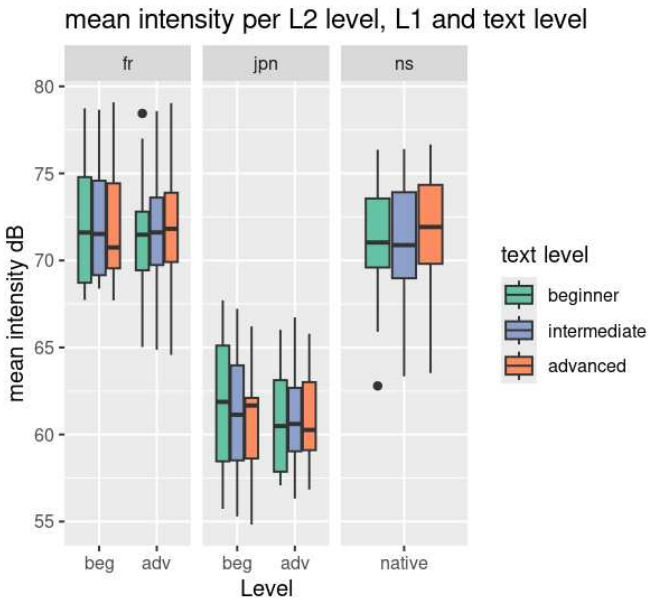

Figure 4.3: Mean intensity per text level, L2 level and L1

In order to analyze intensity metrics, the same Praat script was used as for the pitch extraction in the previous section. It extracts intensity (in dB) features every 10 ms of the sound file. Then, the average and standard deviation per text and per speaker were calculated.

The Figure 4.3representing the boxplots of the mean intensity does not show patterns suggesting different intensity depending on L2 proficiency. It illustrates, however, that Japanese speakers (both beginner and advanced) speak lower than French and English native speakers.

The application of a LME (formula below) confirms this observation: the mean intensity of Japanese speakers is about 11 dB lower than that of French speakers (p<0.01) and 10dB lower than that of native speakers (p<0.01). No significant difference was found between French and native speakers and among different L2 level and text levels.

$$meanIntensity \sim textLevel + L2Level + nativeLang + gender + (1 \mid speaker) \tag{4.5}$$

The intensity measures are influenced by multiple factors, such as equipment, microphone distance from the mouth, external noise and other conditions. Although we made sure that the recording equipment and the conditions were similar in all the experiments, some factors could have influenced the recording quality. For example, the recording was realized in a post-covid period, and some Japanese participants preferred keeping their mask on during the experiment. In order to lower the effect on the absolute intensity values, we analyzed its variation among speakers groups.

The Figure 4.4 shows intensity variation among three speaker groups: French speakers have the highest standard deviation of intensity, followed by native speakers and then by Japanese speakers. Intensity variation also changes within texts of varying complexity, which is more salient for Japanese speakers who use less intensity variation when

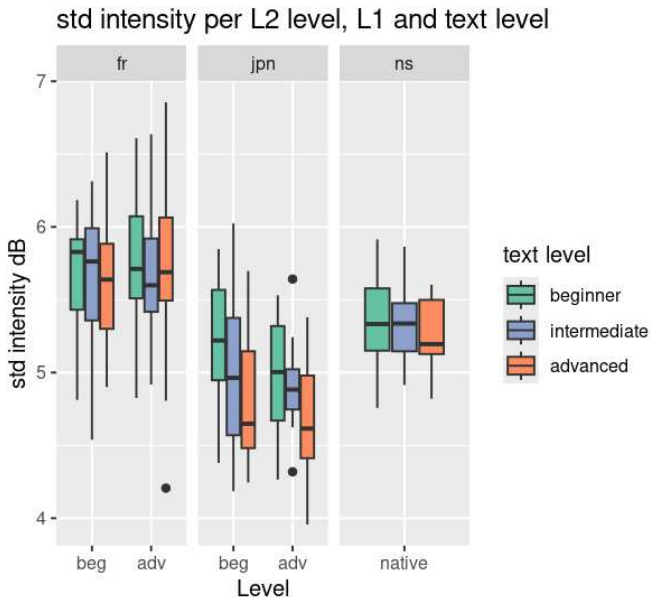

Figure 4.4: Intensity variation per text level, L2 level, and L1

reading more advanced texts.

The LME model (formula below) confirms a higher intensity variation (std) of French speakers compared to Japanese (Estimate=0.89, p<0.01) and native speakers (Estimate=0.42, p<0.01).

$$stdIntensity \sim textLevel + nativeLang + L2Level + gender + (1 \mid speaker) \tag{4.6}$$

The LME model showed that intensity standard deviation of Japanese speakers is dependent on the text complexity. More complex texts are associated with lower intensity variation: intermediate texts compared to beginner ones show an estimate of -0.15 (p=0.23); advanced texts compared to beginner ones show an estimate of -0.33 (p<0.01). The text level is not significant neither for French nor for native speakers. Moreover, the L2 level does not appear as a significant factor.

To sum up, in our dataset, speech intensity is correlated with L1: Japanese speakers read with lower intensity, and French speakers have the highest intensity variation followed by native English speakers. Intensity is also correlated with text complexity for Japanese speakers whose intensity lowers with an increase in text complexity, which can be associated with an increase in cognitive load and/or an expression of a lack of confidence. No significant effect was found for L2 proficiency level.

In general, the results align with those observed in the literature and meet our hypotheses. Later, the intensity measure will be integrated into our L2 level prediction ML model in order to examine whether it contributes to the model's result when added to other speech and eye tracking characteristics.

Disfluencies

This section is dedicated to oral disfluency analysis including truncations, hesitations (or filled pauses) and rep-

native lang	level	total nbr speakers	nbr speakers with repetitions	nbr speakers with truncations	nbr speakers with hesitations
fr	A2	3	3	3	2
	B1	10	10	10	7
	B2	13	13	12	6
	C1	18	18	17	10
	C2	2	2	2	2
jpn	B1	12	12	11	7
	B2	11	11	11	9
	C1	1	1	1	1

Table 4.1: Number of speakers using disfluencies

etitions. The silent disfluencies or silent pauses will be discussed in the next section.

As seen in Chapter 2, disfluencies reflect speech planning strategies (Tavakoli, 2011) and contribute to the listener's judgment about L2 fluency (Rose, 2017). Some of them, such as hesitations, reflect cultural differences, as they are L1 related. For example, in English the hesitations are "hum", "um", in French - "euh", and in Japanese "ano", "eto". Additionally, the disfluencies are also dependent on individual variation as suggested by (Kang et al., 2010; Vasilescu and Adda-Decker, 2006).

Based on the reviewed literature in Chapter 2, we propose the following hypothesis:

Hyp 1: The use of disfluencies is correlated with L2 level and text complexity: the number of disfluencies increases with higher text complexity and lower L2 proficiency

As discussed in Chapter 3 about data processing, the texts read by the participants were manually enriched in disfluencies, then the corrected texts were aligned with speech and the resulting boundaries underwent further manual correction to ensure accuracy. The disfluencies discussed in this section are based on those manually corrected texts.

The Table 4.1 presents the number of speakers who used different types of disfluencies at least once. Repetitions are the most frequently used disfluencies, followed by truncations, while hesitations are the least common ones.

The variation in the percentage of disfluencies is visualized in Figure 4.5. It is calculated based on the total number of pronounced segments including original words, truncations, hesitations, and repetitions compared to the original number of words in the text. The silent pauses are not taken into account. The boxplot 4.5 show that the number of disfluencies increases with text complexity for both native and non-native English speakers. The disfluencies are used less frequently by the native speakers than by the non-native speakers. Among non-native speakers, more advanced French speakers use less disfluencies than less advanced ones, and their disfluency rates approach those of native speakers. This tendency, however, is not observed among Japanese speakers. In order to verify if these observations are statistically significant, we applied the following Linear Mixed Effect (LME) model:

$$sqrt(disfluencyPercent) \sim textLevel + L2Level + nativeLang + (1 \mid speaker) \tag{4.7}$$

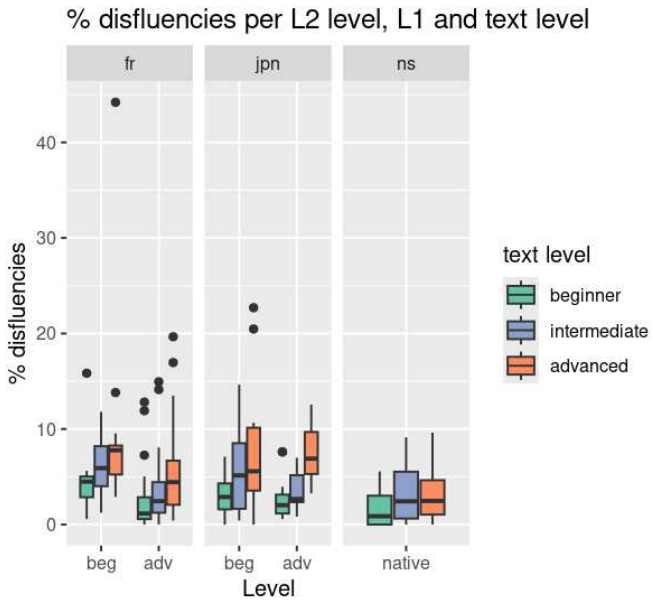

Figure 4.5: Percent of disfluencies per text level, L2 level, and L1

A square root transformation was applied to the dependent variable, disfluency percentage, in order to improve residual distribution and to meet the model assumption. The outliers were cleaned using the 2 standard deviations rule.

We included text level, L2 level, and native language as the model's fixed factors to assess their correlation with the frequency of disfluencies used by the speakers. In addition, we accounted for individual variation by including a random effect for the speaker.

The model confirms our observation that native speakers use less oral disfluencies compared to French (estimate=0.5, p=0.04) and Japanese speakers (estimate=0.6, p=0.01).

Advanced French speakers produced more disfluencies compared to less advanced ones (estimate=0.7, p=0.008). A higher text complexity is associated with more frequent disfluencies: intermediate text compared to beginner one (estimate=0.5, p<0.001), advanced text compared to beginner one (estimate=0.8, p<0.001).

As for Japanese speakers, the difference between advanced and beginner speakers is not significant (p=0.9). Similar to French speakers, more complex texts (advanced vs beginner) are associated with a larger number of disfluencies (estimate=0.9, p<0.001). However, no significant difference was found between beginner and intermediate texts. Adding the interaction term to the model representing the native language and the level of L2 did not contribute to the results.

Moreover, an important part of the model's variability is explained by the random effect of the speaker. It accounts for 42.6% of the variance for Japanese speakers and 60% for French speakers.

In conclusion, there is a significant correlation between the proportion of disfluencies and the speakers' native language, L2 proficiency level, and text complexity. However, a substantial part of model variation is explained by

individual differences among speakers.

In one of the next sections, we will provide a deeper analysis, by investigating the correlation of use of disfluencies with word frequency. And later, features related to disfluencies will be incorporated into the prediction model. Although disfluency related features will serve us to predict the L2 level, their implementation in real-life applications is challenging due to the need of manual intervention for their processing.

Pauses

According to the reviewed literature in Chapter 2, pausing patterns are correlated with L2 proficiency, especially the place of their occurrence (Davies, 2003). In this section, we apply a more general approach by analyzing their number, duration and duration variation across speakers of different native languages and L2 levels. The place of their occurrence (between or inside clauses) is not taken into account in this analysis. However, in an upcoming section, we will investigate how the use of pauses changes before lexical frequent and infrequent words.

The boundaries of the pauses were extracted based on the manually corrected alignment. Then, the pauses were filtered by their duration: only the pauses longer than 200ms were taken into account.

We hypothesize that:

Hyp. 1: advanced speakers use shorter pauses and they are less frequent than those used by beginner speakers

Hyp. 2: advanced speakers realize pauses of similar duration in the same text, on the contrary to beginner speakers who exhibit more variation in pauses duration

The Figure 4.6 illustrates the number of pauses used by French, Japanese and native English speakers. It shows that native speakers use less pauses than non-native speakers, and Japanese speakers use more pauses than French speakers. The number of pauses increases with text complexity and decreases with higher L2 level. The latter is more salient for French speakers.

The Figure 4.7 illustrates the pause percentage. The beginner text elicits a higher pausing rate compared to more advanced texts for all the speakers. It can be due to the number of short phrases used in the beginner text. Additionally, the percentage of pauses decreases for both French and Japanese speakers with L2 proficiency advancement (for intermediate and advanced texts)

The Figure 4.8 represents the difference in mean pauses duration used by French, Japanese and native English speakers. Beginner Japanese speakers use longer pauses compared to other speakers, but the pauses become shorter as their L2 proficiency improves. This trend is not noticeable for French speakers.

The Figure 4.9 illustrates the pauses duration variation. It shows a trend to higher duration variation with an increase in text complexity. The duration of pauses has less variation for more advanced Japanese speakers compared to beginner ones.

To verify the observations, we applied Linear Mixed Models and to investigate the relationship between the

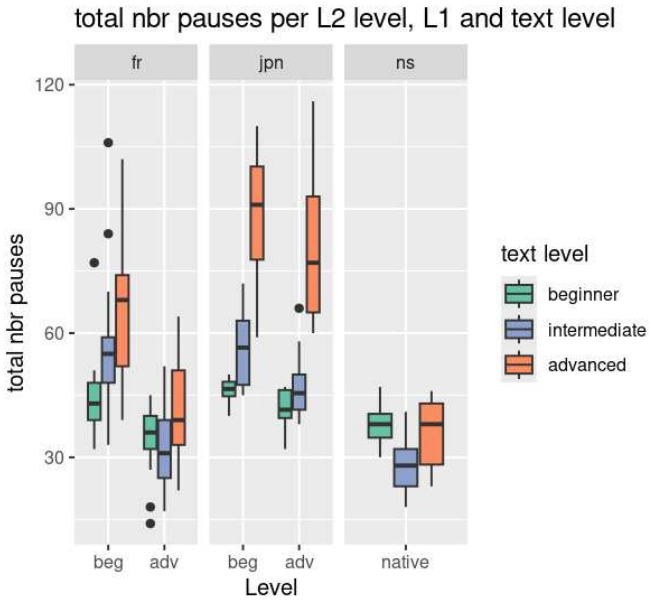

Figure 4.6: Number of pauses by text level, L2 level, and L1

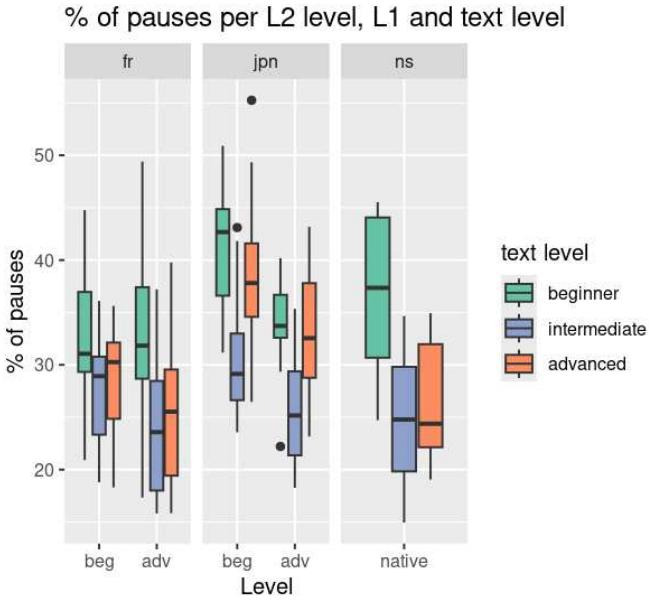

Figure 4.7: Percentage of pauses per text level, L2 level, and L1

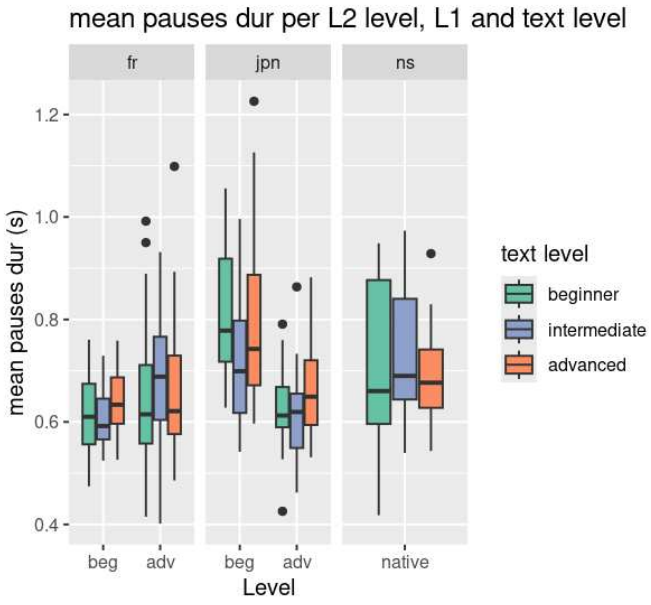

Figure 4.8: Mean pauses duration by text level, L2 level, and L1

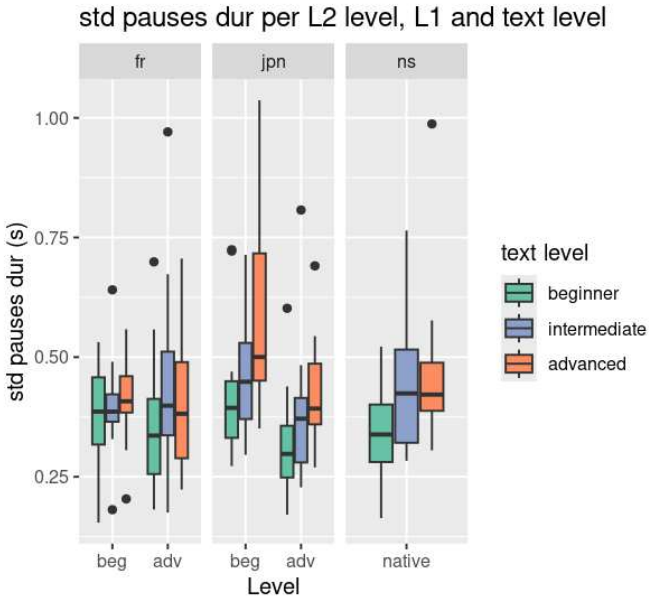

Figure 4.9: Pauses duration variation by text level, L2 level, and L1

duration and frequency of pauses with the L2 level and the native language. The first applied model is the following:

$$percentPauses \sim textLevel + L2Level + nativeLang + (1 \mid speaker) \tag{4.8}$$

The model reveals that the percentage of pauses produced by French speakers is lower than that of Japanese speakers (estimate=-4.87, p=0.002). For the French speakers, no significant difference was found between different L2 levels, however the text level is significant: intermediate vs beginner (estimate=-8, p<0.001), advanced vs beginner (estimate=6.6, p<0.001). As for the Japanese speakers, on the contrary to the French speakers, we observe a decrease in pauses percentage with L2 level improvement: advanced speakers vs beginner speakers (estimate=-6.3, p=0.01). Additionally, intermediate text elicits a lower pause percentage compared to beginner text (estimate=-9.5, p<0.001).

The next applied model is:

$$meanPausesDur \sim textLevel + L2Level * nativeLang + (1 \mid speaker) \tag{4.9}$$

The results suggest that in general, more advanced speakers use shorter pauses compared to less advanced speakers (estimate=-0.15, p=0.001). However, this is not the case for advanced French speakers, whose average pause duration is longer compared to less advanced speakers (estimate=0.19, p=0.001). This result highlights the observation of the Figure 4.8 where the tendency of shorter pauses for more advanced speakers was noticeable for Japanese, but not for French speakers. The model shows that more advanced texts elicit longer pauses for both speaker categories; however, the results are not significant.

Finally, we investigated the correlation of total pausing time with the same factors:

$$totalPausesDur \sim textLevel + L2Level + nativeLang + (1 \mid speaker) \tag{4.10}$$

The total pausing time of Japanese speakers is about 16s longer than that of French speakers (p<0.001) and 20s longer than that of native speakers (p<0.001). Concerning French speakers, total pausing duration is longer for the advanced text compared to the beginner one (estimate=7, p<0.001), and it is 10s shorter for the advanced speakers compared to the beginners (p<0.001). As for Japanese speakers, the total pausing time drops by 14s with L2 level advancement (p=0.01), and increases by 32s with the text complexity (advanced text compared to beginner, p<0.001). Although the fixed terms representing the L2 level and the native language were significant, their interaction is not significant.

4.1.2 Eye Tracking

The reviewed literature showed that eye tracking features, such as fixations and saccades are important factors reflecting L2 proficiency (Augereau et al., 2016a; Berzak et al., 2018), cognitive load (Bax, 2013) related to text com-

plexity and speech planning (Huettig et al., 2011).

In this section, we examine the eye movement behavior in a similar way as we did with speech features in the previous part. Eye movement characteristics, such as fixations, are extracted at the global text level and investigated in terms of their correlation with such factors as native language, L2 proficiency level, text complexity and reading comprehension score. The latter is based on the participants' answers to the multiple choice questionnaire after the text reading.

Based on the literature reviewed in Chapter 2, we hypothesize that:

Hyp 1: Eye movement behavior is dependent on the speaker's L1, in particular native speakers have shorter and less frequent fixations compared to non-native speakers, as they have more language use experience and are eager to anticipate the next word.

Hyp 2: Eye movement characteristics of more advanced non-native speakers approach those of the native speakers

Hyp 3: More advanced texts elicit longer and more frequent fixations as the presence of infrequent words and longer phrases can represent challenges for non-native speakers.

Eye fixations time of occurrence, duration and coordinates were extracted using EyeGotIt software (El Baha et al., 2022). As discussed in Chapter 3 dedicated to data processing, several recording samples have been discarded from the analysis due to calibration problems, that is the number of samples used in this section differs from that used in the previous section for speech analysis.

The Figures 4.10 and 4.11 provide a visualization of the variation of average fixation duration and their number per text level, L2 proficiency level and native language. The mean fixation duration and their number are lower for native speakers compared to non-native ones, and they are lower for French speakers than for Japanese. The fixations produced by advanced French speakers are shorter 4.10 and less frequent 4.11 than those of beginner ones, which could mean that the texts are less challenging for the advanced French speakers. However, Japanese speakers do not show the same tendency for L2 level. Non-native speakers use longer fixations when reading more advanced texts, which is not the case for native speakers 4.10. Indeed, almost all native speakers noticed that advanced text was more natural to read compared to the beginner one. However, they are challenging for non-native speakers. The number of fixations increases with the complexity of the texts for all the speakers 4.11 which could be due to higher number of words in the advanced texts as well as to the presence of rare words.

In order to address the hypotheses, we applied a Linear Mixed Effect model:

$$meanFixationDur \sim textLevel + L2Level * nativeLang + (1 \mid speaker) \tag{4.11}$$

The model was run on the dataset including all the L1s, as it allows to capture differences in mean fixation duration produced by speakers of different L1. It reveals that both French and Japanese speakers show longer average

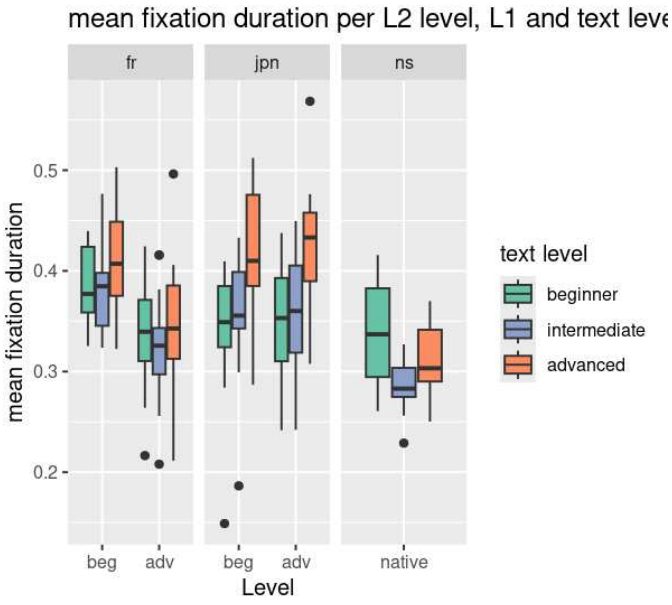

Figure 4.10: Mean fixation duration per text level, L2 level, and L1

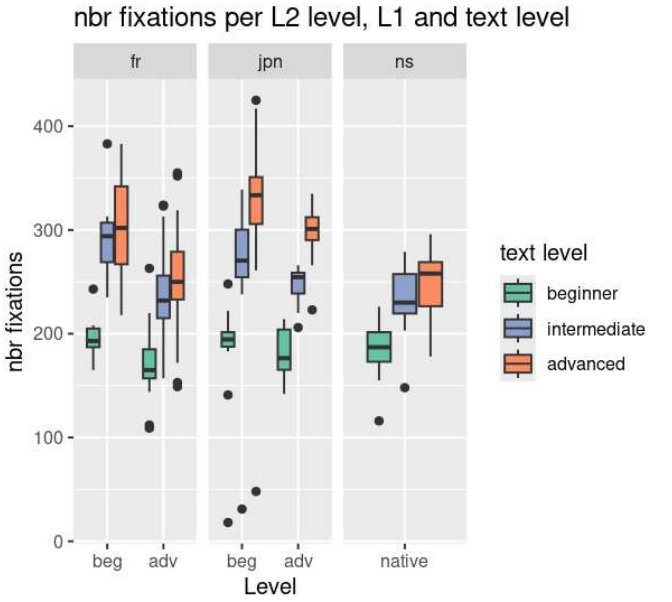

Figure 4.11: Number of fixations per text level, L2 level, and L1

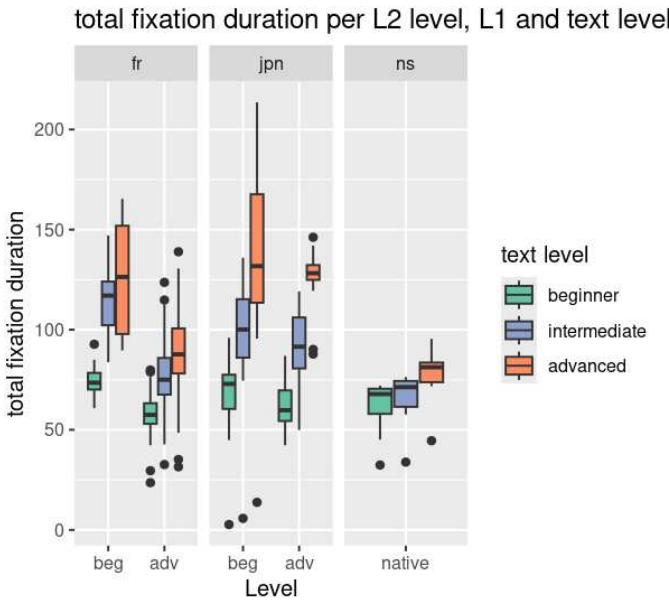

Figure 4.12: Total fixation duration per text level, L2 level, and L1

fixation duration than native speakers: native speakers vs French speakers (estimate=0.8, p=0.002), native speakers vs Japanese speakers (estimate=0.09, p<0.001). The advanced text contributes to longer fixations compared to the beginner one (estimate = 0.02, p<0.001). The interaction term suggests that advanced French speakers use shorter fixations compared to other speaker categories, but this result has a marginal significance (estimate=-0.05, p=0.05).

The following model investigates the correlation between mean fixation duration, L2 level, text complexity and the number of correct answers provided by the participants when answering the text comprehension quiz. The model was run on subsets of data depending on the native language in order to examine the variability within the speakers of the same language.

$$meanFixationDur \sim textLevel + L2Level + nbrCorrectAnswers + (1 \mid speaker) \tag{4.12}$$

For French speakers, an increase in the number of correct answers by 1 is associated with a decrease in mean fixation duration (estimate=-0.005, p=0.002). Being a more advanced speaker is associated with shorter fixation duration by approximately 0.05s compared to less advanced speakers (p=0.001). Mean fixations duration increases with the increase of text complexity: intermediate text vs beginner text (estimate=0.01, p=0.1, the result is not significant), advanced vs beginner text (estimate=0.02, p<0.001, highly significant).

On the contrary to the French speakers, the results of the model run on the Japanese speakers' data does not show a significant correlation between the comprehension score and mean fixation duration. As well as no correlation was found with L2 level. However, the mean fixation duration is related to text difficulty: advanced text requires longer fixations compared to the beginner text (estimate=0.08, p<0.001)

To sum up, the model's results suggest that native speakers use shorter fixations compared to non native speak-

ers. In addition, French speakers produce shorter fixations with L2 level improvement (which confirms the literature). However, this is not the case for Japanese speakers who do not show similar patterns. This can be due to an insufficient size of the Japanese dataset. So, this analysis requires further investigation and the model should be tested on a larger dataset. As for the text complexity, it is a significant factor influencing the mean duration fixation of both French and Japanese speakers. This confirms the results observed in the literature reporting longer fixations in more complex tasks requiring more cognitive effort.

4.1.3 Summary

To sum up, in this section we provided an overview of speech and eye-tracking features at a global level which with respect to our experimental design corresponds to the text level. In particular, we investigated relevant speech features such as as pitch, intensity, pauses and disfluencies, as well as eye tracking features, such as fixation duration. We correlated such features with factors such as with native language, L2 level and text complexity. The analyzed features provide valuable insights into the differences in oral text processing by native and non native speakers, as well as the speaker's L2 proficiency level. It also underling differences among French and Japanese speakers pointing out the relevance of the cultural factor.

In the next section, these features will be investigated at word level: we will examine speech and eye movement variation depending on the word type, for example frequent and infrequent lexical words and function words. Then, in Chapter 5, we will propose a set of features and motivate their relevance for automatic L2 level.

4.2 Word Level

In this section, the analysis is conducted at the word level. It enables us to explore how speakers of different native languages and L2 levels process words both from an oral and visual perspective. It will allow us to get a better understanding of the influence of the speakers' linguistic backgrounds on the processing of frequent or infrequent words and on the difficult word anticipation.

For this purpose, we extract and analyze speech and eye tracking features inside the word boundaries (obtained after a manual correction procedure, as explained in Chapter 3), as well as before the word boundaries in order to account for the differences in word anticipation strategies.

4.2.1 Frequent and infrequent lexical words vs function words

4.2.1.1 Speech features

First, we analyze pronunciation of different word categories in terms of pitch variation, intensity variation and duration. For feature extraction, we used a similar approach as in the section 4.1 where overall pitch and intensity

measures were explored at the text level. The word boundaries and their duration were identified based on the manual correction.

For the word category, 3 labels were used: function (or grammar) words, frequent and infrequent lexical words. The label of frequent or infrequent word is based on the frequency identified using CELEX corpora (Baayen et al., 1995). The decision threshold was set to log frequency lower than 1 which means that the word was found less than 10 times per million words in the corpora.

We hypothesize that:

Hyp1: infrequent words require more processing time: increased pronunciation and fixation duration. The processing time decreases with higher L2 proficiency level. Japanese speakers process infrequent and long words due to a higher cognitive load related to the linguistic distance between English and Japanese.

Hyp 2: longer words are produced with higher pitch and intensity variation

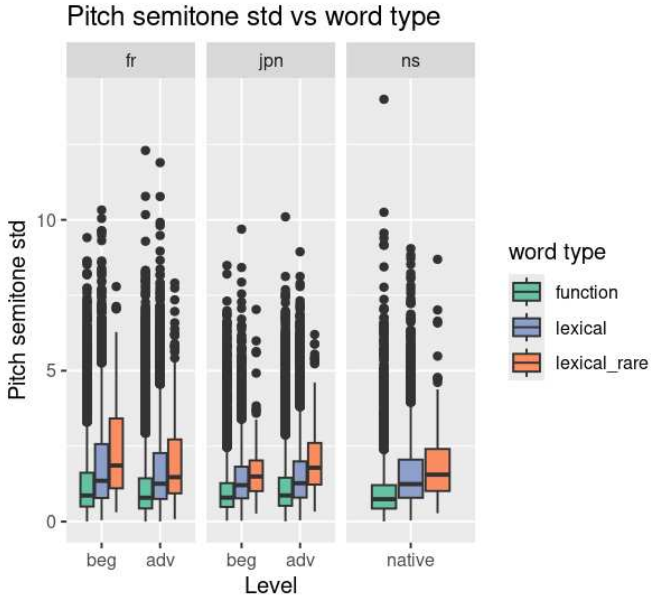

Figure 4.13: Pitch variation (semitone) per word type, L2 level, and L1

The data visualization shows a correlation between word categories and pronunciation features, however a large number of outliers is present. The outliers will be handled later in LME depending on other factors including the speaker and the word. According to the visualizations, infrequent words are produced with higher pitch variation 4.13 and longer duration 4.15 than other lexical or function words. French speakers have higher intensity variation for all the words compared to other speakers and the intensity of the rare words is higher than that of more frequent words 4.14. This tendency is less noticeable among other speakers. French speakers spend less time on the word pronunciation with L2 level improvement, however, this trend is not noticeable among Japanese speakers. According to the plots, the duration is the most salient factor highlighting the differences between word categories among all the speakers. The duration is also related to the speech rate, which is higher among advanced French speakers, as

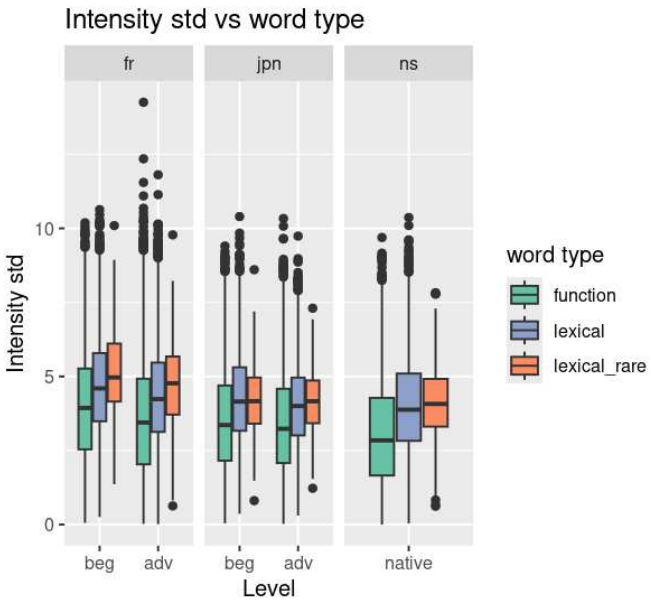

Figure 4.14: Intensity variation per word type, L2 level, and L1

observed in the section 4.1. Longer duration, higher pitch and intensity variation of lexical words is correlated with their number of characters or syllables. In fact, infrequent words are generally longer than other frequent lexical words 4.16. So, the increase in their duration, pitch or intensity variation could be related to their character length or the number of syllables, but also it could demonstrate challenges to access its phonological or semantic form due to their lower frequency.

We built linear mixed effect (LME) models in order to examine the relationship between the audio features (pitch standard deviation in semitones, intensity standard deviation and duration) and their influential factors that were used as fixed effects in our model, such as word type (lexical frequent, lexical infrequent and function), number of characters, number of syllables, text level, L2 level. We also used random factors, such as speaker and word to account for individual and word variation.

Pitch variation

The first model uses pitch variation as a dependent variable. The pitch is transformed in semitone as in the previous section to lower the gender effect. In addition, we applied a square root transformation in order to stabilize the variance and to obtain better residual plots.

As a result, the following model was applied:

$$sqrt(stdPitchSemitone) \sim wordType + nbrChar + nbrSyll + textLevel + L2Level + (1 \mid speaker) + (1 \mid word) \quad (4.13)$$

The models were run separately on each subset of data depending on the speaker's native language to account

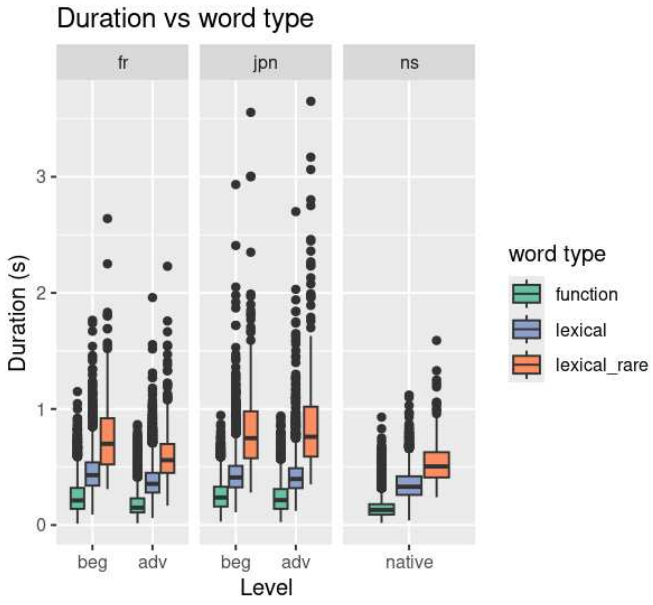

Figure 4.15: Word duration by type, L2 level, and L1

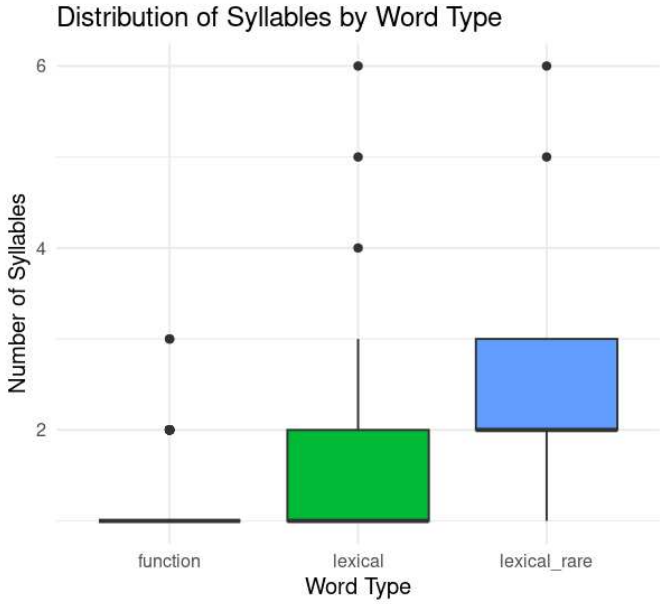

Figure 4.16: Number of syllables per word type

for the variation between speakers of different L2 levels within the same L1 group. After the model was fit, the outliers were filtered using the 2 standard deviation rule, and then, the model was rerun on the updated data.

Only the statistically significant results (p<0.01) are reported below:

Being a lexical word is associated with an increase in the sqrt(stdPitchSemitone) compared to grammar words. For French speakers, the estimated increase is 0.1, for Japanese speakers, it is 0.09, and for native speakers, 0.14. Infrequent lexical words entail a higher increase in pitch variation compared to grammar words: estimate equals to 0.15 for French speakers, 0.16 for Japanese speakers, 0.18 for native speakers.

Each additional character in the word is associated with an increase of about 0.017 units in sqrt(stdPitchSemitone) for native speakers, 0.027 for French speakers and is not significant for Japanese speakers.

The number of syllables was not significant for the speakers of all the three languages.

Beginner-level text is associated with an increase in pitch variability by about 0.058 units compared to the advanced text for the native speakers. However, for the French speakers, the beginner text entailed a decrease in pitch variability by about 0.03. This corresponds to the results found in section 4.1.

L2 level did not show significant results.

We built similar models on subsets of data depending on the word type and included the native language as a fixed factor. It allowed us to compare pitch variations of different word types between native languages. However, no significant difference was found (p>0.05).

In summary, lexical and longer words elicit higher pitch variation among all the speakers, which confirms the Hypothesis 2. L2 is not a significant factor of pitch variation.

Intensity variation

In order to check intensity variation depending on the word type, we applied the following model:

$$sqrt(stdIntensity) \sim +wordType + nbrChar + nbrSyll + textLevel + L2Level + (1 \mid speaker) + (1 \mid word) \quad (4.14)$$

Similar to the previous section, the model was run on subsets of data depending on the native language in order to observe the difference in intensity variation within the speakers of the same language. As in the previous model, we applied square root transformation to the intensity standard deviation to stabilize the variance.

We report here only the significant results (p<0.01):

Lexical words compared to grammar entail an increase in intensity variation by 0.09 in sqrt(stdIntensity) for Japanese speakers, 0.1 for French speakers and 0.17 for native speakers. Function words entail higher intensity variation for non-native speakers compared to native speakers: French (estimate=0.15), Japanese (estimate=0.13). French speakers produce higher intensity variation of frequent lexical words (estimate=0.1) and infrequent words (estimate=0.17) compared to native speakers.

A larger number of characters contributes to an increase of intensity variation of French and Japanese speakers: both estimates equal to 0.03.

L2 level is a significant factor only for French speakers: advanced speakers produce less intensity variation compared to less advanced speakers (estimate=-0.1)

In summary, native speakers realize function words with less intensity variation compared to non-native speakers. The intensity variation increases with a larger number of characters (confirms the Hypothesis 2) and the level advancement (for French speakers).

Duration

In this section, we focus on the duration modeling and the results obtained with the following model:

$$duration \sim wordType + nbrChar + nbrSyll + textLevel + L2Level + (1 \mid speaker) + (1 \mid word) \quad (4.15)$$

Lexical words are longer than grammar words for the speakers of all the languages. What is interesting to notice is that the difference between the duration of grammar and lexical words is larger for native speakers (estimate=0.09) than for French (estimate=0.07) and Japanese speakers (estimate=0.05). It could illustrate the fact that native speakers tend to reduce grammar words when speaking, while non-native speakers of English pronounce them entirely as lexical words. The estimates show that Japanese speakers show less difference between the duration of lexical and grammar words. It can illustrate that they add an epenthetic vowel between the consonants, which can make the words longer. It can also mean the Japanese speakers have difficulty pronouncing both lexical and grammar words.

Both the number of syllables and of characters are important when predicting word duration. As supposed, the word duration increases with a larger number of syllables and characters.

For the native speakers, an increase in one syllable is associated with an increase in duration by 0.05s, and the increase in one character - with an increase in word duration by 0.02s.

For the French speakers, the results are 0.05 and 0.03 respectively.

And for the Japanese speakers, 0.1 and 0.04 respectively.

The analysis shows that the increase in the number of syllables and characters has more influence on the word duration for Japanese speakers. It can also be related to the insertion of an epenthetic vowel that lowers their speech rate and increases word duration.

The advancement in L2 level is a significant factor for French speakers: the word duration decreases by 0.07s with level advancement. It could mean that more advanced speakers read words with more ease. We did not obtain significant results for Japanese speakers.

As in the previous sections, we compared the duration of three types of words among speakers of all three languages. As supposed, the function words are longer for non-native English speakers compared to native speakers:

French speakers spend approximately 0.04s longer than native speakers when pronouncing a grammar word, while Japanese speakers about 0.1s longer.

Lexical words pronounced by non-native speakers are also longer compared to those pronounced by native speakers: estimate for French speakers if 0.04s and 0.08s for Japanese speakers.

Rare words, as being the most challenging ones, have more salient differences between native and non-native speakers. French speakers spend about 0.13s more on rare words than native speakers, while Japanese speakers spend about 0.33s more.

In summary, infrequent lexical and longer words require more processing time by all the speakers. French speakers show improvement with higher L2 level by pronouncing infrequent words faster, while Japanese speakers do not show this trend.

The results confirms the Hypothesis 1 and show that Japanese speakers spend more time processing the words than French and native speakers. It could be due to the insertion of an epenthetic vowel and/or to the difficulty of the English language for Japanese speakers overall.

4.2.1.2 Eye tracking features

In this section, we will investigate the variation of fixation duration depending on the word type.

The gazed words are identified according to the position of the fixation and of the word on the screen, as explained in Chapter 3. The fixation duration is the sum of the duration of all the fixations that the word receives.

The plot 4.17 shows that infrequent lexical words receive longer fixations, while little difference is noticed between frequent lexical and function words. For English native speakers, the difference between the fixation duration on the frequent and infrequent words is less salient than for non-native speakers.

By applying the following LME mode, we examine the fixation duration depending on the word type, number of characters, number of syllables, text level in which the word appears, and L2 level. We also account for individual variation and the variation due to the word itself:

$$totalfixationDuration \sim wordType + nbrChar + nbrSyll + textLevel + L2Level + nativeLang + (1 \mid speaker) + (1 \mid word) \tag{4.16}$$

The model was run on (1) the whole dataset to account for differences due to the L1, and (2) separately on subsets of data depending on the native language to compare the fixation duration within different L2 levels for one native language. The result shows that all the speakers' categories fixate longer on infrequent words compared to other frequent lexical and function words. In particular, French speakers spend about 0.33 s longer on infrequent than on frequent words ($p<0.01$), Japanese speakers fixate on them about 0.75s longer ($p<0.01$), while native speakers about 0.14s longer ($p<0.01$). Both non-native speakers fixate longer on infrequent words compared to native speakers:

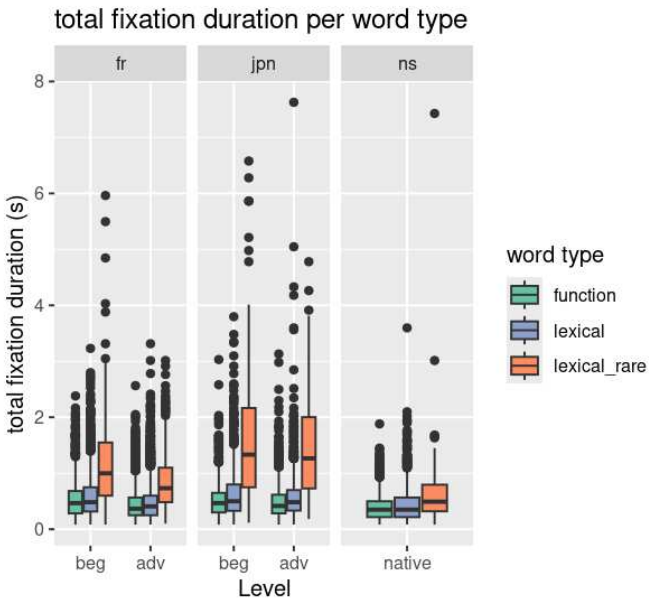

Figure 4.17: Total fixation duration per word type, L2 level, and L1

French speakers' estimate is 0.33 (p=0.01), and Japanese speakers' estimate is 0.9 (p<0.01).

The total fixation duration increases with a larger number of characters in the words for French, Japanese and native speakers. For French speakers, an increase in 1 character is associated with an increase of the fixation duration by 0.038s (p<0.01), while for Japanese speakers by 0.06 (p<0.01) and native speakers by 0.02s(p<0.01).

For French speakers, L2 proficiency impacts the overall fixation duration on words, which decreases with L2 level advancement (estimate=-0.1, p<0.01). However, it is not significant for Japanese speakers. Among beginner French speakers, the difference in fixation time between frequent and non-frequent lexical words is 0.4s (p<0.01), while for advanced speakers, this difference reduces to 0.3s. Japanese speakers also show an improvement with level progression: the difference between the fixations on the frequent and infrequent words is 0.8s for beginner speakers and 0.7s for advanced speakers. The difference in fixation duration between function and frequent lexical words is not significant for the analyzed speakers.

To sum up, both native and non-native speakers realize longer fixations on the infrequent than on frequent words, which confirms the Hypothesis 1. The difference is more salient for non-native speakers, especially for Japanese speakers. French speakers show more improvement with L2 level advancement by using shorter fixations. No significant difference was found between the processing of function and frequent lexical words.

4.2.2 Difficult words

In the previous section, we observed the differences in visual and oral processing of frequent, infrequent lexical and function words by the representatives of different cultures, languages and L2 levels. This section is dedicated

to difficult word processing analysis. The results will help us to get insights into the strategies that the participants implement depending on their cultural and linguistic backgrounds. In future work, this analysis can be implemented for reading challenges anticipation and can serve as a foundation for proposition nudging strategies to improve L2 acquisition process, in particular word pronunciation and comprehension.

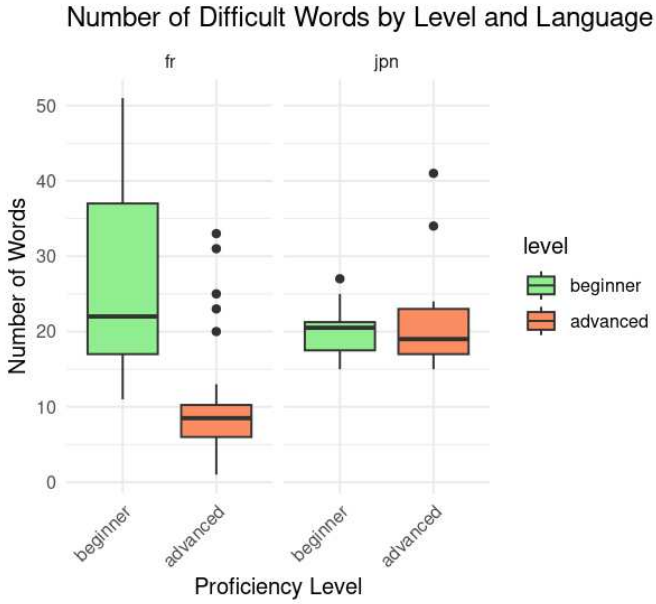

Figure 4.18: Number of difficult words reported by French and Japanese speakers

In the experiment, discussed in Chapter 3, after the reading aloud task, the participants were asked to highlight words which were difficult to understand and/or to pronounce (see Chapter 3 about the experimental protocols). These words represent the target of the present section and are referenced hereafter as “difficult words”. Unfortunately, at this work stage, no distinction was made between the challenges related to the phonetic or the semantic decoding, so the word reading challenge will be considered from a global perspective. Future work could provide a more fine-grained analysis including the distinction between reading challenges as function of linguistic levels (phonetic, syntactic, semantic etc.).

The challenging vocabulary reported by the participants includes such words as:

deleterious, inconsequential, bedevil, abhor, singsong, acquiesce, boisterous, etc.

Most of the difficult words correspond to infrequent lexical words and appear in the advanced text (see Appendix A).

We start the analysis by calculating the number of difficult words reported by French and Japanese speakers depending on their L2 level. The boxplot 4.18 illustrates that for the French participants, the number of difficult words decreases with the L2 level advancement. However, Japanese speakers report a similar number of difficult words regardless of their L2 level. The mean number of words reported by beginner French and Japanese speakers is similar. However, beginner French speakers show more variability.

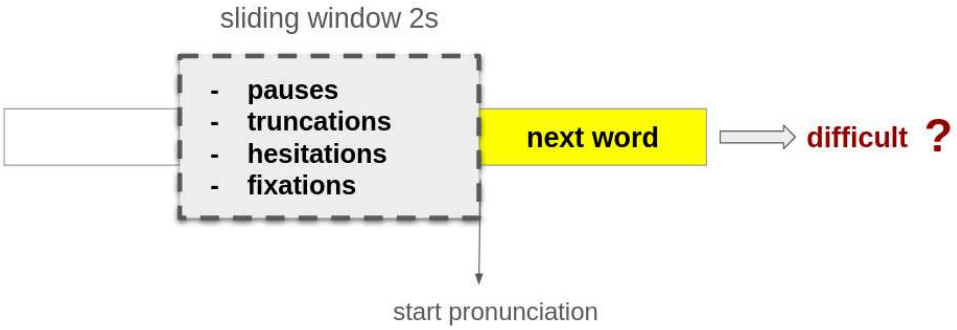

Figure 4.19: Difficult words analysis: sliding window

In order to obtain an in-depth understanding of the strategies implemented by the speakers in order to anticipate the pronunciation of difficult words and to determine the difference from those applied before other non difficult words, we relied on a sliding window technique.

Before the start of pronunciation of each word (which was determined according to the manually corrected boundaries), we took a window of 2 seconds and calculated the corresponding eye tracking and speech features 4.19. The size of the window was close enough to the pronounced word and was sufficient to determine the difference in visual and oral processing of difficult and non-difficult words, as will be shown in the analysis below.

Only intermediate and advanced texts were used for the analysis, because no difficult word were reported in the beginner text. As for the previous analysis including eye-tracking features, speakers with poor calibration were not included in the analysis. In total, using the sliding window technique, we obtained 675 samples of difficult and 20010 of non-difficult words for French speakers, and 287 and 7103 for Japanese speakers respectively (Table 4.2])

language	nbr of samples with difficult words	nbr of samples with non-difficult words
French	675	20010
Japanese	287	7103

Table 4.2: Sample count using sliding window technique

Then, we calculated the number of cases in which the participants used disfluencies. The analysis shows that pauses are the most commonly used disfluencied before both difficult and non-difficult words (more than 60% of cases), but they are more frequent before difficult words 4.20. The truncations 4.21 and hesitations 4.22 are less frequently used: truncations being used in less than 20% of cases and hesitations - less than 10%. However they are more frequent before difficult words than non-frequent ones.

The pauses are longer before difficult than non-difficult words for both French and Japanese speakers 4.23. The plot shows that the difference between pauses duration before two types of words is less salient for French speakers, and the pauses become shorter with L2 level advancement 4.23.

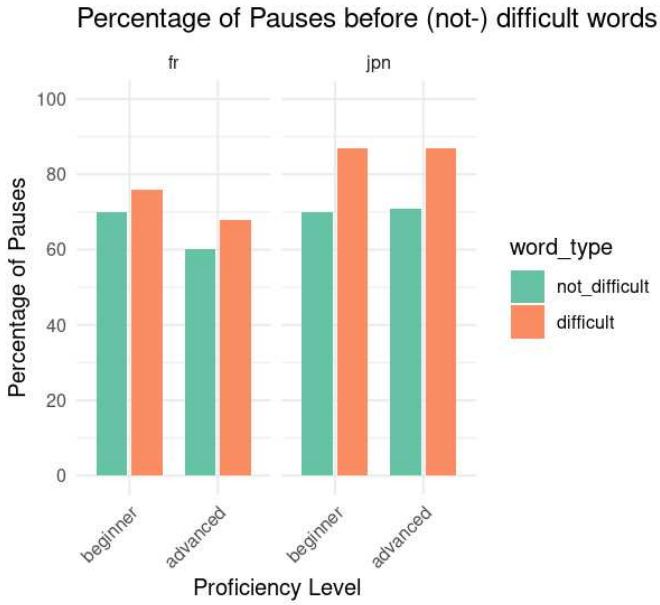

Figure 4.20: Percentage of cases with pauses before (non-)difficult words

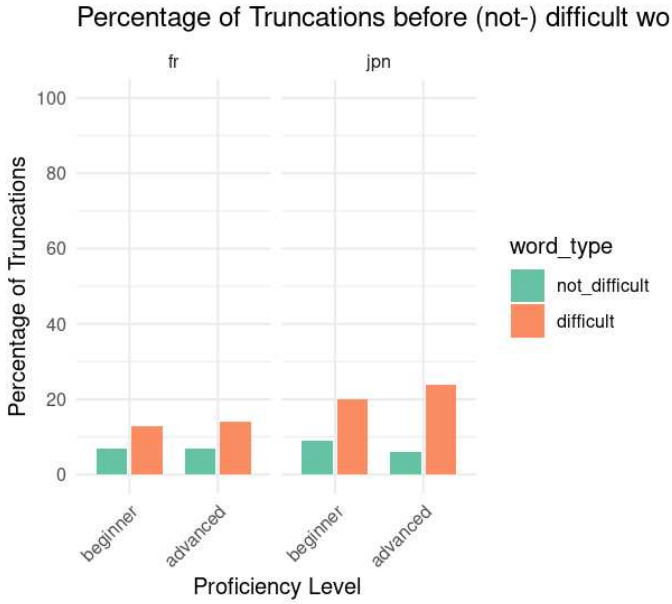

Figure 4.21: Percentage of cases with truncations before (non-)difficult words

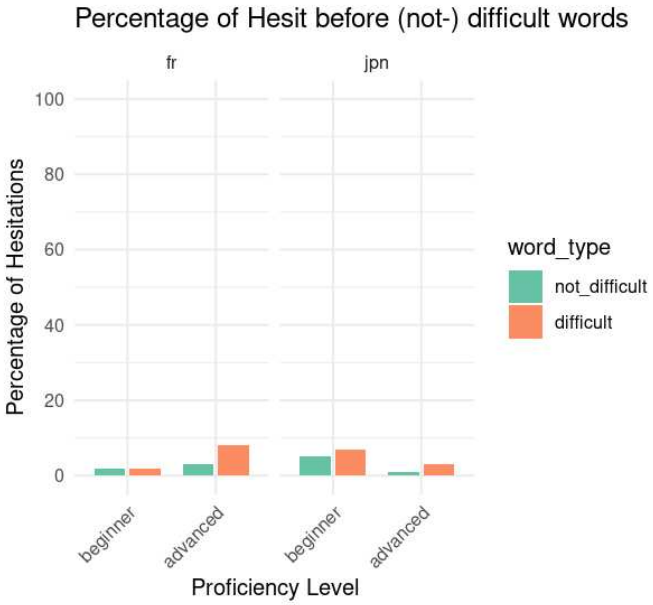

Figure 4.22: Percentage of cases with hesitations before (non-)difficult words

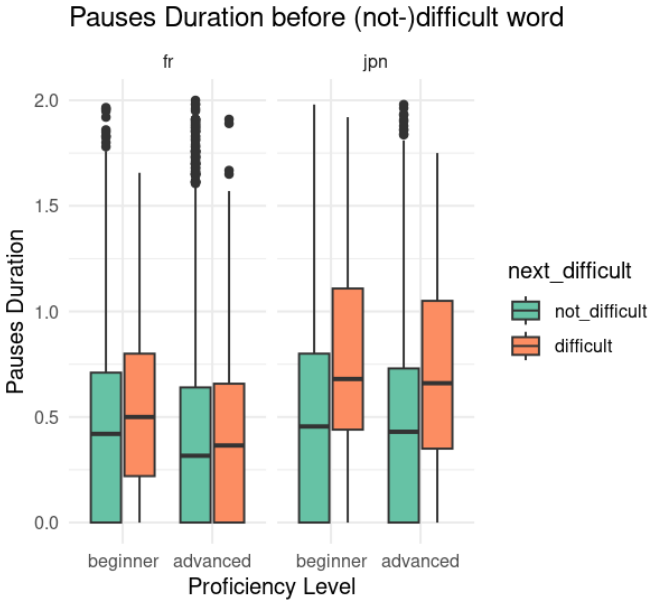

Figure 4.23: Pauses duration before (non-)difficult words

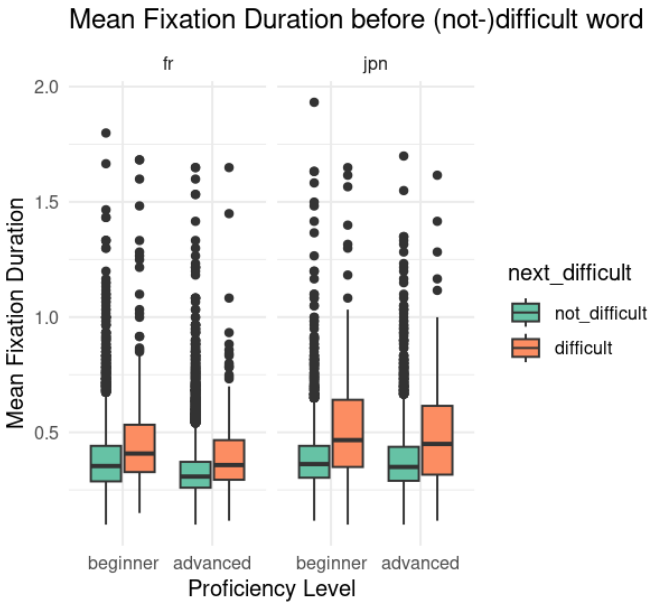

Figure 4.24: Mean fixation duration before (non-)difficult words

As for the eye movement, the fixations are longer 4.24 and less frequent 4.25 before difficult words. It can be related to the fact that when the word is difficult, the participants tend to fixate longer on the same word, so the number of fixations is smaller. It is in accordance with the results obtained in the previous section: the infrequent words which are also commonly considered as difficult, receive longer fixations.

Summary

Tu sum up, we observed that the oral and eye movement behaviors linked to the occurrence of a difficult word in the text by comparison with non difficult words are different. The statistical differences described above allow us to hypothesize that both oral and eye movement features can be used for automatic prediction of reading challenges in future work.

However, one needs to keep in mind that the results obtained and described in this section can be improved by adding more data, especially for Japanese speakers, and by adding the information about pupil dilation which reflects cognitive load.

4.3 Phone level

In the previous sections, we investigated the differences in speech and eye tracking features between L1 and L2 speakers of different proficiency levels. The analysis was conducted at global textual level and at word level. It enabled us to get insights into reading challenges that the participants faced during the experiment, their strategies of coping with them (such as the use of disfluencies and longer fixations) and to correlate the results with their linguistic background. This section is dedicated to a more fine-grained analysis at the phone level.

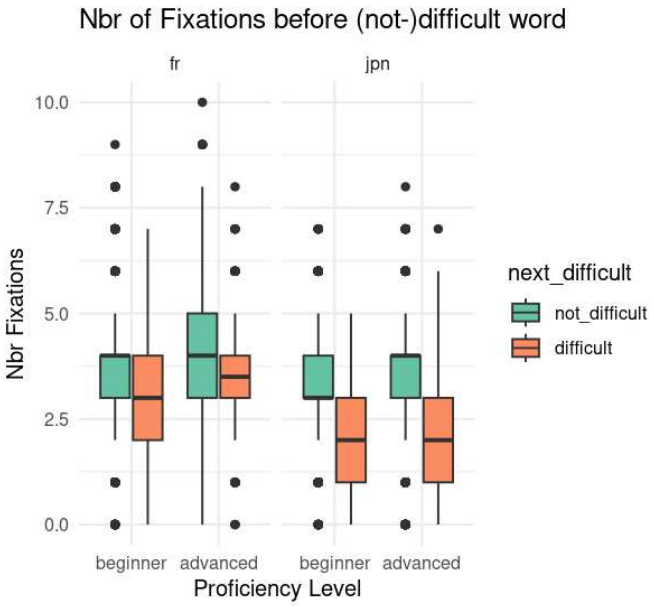

Figure 4.25: Number of fixations before (non-)difficult words

According to the reviewed literature in Chapter 2, L1 influences the pronunciations of L2 sounds. The goal of this section is to estimate at which extent L1 influences the L2 in our protocol and which factors are the most significant: culture and/or the L2 proficiency level. Furthermore, our goal is to estimate if fine-grained phonological characteristics can be easily implemented in ML models for L2 level detection.

As discussed in Chapter 3 about data acquisition and processing, manual correction of word and phoneme boundaries was necessary due to a high number of errors made by the forced alignment system. Given that manual correction is a time consuming process, the analysis of this section is focused only on a pair of vowels /ʊ/ lax and tense /u/. Tense vowels are usually longer than lax ones and require more articulation tension to be produced. The distinction between tense and lax vowels is absent from both French and Japanese languages, so we expect the speakers to substitute them with vowels from their phonological inventories (see Chapter 2). In this analysis, we examine whether the speakers pronounce these vowels more distinctly with L2 level advancement.

The vowel space below 4.26 shows the location of the vowels in English (black), Japanese (red) and French (blue). The axes of the vowel space are represented by the first two formants: F1 on the vertical axis which is related to the degree of openness (the higher the F1 the more open is the vowel); and F2, on the horizontal axis responsible for the degree of backness (the higher the F2 the more front is the vowel). According to the plot, the Japanese vowel /ɯ̟/ is more central than the French vowel /u/ and has a higher F2. The English lax vowel /ʊ/ has higher F1 compared to tense English /u/, French /u/ and Japanese /ɯ̟/. Previous studies examined the role of L1 and L2 level in tense/lax vowel distinction. For example, (Schwartz, 2019) found that more advanced English speakers used voice quality differences to distinguish between tense/lax vowels while it was not the case for less proficient speakers. The age of language acquisition influences the phoneme pronunciation intelligibility (Flege, 1992), in particular the distinction

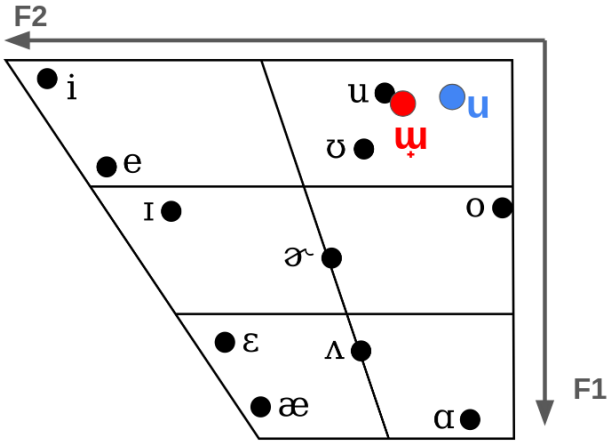

Figure 4.26: [u] location in the vocal space of English (black), Japanese (red) and French (blue) languages

of tense and lax vowels (Chang and Weng, 2013). Native English speakers rely on formant characteristics rather than on the duration to recognize tense/lax vowels (Bohn and Flege, 1990). However, for Japanese speakers, the duration characteristics are more salient (Tsukada, 2009, as cited in De Weers and Munro, 2018).

Based on the literature and the comparison of vowel spaces of English, French and Japanese 4.26, the following hypotheses are introduced:

Hyp. 1: Beginner speakers neutralize the vowel distinction more than the advanced speakers (vocal space ellipses of beginner speakers overlap more than that of advanced ones)

Hyp. 2: Given the observed vocal spaces, we expect that Japanese speakers pronounce English /u/ with higher F2 compared to French speakers.

Hyp. 3: Japanese high back vowel /ɯ/ is unrounded but French high back vowel /u/ is rounded. Lip rounding lowers F3, so we expect French speakers to pronounce the /u/ sound with lower F3

Hyp. 4: Japanese speakers use duration rather than formant frequencies for tense/lax distinction

To verify these hypotheses, we extracted the formants F1-F3 and duration of the corresponding vowels using a Praat (Boersma and Weenink, 1992–2022) script [2]. The script extracts the F1-F3 measures every 5% of the vowel, that is 20 measures per vowel. In order to avoid coarticulation at the beginning and at the end of the vowel, the measures were taken only at midpoint. Then, the outliers were deleted using 2*standard deviation filtering by speaker. The initial and filtered number of data samples are presented in Table 4.3. The proportion of filtered data does not exceed 2%.

The resulted measures of F1 and F2 were plotted to represent /u/, /ʊ/ vowel space by language and by level. The Figure 4.27 represents French native speakers divided into two levels: beginner (A2_B1) and advanced (B2_C1_C2); the Figure 4.28 represents Japanese speakers; and the Figure 4.29 - native speakers.

As expected, the vowel space of native speakers shows clearer distinction between the two vowel timbres, com-

[2]written by Katherine Crosswhite, Changed by Andries W Coetzee (8/15/2012)

speakers	phone	nbr before filtering	nbr after filtering	nbr filtered	% filtered
fr	u:	1812	1773	39	2%
	ʊ	575	574	1	0,2%
jpn	u:	514	506	8	1,6%
	ʊ	235	233	2	0,8%
native	u:	447	439	8	1,8%
	ʊ	152	151	1	0,6%

Table 4.3: Number of vowels before/after filtering

pared to French and Japanese speakers whose plot shows more overlap.

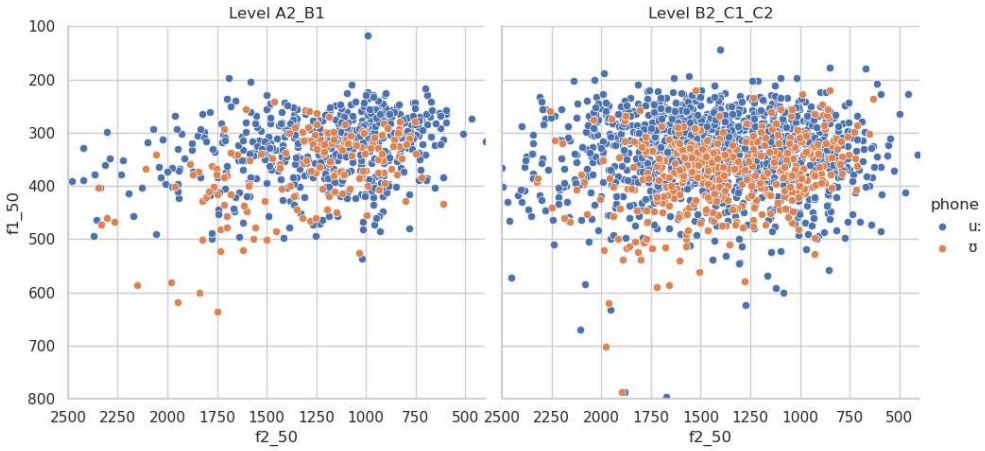

Figure 4.27: Vocal space of /u/ and /ʊ/ of French speakers

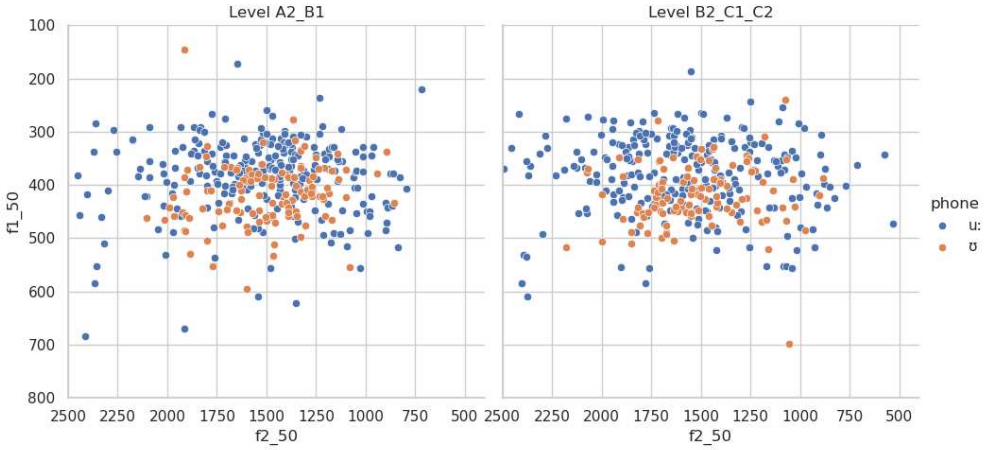

Figure 4.28: Vocal space of /u/ and /ʊ/ of Japanese speakers

In order to calculate the vowel space overlap, we relied on the Pilai score (Pillai, 1955), which is a commonly used measure used for this purpose (Kelley and Tucker, 2020; Nycz and Hall-Lew, 2014). Higher value means that there is less overlap between the vowel spaces, lower value corresponds to increased overlap.

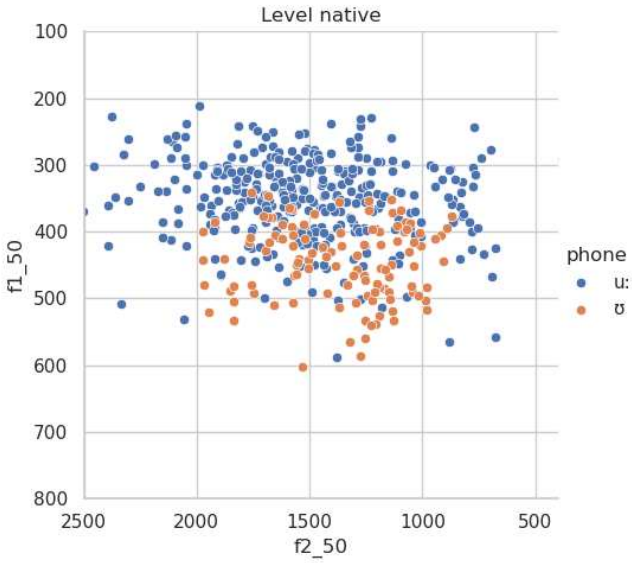

Figure 4.29: Vocal space of /u/ and /ʊ/ of English speakers

For its calculation, we relied on the script provided in the tutorial by Joey Stanley [3]and the MANOVA test in R. The obtained results are reported in Table 4.4. The results show that, the native speakers distinguish better the /u/ vs /ʊ/ specific timbres, as they have the largest Pilai score (0.4), followed by Japanese (0.07) and then by French speakers (0.04). French speakers show an improvement of lax/tense distinction with level advancement: we observe a higher Pilai score for more advanced French speakers of English. However, this result is not confirmed by Japanese speakers of English, for whom the score is slightly lower with level improvement (0.07 for advanced speakers and 0.08 for beginners). Although the distinction between lax/tense vowels improves with L2 level progress for French speakers, the distinction between these two vowel categories is more salient for Japanese speakers.

Native lang	Level	Pilai score	Significance
NS		0.4	***
FR	all	0.04	***
	A2_B1	0.02	*
	B2_C1_C2	0.05	***
JPN	all	0.07	***
	A2_B1	0.08	***
	B2_C1_C2	0.07	**

Table 4.4: Pilai score results: vowel space overlap

The difference in the phone distinction by the speakers of the three languages is also observable in the boxplots 4.30 and 4.31 Indeed, the boxplot shows that English native speakers clearly distinguish the two phones by both F1 and F2, while French and Japanese speakers have larger overlap for both formants. The boxplot shows a smaller vowel overlap for Japanese speakers suggesting that they distinguish these two vowels better than French speakers.

The results of the Pilai score and the boxplot observation confirm only partially our first hypothesis (**Hyp. 1**). The

[3]https://joeystanley.com/blog/a-tutorial-in-calculating-vowel-overlap/

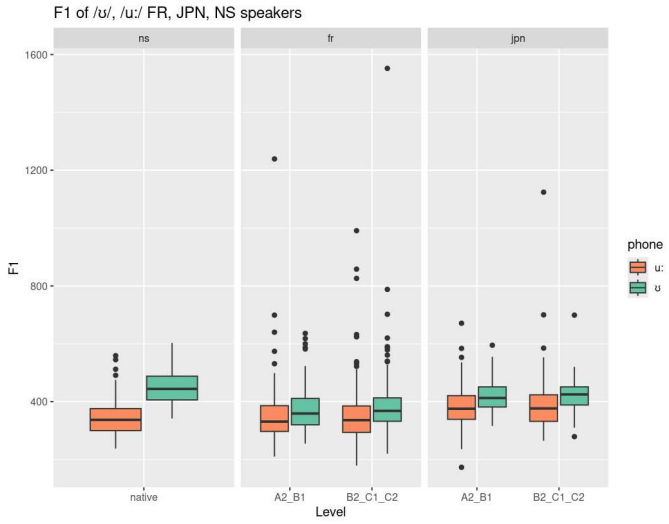

Figure 4.30: F1 of /u/ and /ʊ/ realized by French, Japanese and English speakers

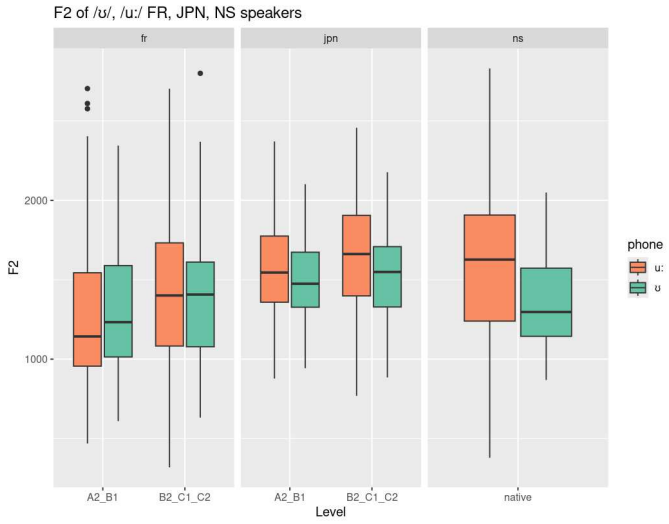

Figure 4.31: F2 of /u/ and /ʊ/ realized by French, Japanese and English speakers

fact that more advanced speakers distinguish the phones better was not confirmed for Japanese speakers. It could be explained by a smaller number of data of Japanese speakers and needs further investigation.

In order to investigate in detail the vowel production and its correlation with the native language and the L2 level, LME models are applied in R. The formant (F1 or F2) is used as a dependent variable, L1 and L2 level as fixed effect, word and speaker as random factors. The model was run separately for each subset of data depending on the speaker's native language, that is native English, L2 English by French speakers, L2 English by Japanese speakers.

$$F1 \sim phone + nativeLang + L2Level + (1 \mid speaker) + (1 \mid word) \tag{4.17}$$

$$F2 \sim phone + nativeLang + L2Level + (1 \mid speaker) + (1 \mid word) \tag{4.18}$$

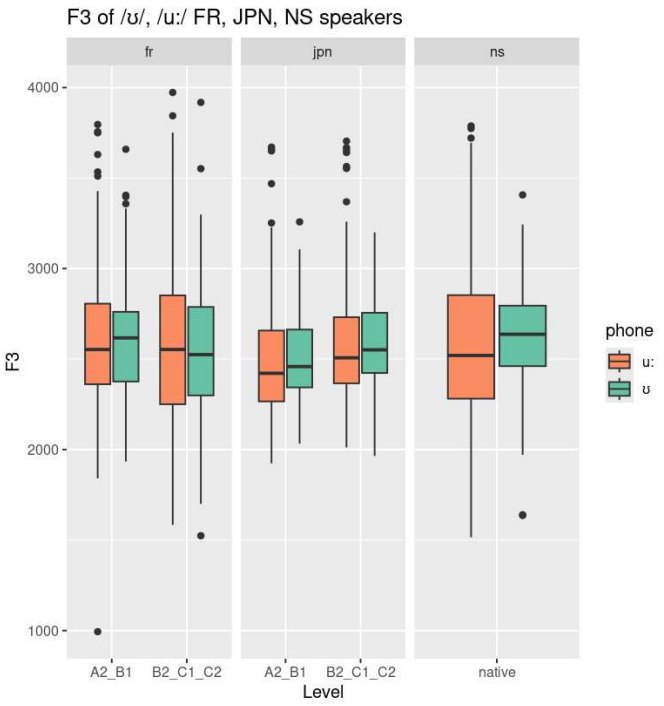

Figure 4.32: F3 of /u/ and /ʊ/ realized by French, Japanese and English speakers

The results show that the F1 of /ʊ/ produced by native speakers of English is 93 Hz ($p<0.01$) higher than /u/ and there is no significant difference in F2 between two vowels. The result is in line with the vowel location on the vowel space 4.26 and the boxplots 4.30 and 4.31: there is a larger overlap of F2 than F1 on the boxplots and the difference between these vowels is more salient on the F1 axis than on F2 axis in the vocal space 4.26.

French speakers realize /ʊ/ with higher F1 than that of /u/ (estimate=30, $p<0.01$). The F2 of /ʊ/ is 163 Hz higher for more advanced levels ($p<0.01$).

As for the Japanese speakers, no significant effect was found neither for F1 nor for F2 realization, which can be due to a small dataset with high variability.

The comparison of Japanese with French speakers' production shows that F1 of Japanese speakers is 48 Hz higher than that of French speakers for /u/ and 44 Hz higher for /ʊ/. Moreover, F2 of Japanese speakers is 176 Hz higher than that of French speakers for /ʊ/ and 237 Hz higher for /u/, which confirms the Hypothesis 2.

No significant results were found for F3 neither for Japanese nor for French speakers. The boxplot 4.32 does not display either significant differences. Therefore, the Hypothesis 3 advancing the role of lip rounding is not confirmed.

We verified the correlation of the phone duration with the phone type using the T-test in R and found out the following results.

For all the speakers, the mean duration of /u/ is longer than that of /ʊ/, which supports the observation that tense vowels (/u/ in our case) are longer than lax vowels (/ʊ/). Both native and non-native English speakers are able to make the difference between these two vowels based on the duration pattern. More precisely, for the native

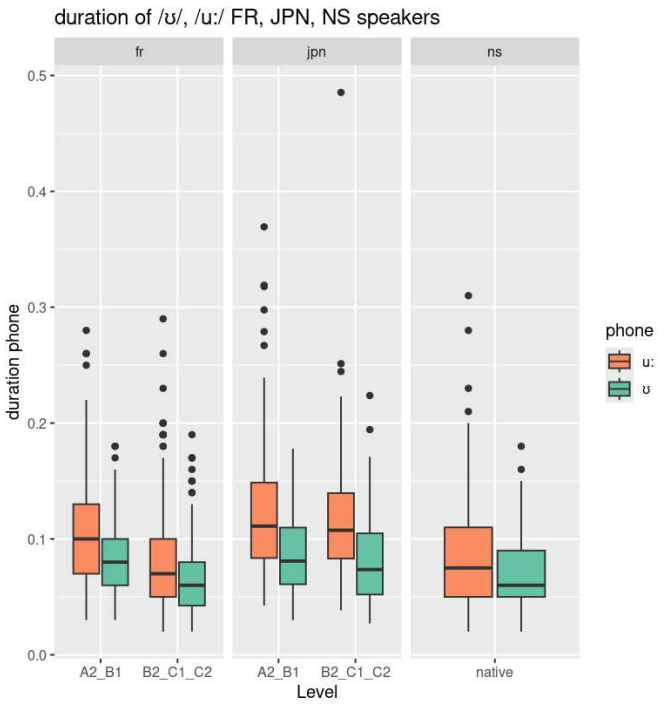

Figure 4.33: Vowel Duration of /u/ and /ʊ/ realized by French, Japanese and English speakers

speakers the mean /u/ duration is about 0.017s higher than /ʊ/ duration ($p<0.01$); for the French speakers it is about 0.014s higher ($p<0.01$) and for Japanese about 0.04 higher ($p<0.01$). The analysis reveals that the difference in the duration of the lax/tense vowels is more salient for Japanese speakers, which confirms our fourth hypothesis and the study of (Tsukada, 2009, as cited in De Weers and Munro, 2018) suggesting that Japanese speakers rely more on duration characteristics to distinguish between tense/lax vowels.

As for the difference in duration within the L2 levels, the more advanced French speakers pronounce shorter /u/ and /ʊ/ than the beginners which supports the hypothesis that more advanced speakers read faster. However, this is not the case for Japanese speakers, for whom the change in vowel duration within the levels is not significant. This is also visible on the boxplot 4.33.

Summary

To sum up, we conducted a comparative analysis of a pair tense-lax vowel realization: /u/ and /ʊ/ by French, Japanese and native English speakers. The distinction between tense/lax vowels is absent in both French and Japanese languages, therefore, our goal was to investigate whether the participants are able to distinguish theses vowels and which features they rely on to make the distinction (formant or duration). Our first hypothesis concerning the vowel space overlap depending on the L2 level is partially confirmed. The analysis shows that French speakers are able to better distinguish between tense and lax vowel with L2 level advancement, however this was not the case for Japanese speakers. The second hypothesis was confirmed: Japanese speakers pronounce both vowels with higher F2 compared to French speakers. The hypothesis 3 concerning the lip rounding translated by the vowels' F3 is not

confirmed: this parameter was not significant neither for French nor for Japanese speakers. Finally, our fourth hypothesis was confirmed: Japanese speakers rely on duration characteristics rather that on formants to distinguish between tense/lax vowels. In conclusion, we observed differences in tense/lax vowel realization by French and Japanese speakers, and we will include the vowel characteristics into the prediction model (Chapter 5) in order to examine whether they contribute to the L2 level prediction when combined with other speech and/or eye-tracking features. Future work consists in enriching the present analysis by adding other vowels and consonants to get a better understanding of the pronunciation challenges faced by French and Japanese speakers and their progression with L2 level advancement.

Chapter 5

L2 level prediction based on speech and eye tracking features

In Chapter 3, we described several experimental protocols implemented to address our research questions. More specifically, we collected speech and eye tracking data from French and Japanese speakers of English, as well as a control set of corresponding data produced by native English speakers, using a reading aloud setup allowing to simultaneously record the two modalities. Then, we examined a range of features extracted from speech and eye movement, that we correlated with the participants' self-reported L2 proficiency levels, text comprehension scores, and native language. The results reported in Chapter 4 show that both eye tracking and speech patterns correlate at different levels of significance with the participants' L2 level and mother tongue. In the present chapter, we explore whether these features, when combined, can contribute to an automatic prediction of L2 proficiency. Moreover, we identify the most salient predictors and examine their variation based on the participants' L1 background.

Our hypotheses are as follows:

Hyp 1: A combined feature set that integrates both eye-tracking and audio characteristics is more effective for L2 level prediction than using these feature sets independently.

Hyp 2: The L1 filter as a marker of cultural differences reflects in the prediction results: the features contributing the most to the L2 level detection differ for French vs Japanese speakers.

5.1 Data pre-processing

To test these hypotheses, we created three subsets of features (audio, eye-tracking, and a combination of both) and applied them to three datasets: Japanese speakers, French speakers, and a combined group of Japanese and French speakers. The target is represented by the 2 levels, ranging from A2 to C2, grouped into two categories: A2_B1

audio feature	eye movement features	other
number of pronounced syllables (including disfluencies)	number of fixations	text level
number of units pronounced (words + oral disfluencies)	total fixations duration	page number
number of pauses (>200ms)	mean fixations duration	gender
pauses total duration	std fixations duration	
pauses duration std	number of forward saccades	
speech duration	number of backward saccades	
phonation duration (speech duration - pauses duration)	total forward saccades duration	
speech rate (number of syllables/duration)	total backward saccades duration	
articulation rate (number of syllables/phonation duration)	mean forward saccades duration	
F1 mean of /u/	mean backward saccades duration	
F1 std of /u/	std forward saccades duration	
F2 mean of /u/	std backward saccades duration	
F2 std of /u/	mean length of forward saccades	
mean duration of /u/	mean length of backward saccades	
std duration of /u/	std length of forward saccades	
F1 mean of /ʊ/	std length of backward saccades	
F1 std of /ʊ/		
F2 mean of /ʊ/		
F2 std of /ʊ/		
mean duration of /ʊ/		
std duration of /ʊ/		
pitch mean in Hz		
pitch std in Hz		
pitch mean in semitone		
pitch std in semitone		
intensity mean		
intensity std		

Table 5.1: Feature sets for L2 level prediction (in red deleted redundant features)

(beginners) and B2_C1_C2 (advanced).

These features were extracted inside the boundaries of a page of text based on the manually corrected speech alignment. Then, the features were assembled into vectors, each corresponding to a page read by a speaker. Each page contains 6 lines, one line including up to 24 words. Each speaker contributes 8 vectors: 2 vectors for the beginner text (excluding the third page due to its brevity: it includes one line), and 6 vectors for the intermediate (3) and advanced texts (3).

In total **42 features** were extracted as mentioned in Table 5.1 and among them:

- **26 speech features**, including those corresponding to prosodic measures, vowel pronunciation, verbal disfluencies and silent pauses. Among the audio features, the number of pronounced syllables and the total pauses duration were excluded due to the redundancy with speech rate and phonation duration.
- **16 eye movement features**, including duration and number of fixations, saccades, and regressions

The feature sets for the classification are listed in Table 5.1

As discussed in Chapter 3 about data collection and processing, the original dataset was reduced due to poor eye tracker calibration for some speakers. Specifically, the data of 3 French and 9 Japanese speakers was dropped

Language	Beginner Level		Advanced level		Total nbr vectors
	Nbr speakers	Nbr vectors	Nbr speakers	Nbr vectors	
French	13	104	30	240	340
Japanese	6	48	9	72	120

Table 5.2: Number of speakers and vectors used for L2 level prediction

from the eye movement analysis provided in Chapter 4. In the present Chapter, in order to maintain consistency in results comparison, the same speakers were excluded when evaluating the model using audio features. As a result, we obtained 340 vectors for French speakers and 120 for Japanese speakers. The detailed number of speakers and vectors per language and per level are presented in Table 5.2

5.2 Model selection

The research community has explored several approaches to dealing with L2 level prediction tasks using audio. Early methods for modeling speech relied on designing low-level features based on vocal signals to capture the acoustic modulations in speech; including statistics on pitch, rhythm, individual vocal components, and slopes.

In order to predict speakers' L2 level with a combined feature set that integrates both eye-tracking and audio characteristics, we selected the Random Forest Classifier. This model was chosen for its ability to handle non-linear data, making it well-suited for our classification task. It is also suitable for small datasets, such as the one that we used. Additionally, it provides information related to feature importance which is essential in our analysis, as we are interested in investigating which features contribute best to the model decision when defining L2 level.

The default Random Forest model was applied using scikit-learn (Pedregosa et al., 2011) Python library, where the number of trees (or estimators) is 100, the number of features considered for splitting corresponds to the square root of the total number of features, the maximum depth is 8. To choose better hyperparameters, we used the grid search technique. However, the default model with the parameters reported above, provided better performance.

We applied a stratified k-fold cross-validation technique in order to evaluate the model on different subsets of data and to ensure that it is able to generalize to unseen data. The data was split into 5 folds of equal size, each fold was used once as the test set, while the remaining folds were used to train the model. For the data splitting, we ensured that the data related to one speaker appeared only in the train or test set, but never in both sets. Additionally, to maintain consistent distribution data representation, we applied stratification by L2 level.

The model performance was evaluated on each fold using the F-score. The metrics were calculated overall at the level of individual vectors, each of them representing one page read by one speaker, as well as per speaker, employing the majority vote technique (Figure 5.1). Below we report weighted F-score, as this metric is more adapted for unbalanced datasets, accounting for the size of each class. We calculate its average value across the 5 folds and the standard deviation.

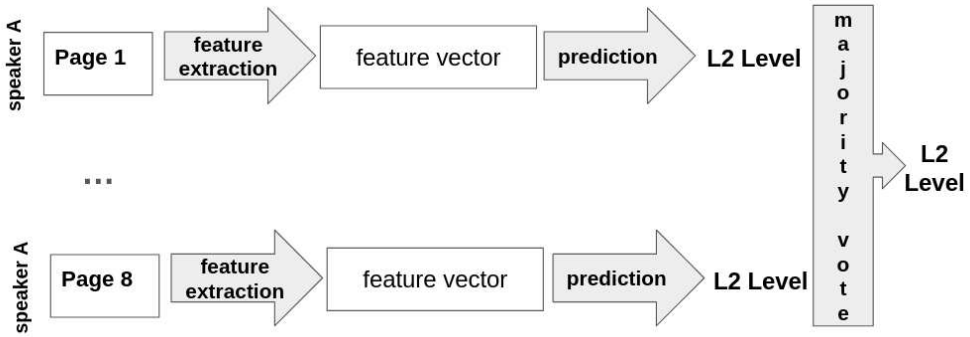

Figure 5.1: L2 level prediction per page and per speaker

In order to provide a comparison to our model, we provide a baseline model. For this purpose we used a Dummy Classifier from scikit-learn library (Pedregosa et al., 2011) that does not take into account any features or correlations in the data, but consistently predicts the majority class. In our case, the majority class for both French and Japanese speakers is the advanced one, which is represented by the speakers having B2, C1 and C2 levels. The performance of the baseline model was evaluated using the average F-score calculated across the same 5 folds of the data, as when using the Random Forest model. The results of the baseline model are also presented per vector (per page) and by speaker using the majority vote technique. The results of the baseline model were added as a comparison to all the results in the next sections of this chapter.

5.3 Results: all features

The Table 5.3 illustrates the performance of the baseline model compared to the Random Forest model when using speech, eye-tracking features or their combination. At this stage, no feature selection technique was applied which means that all the extracted speech or eye-tracking characteristics were used. The baseline model shows an F-score of 59% for French speakers when using per vector (per page) classification and 82% when using the majority vote technique. As for the Japanese speakers, the average F-score per vector is <50% which is lower than a random chance, and the performance per speaker is 72%. The baseline model performed worse than a random chance (<50%). The results obtained with Random Forest classifier presented in Table 5.3 indicate that the model can predict L2 proficiency level of French speakers with an overall F-score of 72% using audio features, 75% when using eye tracking features, and 74% when using a combination of speech with eye tracking features, which outperforms the baseline model. However, the result of the baseline model calculated per speaker outperform our model. As for Japanese speakers, the model obtained an F-score lower than 50% when using audio features, 61% with eye tracking features, and less than 50% with a combination of speech and eye tracking features. For Japanese speakers, the model used with eye-tracking features outperforms the baseline used for per vector (per page) classification.

Our model's performance for French speakers is homogeneous among different feature sets, while for Japanese speakers, eye tracking features seem to be more efficient than audio features or the combination of both. However, the results obtained using eye movement features are better for French speakers (75%) compared to Japanese speakers (61%). This can be due to a smaller amount of Japanese speakers' data both for training and testing sets. Indeed, splitting data using stratified 5-fold cross validation resulted in about 34 French speakers in the training set and 9 French speakers in the test set, while for Japanese speakers there were 12 in the training set and 3 in the test set. The data is also unbalanced in terms of the number of speakers per L2 level, advanced speakers representing the majority for both French and Japanese speakers.

Lang	Baseline		Audio Features		Eye Movement Features		Audio + Eye Features	
	Per Vector	Per Speaker (Majority Vote)	Per Vector	Per Speaker (Majority Vote)	Per Vector	Per Speaker (Majority Vote)	Per Vector	Per Speaker (Majority Vote)
FR	0.59 ± 0.17	0.82 ± 0.09	0.72 ± 0.05	0.69 ± 0.07	0.75 ± 0.03	0.77 ± 0.02	0.74 ± 0.03	0.77 ± 0.06
JPN	<0.50	0.72 ± 0.19	<0.50	<0.50	0.61 ± 0.13	0.61 ± 0.07	<0.50	<0.50

Table 5.3: Results of L2 level prediction without feature selection

On the contrary to our expectations, the combination of audio with eye tracking features did not contribute to the results' improvement. For French, the results are similar to those obtained with audio or eye tracking feature sets separately and represent the F-score of 74%. In case of Japanese speakers, the combination of two feature sets worsens the results compared to those obtained with eye tracking features only: F-score of 61% for eye tracking features versus F-score lower than 50% for a combination. These results are however preliminary, as we hypothesize that they can be improved by using a feature engineering technique discussed in the next section.

As observed in Chapter 4, it is challenging to select speech features reflecting the progression from beginner to more advanced levels for Japanese speakers that are comparable to the ones characterizing the French group of speakers. For example, across different L2 levels, the number and duration of pauses is rather homogeneous, small variation of the speech rate (in the sens of expected higher rate for more advanced speakers) is also observed, and the variation in vowel pronunciation is not significant with higher L2 level. One can hypothesize that such patterns are related to a slower progression, but also they can be simply a language-dependent pattern: the progression may be correlated to other measures and/or other linguistic levels that we did not consider in our study and can be the result of cultural differences. A hypothesis interesting to explore in further work is the in-depth correlation between the reading and speaking patterns shown by Japanese speakers and the linguistic distance between English and Japanese languages, both genetic and typological. Such distance studied in both historical and typological linguistics potentially correlates with processing time and may translate in slower progression of "naturalness" with higher L2 levels. From a prediction standpoint, we hypothesize that in order to improve the results, other audio metrics should be used to assess their L2 level.

Features that contributed the most to the model's decision-making differ depending on the speakers' mother

tongue. For example, concerning audio features, for French speakers (Figure 5.2), articulation rate and speech rate are the two most salient factors influencing model's decision and represent respectively 16% and 11% of importance. Among other important features, one can mention the number of pauses, the number of pronounced units (which also includes disfluencies such as truncations, hesitations and repetitions), the realization of /ʊ/ vowel' second formant (F2), its mean and standard deviation, and the vowel duration. These features were also found significant when we analyzed speech characteristics' correlation with L2 proficiency level in Chapter 4.

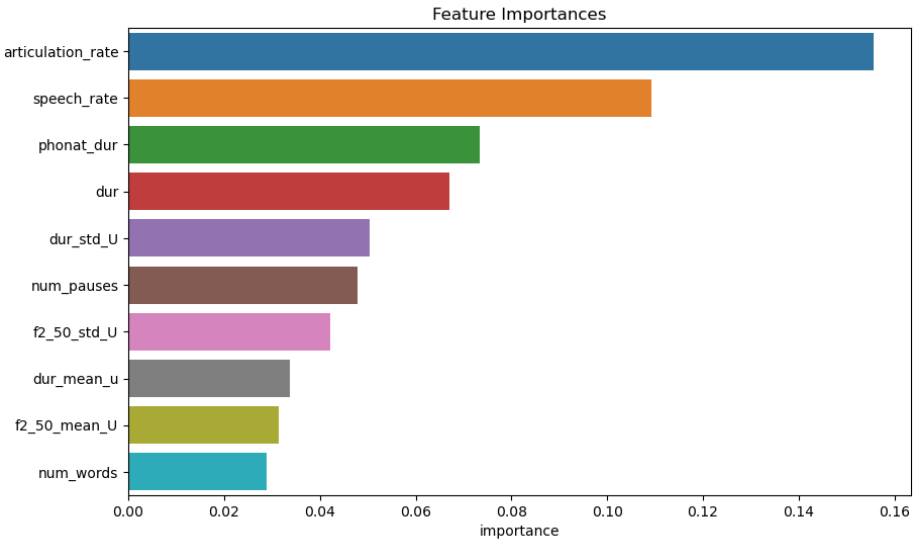

Figure 5.2: Important audio features for L2 level prediction (French speakers)

As for Japanese speakers (Figure 5.3), the most important feature is the standard deviation of the third formant of the /ʊ/ vowel, contributing up to 17,5% to the model's decision. This feature is related to lip rounding, and was not found to be an important factor correlated with L2 proficiency neither for Japanese nor for French speakers. Among other influencing features, one can also mention the standard deviation of the vowel /ʊ/'s F1, F2 and the duration of the vowel. We hypothesize that the L2 level identification for Japanese speakers is related to their pronunciation stability, such as the stability of the formant production and vowel duration. However, the model based on audio features does not show enough precision for Japanese speakers, the results, including the influential features, need further investigation and improvement.

As for the eye movement feature for the French speakers (Figure 5.4), the model reports the following most important features: total saccade duration (13%), total fixation duration (11%), mean fixation duration (11%), mean saccade duration (10.5%). The average and standard deviation of the backward saccades (regressions) length contribute about 5% to the model decision. For the Japanese speakers (Figure 5.5), the most important feature is standard deviation of saccade length, representing 10% of the model's decision. The following 5 features are related to backward saccades (regression) and contribute between 5-7% to the model's decision.

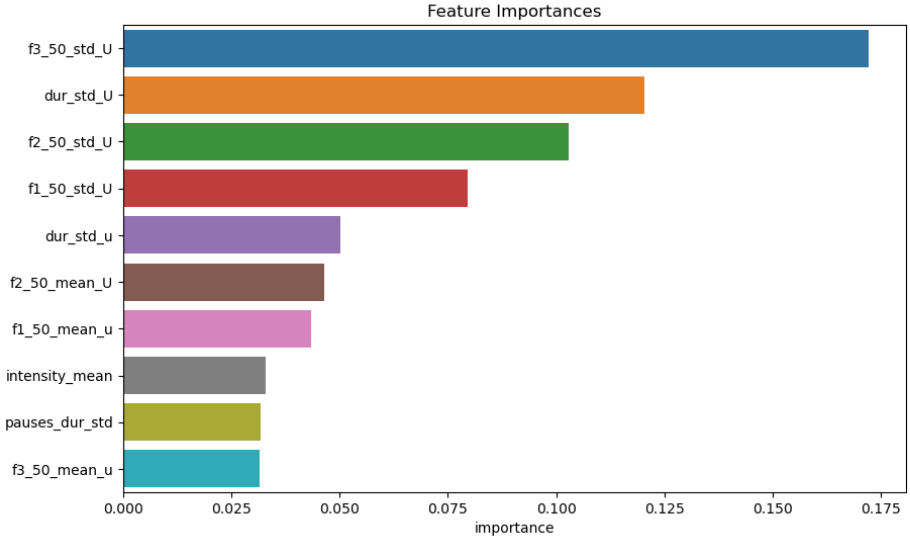

Figure 5.3: Important audio features for L2 level prediction (Japanese speakers)

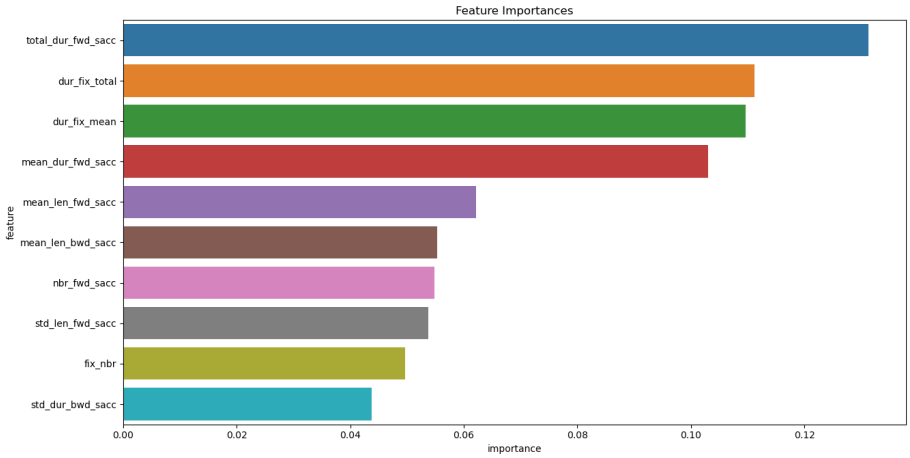

Figure 5.4: Important eye tracking features for L2 level prediction (French speakers)

Summary

To sum up, the preliminary results obtained using Random Forest model including all the extracted features (speech and/or eye tracking) show that both kinds of features are efficient for L2 level prediction of French speakers. However, for Japanese speakers, speech feature set shows poorer performance compared to eye-tracking feature set. While eye tracking features seem to contribute better to the model performance for Japanese speakers, the results illustrate high variability among the folds. We hypothesize that these results can be due to a smaller dataset and/or to the feature set which is not enough representative of the Japanese speakers' performance across different L2 levels. We hypothesize that the results can be improved for both speaker groups by applying feature selection technique, which will be discussed in the next section.

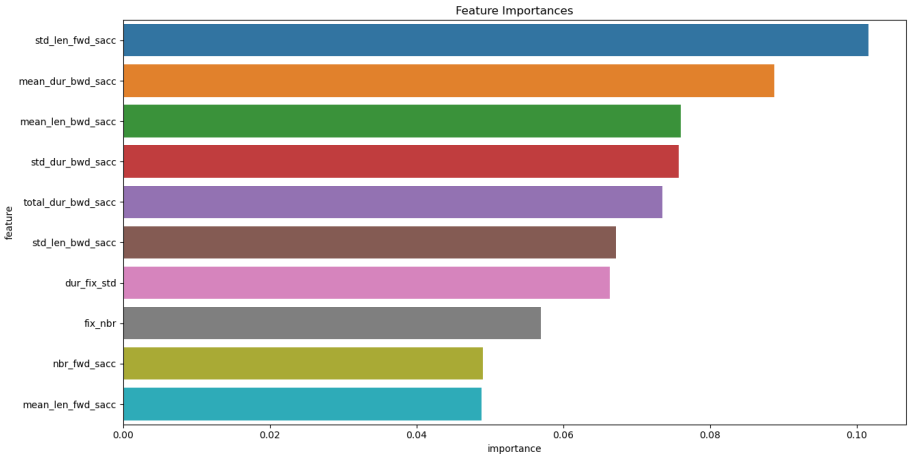

Figure 5.5: Important eye tracking features for L2 level prediction (Japanese speakers)

5.4 Results: selected features

Based on the model's prediction results, the analysis of the most important features and the analysis of the main statistical patterns with respect to speech and eye tracking characteristics (Chapter 4), in this section we select a subset of features with the aim to improve the model's performance for both French and Japanese speakers. For this purpose, we conduct several tests using different subsets of features described in Table 5.1 and as a result, select the features that contributed the most to the model's decision-making. In this section, we describe the feature selection process, compare the results with those obtained with the initial sets of features, and investigate the most important features for both French and Japanese speaker groups.

5.4.1 Speech features

Concerning audio features for French speakers, the following 6 features were selected: **speech rate, articulation rate, number of pronounced units (including disfluencies such as truncations, hesitations, repetitions), standard deviation of pauses duration, phonation duration, mean F2 of the vowel /u/**. The feature set also included the information about text level and page necessary for the prediction. Indeed, the analysis of the speech characteristics in Chapter 4 showed that more advanced French speakers speak faster, with fewer disfluencies, less frequent and shorter pauses, they also realize higher F2 of /u/ with L2 level improvement and they come closer in their pronunciation to that of native speakers.

The Table 5.4 shows the results (F-score) obtained using the updated audio feature set (including 6 features plus text level and page) compared to the initial one (including 26 audio features plus text level and page). The new results outperform those obtained with the initial feature set: 72% compared to 78% with selected features. The results also outperform those obtained with the baseline model calculated per vector (59%) and are similar to the

baseline results calculated per speaker (82%). The feature importance analysis (Figure 5.6) shows that the model relied mostly on articulation (25%) and speech rate (23%) for its decision, while the pronunciation related feature (F2 of /u/) contributed less to the model decision (7%).

Lang	Baseline		Initial Audio Features (+ Text Level, Gender)		Updated Audio Features (+ Text Level, Gender)	
	Per Vector	Per Speaker (Majority Vote)	Per Vector	Per Speaker (Majority Vote)	Per Vector	Per Speaker (Majority Vote)
FR	0.59 ± 0.17	0.82 ± 0.09	0.72 ± 0.05	0.69 ± 0.07	0.78 ± 0.04	0.82 ± 0.08
JPN	<0.5	0.72 ± 0.19	<0.5	<0.5	0.64 ± 0.23	0.67 ± 0.28

Table 5.4: L2 prediction results with updated speech features for French (6 features) and Japanese speakers (9 features)

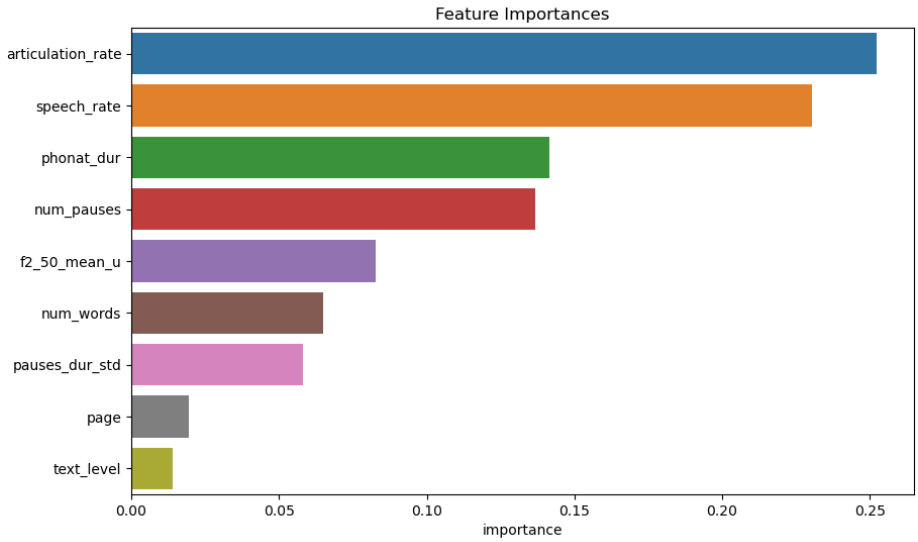

Figure 5.6: Important updated speech features for L2 level prediction (French speakers)

For Japanese speakers, the same feature selection logic was implemented. The updated feature set includes 9 features: **speech rate, articulation rate, number of pauses, standard deviation of pauses duration, mean intensity, pitch standard deviation (in Hz and semitone), standard deviation of phone duration /ʊ/, mean F1 of /ʊ/,** text level and page. The analysis of speech characteristics in Chapter 4 highlighted that the L2 level progress is, in general, less salient among Japanese speakers compared to French speakers, so the feature selection process was more challenging for them. For the pronunciation features we relied on the vowel space of Japanese and English speakers (Figure 2.2) and on the most important features in the previous model. However, while the F3 of /ʊ/ was the most significant factor of the previous model, it did not contribute to the model improvement, so was not included in the updated feature set. In the updated feature set, the highest importance was attributed to the standard deviation of the duration of /ʊ/, representing 32% of model decision [Figure 5.7]. The phone duration is followed by pauses duration variation and F1 of /ʊ/ representing about 15% of the decision-making. As for the articulation rate and

speech rate that were highly important for French speakers, it contributed less than 5% to the model decision for Japanese speakers.

The comparison of the initial results with the updated ones is presented in Table 5.4. The model's performance for Japanese speakers has improved: the F-score is now equal to 64% compared to <50% obtained with the initial feature set. It also outperforms the baseline results calculated per vector. However, the new results report high variation among folds, which means that the model is still not able to generalize well on unseen data and is highly dependent on the data in the train and test sets. This can be due to an insufficient amount of data of Japanese speakers, its high variance, and/or not sufficiently adapted speech metrics for Japanese data.

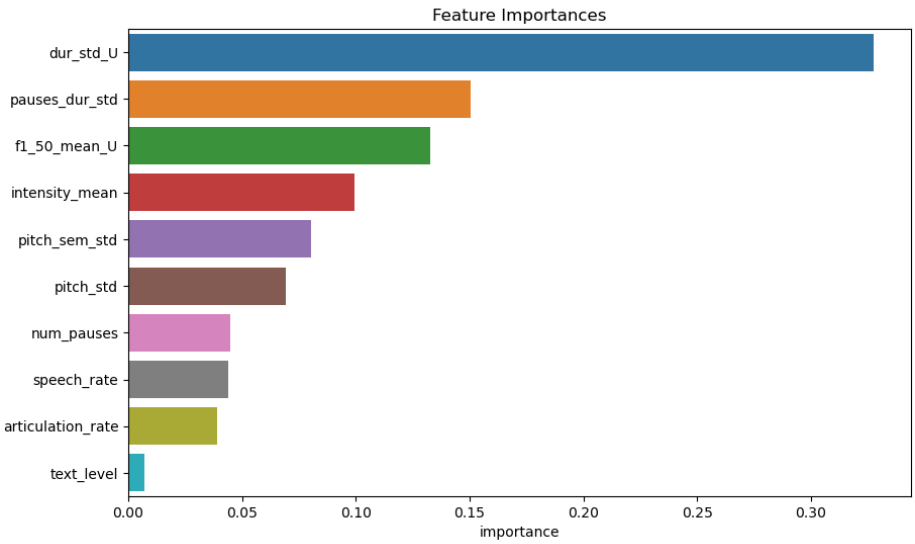

Figure 5.7: Important updated speech features for L2 level prediction (Japanese speakers)

5.4.2 Eye-tracking features

A similar feature selection procedure was implemented for the eye tracking features. For French speakers, the following 9 features showed the best performance across the folds: **mean and total fixation duration, standard deviation of fixation duration, mean and standard deviation of backward saccades (regressions) duration, mean and standard deviation of forward saccades duration, number and total duration of forward saccades**, plus text level and page. The reduction of the feature set allowed us to obtain better performance per speaker with the majority vote technique: F-score of 77% for the initial feature set compared to 81% for the reduced one (Figure 5.5). The results outperform the baseline model evaluated per vector. However the results evaluated per speaker are similar for both the baseline and the Random Forest. The most important features of our model are the total duration of saccades and fixations accounting respectively for 20% and 15% of the model's decision (Figure 5.8).

For Japanese speakers, another eye-tracking feature set appeared to be the most suitable and included the fol-

Lang	Baseline		Initial Eye Features (+ Text Level, Gender)		Updated Eye Features (+ Text Level, Gender)	
	Per Vector	Per Speaker (Majority Vote)	Per Vector	Per Speaker (Majority Vote)	Per Vector	Per Speaker (Majority Vote)
FR	0.59 ± 0.17	0.82 ± 0.09	0.75 ± 0.03	0.77 ± 0.02	0.75 ± 0.05	0.81 ± 0.06
JPN	<0.5	0.72 ± 0.19	0.61 ± 0.13	0.61 ± 0.07	0.62 ± 0.11	0.73 ± 0.16

Table 5.5: L2 prediction results with updated eye-tracking features for French (9 features) and Japanese speakers (8 features)

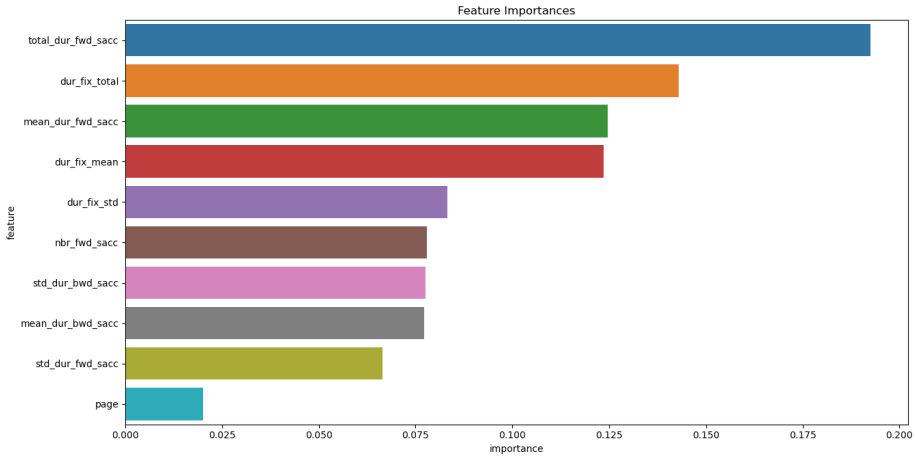

Figure 5.8: Important updated eye-tracking features for L2 level prediction (French speakers)

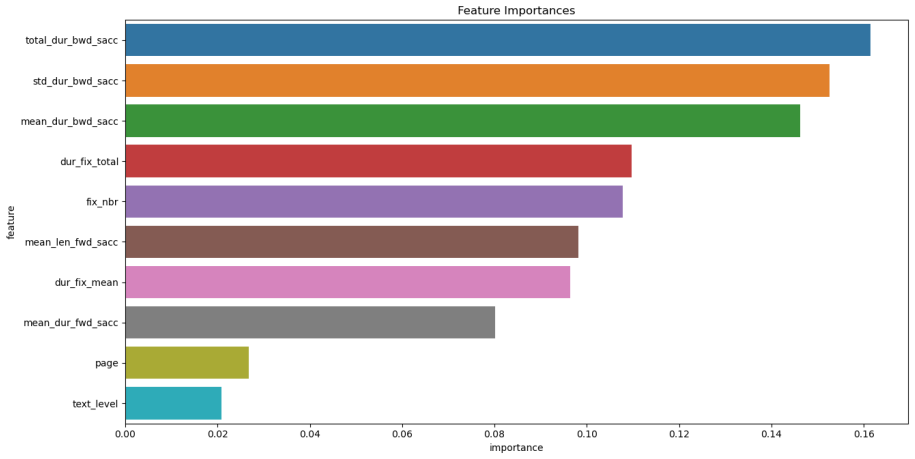

Figure 5.9: Important updated eye-tracking features for L2 level prediction (Japanese speakers)

lowing 8 features: **mean and total fixation duration, number of fixations, mean length of forward saccades, mean duration of forward saccades, mean and standard deviation of backward saccades (regressions) duration, total duration of backward saccades**, plus text level and page. The reduction of the eye-tracking feature set contributed to the improvement of per speaker performance with the majority vote technique: F-score of 61%

for initial set compared to 73% for the updated feature set (Figure 5.5). The model also outperformed the baseline evaluated both per vector (62% vs <50%) and per speaker (72% vs 73%). The most important features are those related to backward saccades and contribute to the model's decision about 15% each (Figure 5.9). Besides, the reduction of the feature set improved the results for Japanese speakers, there is still high variability among different folds, meaning that the model is not able to generalize well to unseen data and is dependent on the data in the train and test sets. The same trend was observed earlier with audio features, and this can be due to insufficient amount of Japanese speakers' data.

Based on the selected features and the analysis of their relevance, it appears that the backward saccades (or regressions) are more decisive factors for L2 level detection for Japanese speakers, while they are less important for French speakers, whose L2 level is better reflected by the use of fixations and forward saccades.

5.4.3 Speech and eye tracking features

In the previous section, we have observed the way in which an accurate selection and the reduction of the number of speech and eye tracking features improve the results of L2 level prediction for both Japanese and French speakers. We also examined and compared the most important features influencing the model decision for both types of speakers and discovered that the salient features were different for French and Japanese speakers. In this section, we present the results of the combination of the updated feature sets. Based on the results obtained in the previous section, we selected an equal number of speech and eye tracking features to combine them and to evaluate the model performance. Specifically, 4 most important features of each set were selected and resulted in vectors of size 8 for both French and Japanese speakers. Different combinations of the number of speech and eye tracking features were tested, but led to similar results. We suppose that the model decision is driven mostly by the most important features, and the augmentation of the number of features does not necessarily improve the results.

The Figure 5.6 presents speech and eye-tracking features used for French and Japanese speakers. These feature sets were also enriched by the information about the text level and the page number.

lang	speech features	eye features
French	articulation rate speech rate phonation duration num pauses	mean fixation duration total fixation duration mean duration forward saccades total duration forward saccades
Japanese	std duration /ʊ/ mean f1 /ʊ/ std pauses duration mean intensity	mean duration backward saccades std duration backward saccades total duration backward saccades total duration fixations

Table 5.6: Updated and reduced speech and eye-tracking feature sets for French and Japanese speakers

The Table 5.7 summarizes the results obtained using a combination of speech and eye tracking features before and after feature reduction for French and Japanese speakers. For French speakers, the results are improved for per

vector performance, those for per speaker performance remain similar with a slightly higher variability when using the updated feature set. However, the best results for the French speakers are obtained using the updated speech feature set alone: F-score of 78% per vector and 82% per speaker (Table 5.4).

As for the Japanese speakers, the combined and reduced feature set of speech and eye tracking features improved the model's performance compared to the results of the combined features without selection. The results improved from the performance lower than 50% to F-score of 67% per vector and 63% per speaker. However, we still observe a relatively high variation among the 5 folds. Among all the tested feature sets, the highest performance per speaker was obtained using the updated set of eye-tracking features alone and resulted in an F-score of 73% (Table5.5). For both speakers, the results obtained with the updated feature sets outperform the baseline model evaluated per vector, but not the performance evaluated per speaker. However, the baseline model fails to generalize to the minority class on the contrary to the Random Forest model (Table 5.8).

The feature importance analysis shows that both models for French and Japanese speakers rely more on the audio features than on eye-tracking features when combined in equal proportions. Specifically, the model for the French speakers relies most on the articulation rate that contributes more than 25% to the model's decision and the speech rate contributing about 16%. Eye tracking features contribute about 7% each (Figure 5.10). As for Japanese speakers, the model relies mostly on the standard deviation of the /ʊ/ duration (>35% of importance), while the other features contribute less than 15% to the model's decision-making (Figure 5.11).

Lang	Baseline		Initial Set Audio + Eye Features		Updated Set Audio + Eye Features	
	Per Vector	Per Speaker (Majority Vote)	Per Vector	Per Speaker (Majority Vote)	Per Vector	Per Speaker (Majority Vote)
FR	0.59 ± 0.17	0.82 ± 0.09	0.74 ± 0.03	0.77 ± 0.06	0.78 ± 0.02	0.77 ± 0.07
JPN	<0.5	0.72 ± 0.19	<0.5	<0.5	0.67 ± 0.30	0.63 ± 0.40

Table 5.7: L2 prediction results with updated combined speech and eye-tracking features (French and Japanese speakers)

To sum up, we selected an equal number of speech and eye-tracking features to analyze the model's performance and to account for the contribution of each kind of features to the model's decision-making. The results show that different features are required to assess the L2 level of French and Japanese speakers. Therefore, we need to take into account the L1 specificities when assessing L2 level, and possibly create different evaluation models depending on the speaker's L1. The results obtained on our dataset do confirm the Hypothesis 2. The Hypothesis 1 is however not confirmed. Although the results obtained with the combination of both speech and eye tracking features allows to predict speakers' L2 level, the best performance is obtained using the speech feature set for French speakers and eye-tracking feature set for Japanese speakers. We need to take into account that the dataset on which the models were applied is relatively small and required data reduction due to eye-tracker calibration issues. Therefore, the obtained results show the tendencies and allow us to formulate new hypothesis about the selection of the features which are the most representative of the L2 level and L1, as well as the choice of the appropriate model.

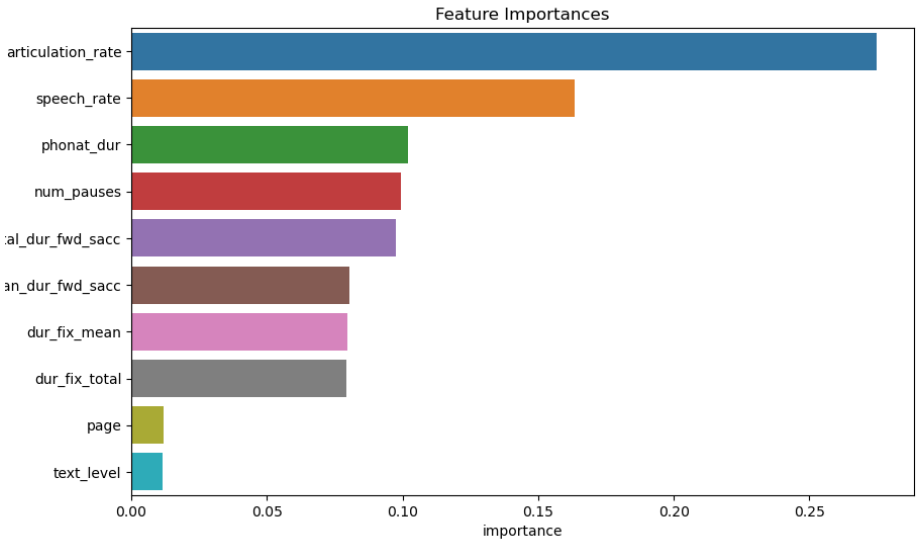

Figure 5.10: Important updated combined speech and eye-tracking features for L2 level prediction (French speakers)

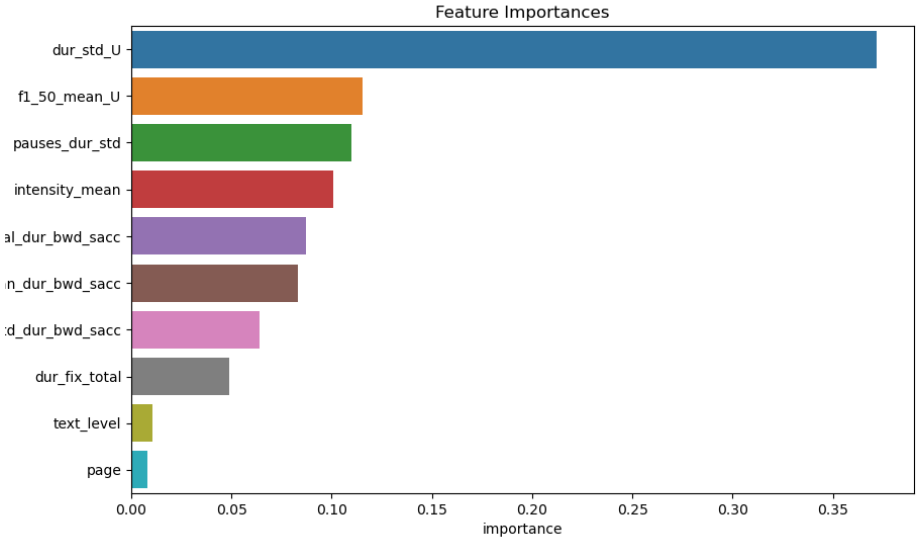

Figure 5.11: Important updated combined speech and eye-tracking features for L2 level prediction (Japanese speakers)

Given the presented performance, we aimed at investigating the model generalization for each class of speakers: beginner and advanced. The Table 5.8 represents the performance for the last model that combines the updated speech (4) and eye-tracking (4) features. It shows the F-score calculated per vector (per page) and compares the performance of the Random Forest model with the baseline. According to the Table, the class representing the beginner speakers is less well predicted compared to the advanced speakers for both French and Japanese participants. The class of the beginner speakers is also the minority class. Indeed, according to the Table 5.2 representing the number of speakers per L2 level, there are fewer beginner speakers among both French and Japanese participants. There-

fore, we hypothesize that the data balancing in the training set could improve the prediction results for the minority class. In order to verify this hypothesis, we conducted three tests where in the training set:

- the number of samples representing the advanced class is larger than that of the beginner class (the original setup)
- equal number of samples for advanced and beginner classes (undersample the advanced class to the beginner one)
- the number of samples representing the beginner class is larger than that of the advanced class (undersampling the advanced class)

These tests are performed on the French speakers only, because more data is available. As a baseline, we take the last model that combined 8 speech and eye-tracking features. The Table 5.9 illustrated the comparison of per class performance in the three setups proposed above. The results show that there is a gap between the performance on the beginner and advanced speakers classes in all the tests, where the beginner speakers are less well predicted. The performance for the beginner class remains unchanged in all the three setups. However, the performance for the class representing the advanced speakers lowers when less data representing the advanced class is available. Given these observations, we hypothesize that the poorer model's performance on the beginner speakers class can be due to two issues. The one is the small data size which is not sufficient to learn the model to generalize well for both classes. The second possible reason is the selected features that are not sufficiently representative of the beginner class and/or are better suited for the representation of the advanced speakers. It can also be the combination of both causes. In order to cope with this issue, several steps need to be undertaken in future work. For example, there is a need to collect more data for both French and Japanese, beginner and advanced speakers in order to obtain more robust results. Then, we consider providing a more in-deep feature engineering technique and choosing feature sets that are more representative of the spoken and eye movement behavior of the speakers of both classes. Another option to consider is to apply a data-augmentation technique for the minority class.

Lang	Class	Baseline	Audio + Eye (8 Best)
FR	Beginner	0.0	0.58 ± 0.1
	Advanced	0.81 ± 0.09	0.84 ± 0.4
JPN	Beginner	0.0	0.51 ± 0.4
	Advanced	0.52 ± 0.3	0.65 ± 0.3

Table 5.8: Per class performance for the feature set (speech + eye-tracking)

Lang	Class	Adv >Beg	Adv = Beg	Adv <Beg
FR	Beginner	0.58 ± 0.1	0.57 ± 0.1	0.59 ± 0.2
	Advanced	0.84 ± 0.4	0.79 ± 0.03	0.78 ± 0.2

Table 5.9: Per class performance with undersampling the majority class (French speakers)

features related to	LME		RF	
	French	Japanese	French	Japanese
speech rate	yes	yes	yes	yes
pitch	no	no	no	yes
intensity	no	no	no	yes
oral disfluencies	yes	no	yes	no
pauses	no	yes	yes	yes
vowel duration	yes	no	no	yes
vowel F1	-	-	no	yes
vowel F2	yes	no	yes	no
vowel F3	no	no	no	no
fixations	yes	no (overall) yes (unfrequent words)	yes	yes
forward saccades	-	-	yes	no
backward saccades	-	-	no	yes

Table 5.10: Features detected with Linear Mixed Effect model (LME) vs features detected with Random Forest classifier (RF)

5.5 Summary

In conclusion, in this chapter we investigated automatic L2 level proficiency prediction for French and Japanese speakers using Random Forest classifier. We selected a variety of speech and eye-tracking factors correlated with the L2 proficiency according to the literature discussed in Chapter 2 and to the results obtained in Chapter 4. We applied feature engineering techniques in order to select the most significant factors in order to reduce the feature set and to improve the model's performance. The results show that speech and eye tracking features are efficient for L2 prediction for French speakers: the model achieves an F-score superior to 70% for all feature sets. For Japanese speakers, the model achieves an F-score of about 60%, however a higher variance is present due to a small size of the database.

The combination of the speech with eye tracking features did not significantly improve the results for French or Japanese speakers, however both types of features contributed to the model's decision. The best results were achieved using speech feature set for French speakers and eye-tracking feature set for Japanese speakers.

The Table 5.10 presents an overview of features that were tested in Chapter 4 using LME models and those that showed their importance in the prediction model of L2 level. This table does not illustrate a detailed set of tested features, instead it represents their categories. For example, instead of including all the features related to pauses, such as their number, total duration, its average and standard deviation, we combine them in one category "features related to pauses". The Table shows whether the features related to the given category were correlated with L2 level when tested with LME, and whether they showed their efficiency in a Random Forest (RF) model to L2 prediction when combined with other features. For this purpose, the following color codes are used: green, if the feature effectiveness corresponds to our expectations based on LME, and red, otherwise. The categories of features that were not tested using LME, such as those related to forward and backward saccades are marked with "-" and should be

included in future work. The features that we expected to contribute to the model decision and were actually effective include: speech rate (for both group of speakers), number of disfluencies (for French speakers), F2 of the vowel /u/ (for French speakers), pauses (for Japanese speakers), and fixation duration (for both group of speakers). On the contrary to our expectations, features related to pitch and intensity were efficient for Japanese speakers. Features related to vowel duration were expected to be efficient for French speakers, but not for Japanese ones, however the results are inverse to our expectations. The results presented in this chapter show that the models used for French and Japanese speakers rely on different features depending on the L1, which highlights an importance of taking into account speaker's mother tongue when assessing the L2 level. Indeed, speakers of different mother tongues have varying difficulties when learning a foreign language which may depend, beyond the speaker dependent features and cultural learning habits, on the linguistic proximity between L1 and L2, both genetic ans typological. French and English are closer genetically, as both are Indo-european languages whereas Japanese is considered a language isolate and is not related to the branch that lead to French and English. French and English are also typologically, and what is more, geographically close: they share a common European history that translates in borrowings at all linguistic levels. We hypothesize that an in-depth analysis of the impact of the linguistic patterns with respect to the relation between L1 and L2 could help selecting the most representative features in order to assess L2 skills of speakers having different mother tongues.

The results presented in the chapter provide preliminary estimation of features that can be taken into account when assessing L2 level and need to be tested on a larger and more balanced dataset.

Chapter 6

Towards using culturally adapted nudging strategies in L2 learning: example of manga

As we pointed out in the introduction, nudges are subtle influential strategies allowing to modify people's behavior (Thaler and Sunstein, 2008). The term comes from economic theory and is increasingly used in education, including L2 learning. In Chapter 2, we provide a literature review concerning the use of nudging technologies in education.

In this thesis, we relied on the speech and eye tracking data collected from representatives of different mother tongues and cultures, French and Japanese speakers, in order to decipher cultural differences in terms of L2 pronunciation and reading challenges. The results showed that speakers of different cultures have different pronunciation challenges and implement different strategies to cope with difficult words from the point of view of both pronunciation and comprehension. Based on these observations, we aimed at investigating the possibility of implementation of culturally adapted nudging strategies which could improve L2 acquisition, including text comprehension and pronunciation. For this purpose, we implemented a preliminary research protocol to test this hypothesis: in this purpose we employed manga as a tool to implement nudging strategies.

Manga are Japanese comic books which gained their popularity worldwide, including France. In Japan, manga is not only a popular leisure mode, but also an educational tool. There exists a variety of manga designed for learning various subjects, from math to foreign languages. They are known for improving the comprehension of a particular subject and increasing learning motivation due to their captivating stories and a combination of textual and visual cues. Moreover, they have been already used by the Japanese partners of the LeCycl project in previous studies to evaluate comprehension through eye tracking measures (Rigaud et al., 2015; Augereau et al., 2017), as well as for unknown word estimation (Takaike et al., 2023).

Japanese comic books are characterized by dialogue structures associated with images, containing mostly short phrases and requiring a specific reading direction, which can be challenging for inexperienced manga readers: from left to right, from top to bottom. As a literary medium, they are close to Western culture comics, however the reading strategy is different and may require an adaptation effort for non Japanese readers. In addition, cultural references, such as those related to traditions, Japanese society, gastronomy, and language reflecting the relationships between people, differ from Western culture. These and other characteristics increase the reading complexity for inexperienced and/or non-Japanese readers.

While aware of the challenges posed by this medium, we nonetheless adopted manga as a preliminary tool for testing nudges for the reasons mentioned: there have already been comparable experiments carried out by our colleagues from the University of Osaka, both the French and Japanese cultures are familiar with the medium, albeit to different degrees, and it is a medium that lends to multimodality, as text and images contribute to understanding the content conveyed by the book.

The experiment conducted for this purpose is a pilot study that requires improvements and further investigation. It opens the potential for using manga as nudges in L2 acquisition.

Our goal thus was to investigate:

1. whether manga improves oral production by enhancing text comprehension among Japanese speakers

2. whether we can easily transfer this cultural education tool to another culture, less familiar with it as learning medium: from Japanese to French culture

In this Chapter, we will present the research protocol implemented among Japanese and French speakers of English using manga as a reading support. Then, we will show preliminary results obtained on this data, and, finally, we will discuss the limitations of the experiment and new research directions that the experiment opens.

6.1 Data acquisition and processing

We recall that prior to the LeCycl project, OMU partners have been using manga for a variety of reasons, not necessarily related to learning, for example: text-independent speech balloon segmentation (Rigaud et al., 2015), improving comic visualization of smartphones (Augereau et al., 2016b), emotion estimation during comic books reading (Matsubara et al., 2016). For the LeCycl project, where the focus is on learning, they implemented a research protocol aiming at unknown words estimation during reading aloud comic books in English as a foreign language (Takaike et al., 2023). In line with this experiment, as well as the experiments described in this thesis, we built a new research protocol described below.

Participants: Two groups of participants were recorded for this study:

(1) 24 Japanese speakers, who participated in the previous experiment described in this thesis, involving reading aloud texts of different levels. The experiment resulted in 36h of collected speech (see Chapter 3 Table 3.1).

(2) 14 French speakers, 7 of whom participated in the previous study, and 6 of them have little or no experience in reading manga. The experiment resulted in 10,5h of recorded speech (see Chapter 3 Table 3.1).

Manga choice: The manga for the experiment were empirically chosen, by estimating the number of unknown words for Japanese and French speakers. Therefore, the representatives of two cultures were asked to read different manga during the experiment. The manga texts are also longer compared to those used in the previous experiments described in the thesis. The manga proposed in this experiment are quite popular among Japanese readers and included:

- "Barrage" by Kohei Horikoshi, a story of an orphan saving the planet Industria from alien ennemies
- "School Judgement" by Takeshi Obata, a story about elementary school students settling their disputes in a courtroom-style
- "Delicious in Dungeon" by Ryoko Kui, a story of a group of friends exploring a dungeon and cooking dungeon monsters to survive

Task: Japanese speakers read an extract of manga "Delicious in Dungeon" (including 973 words see Table 6.1), both in its original (with images) and textual form. Its reading took about 10-15 min. The other manga were read entirely in their original form, and the experiment took 2 hours divided into 1h sessions. French speakers read manga only in their original form. The experiment for French speakers was shortened and lasted about 1h-1h30. The task was similar to that of the previous experiments discussed in this thesis: to read aloud and to highlight unknown words. No comprehension questions were asked after manga reading. As in the previous experiments, the recorded data included speech and eye movement.

Equipment: Due to time and location constraints, slightly different recording equipment was used for the data collection. For the recording of the Japanese speakers reading an extract of manga in its textual form and of the French speakers, the microphone was the same, as in the previously described experiments including text reading of different levels. As for the experiment involving Japanese speakers reading manga in its original form, a different microphone was used. In addition, different versions of Tobii eye tracker were used for both experiments, as well as the type of screen: portable or fixed. These differences in the used equipment challenged the data processing procedure and the results comparison.

Data processing:

The processing of manga differs from text processing because it includes both visual and textual information.

For example, the word counting in the manga is challenging and requires manual intervention. For instance, our OMU colleagues relied on Google Vision to retrieve text from manga. While this approach was sufficient for

Table 6.1: Number of words in manga read by French and Japanese speakers

speakers' L1	manga ID	total nbr words	total nbr of unique lemmas	nbr of rare unique lemmas (CELEX)	nbr of unique japanese words
French	mng_1	1671	499	35	0
	mng_2	2177	741	78	14
	mng_3	1647	565	52	0
Japanese	mng_img mng_txt	973	367	21	0

their research goals which only concerned the eye tracking modality, its output was not precise enough for ours where speech is a relevant part requiring appropriate processing analysis. We thus considered a different option, that is to use an ASR tool on the most proficient participant's speech, to use the result as the baseline, and then, manually correct the texts for the other speakers. The results of word calculation are presented in Tables 6.1. We also calculated the number of infrequent words using CELEX corpora (Baayen et al., 1995), as in Chapter 4, in order to account for text's lexical complexity.

Another data processing challenge is related to the use of different equipment as described in the previous section. For instance, one can hypothesize that the different microphones will impact at least some speech parameters. To anticipate this issue, such speech characteristics as pitch and intensity were not analyzed as they may be dependent on the type of the equipment. Instead, we calculated speech features which we assume to be independent of the recording conditions, which included speech rate, articulation rate, phonation time, and percentage of silence.

As discussed in previous chapters, these features were calculated based on the manual correction of texts and word boundaries in order to obtain reliable results. Indeed, this processing technique was necessary for the precision required for our research goals. However, it is a time consuming task and therefore unrealistic from the perspective of a real-time application. Therefore, because of time and resource constraints, only a part of manga data (those read by French speakers) was manually enriched in disfluencies and aligned. This work was completed by our Japanese partners: Motoi Iwata and his student Hayato Seki. It allowed us to calculate the number of disfluencies and to compare their percentage used when reading texts and manga.

As for the other features (phonation time, articulation rate, silence percentage), they were automatically calculated using a Praat script designed for this purpose (de Jong and Wempe, 2009), in order to reduce processing time and to obtain preliminary results. The script estimates the number of syllables, and thus, the speech rate by relying on intensity peaks. The script was applied to all the data in order to make the results comparable.

6.2 Preliminary results

The features mentioned earlier were compared in two setups:

(1) among Japanese speakers reading an extract of manga in its original form including images and in its textual

form. This setup allows to account for the role of image in reading aloud production.

(2) among French speakers reading three different manga in their original form in comparison with the reading production of three texts (beginner, intermediate, and advanced) from the first protocol. As mentioned in the previous section, French speakers read manga only in the original form, and the manga read by Japanese speakers differs from those read by French speakers. Therefore, in order to get insights into the overall reading performance of the French speakers when reading manga, we compared it to the results obtained for text reading from the previous experiments. We are aware that this comparison is not reliable, but it allows us to get a general idea about speech rate and pausing time when reading text with or without images.

To sum up, in this section, the results are provided for the sake of comparison, to get a general idea of Japanese and French speakers' performance when reading aloud plain texts or manga. The dataset was acquired empirically, and the data is not directly comparable. Therefore, no statistical modeling was implemented. Being aware of the study's limitations, we interpret the results with caution, formulating hypotheses and proposing an improvement of the research protocol for future work.

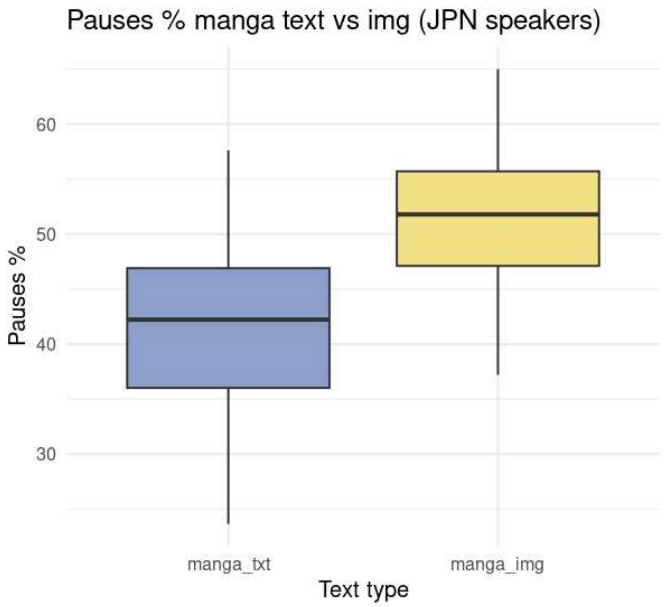

Figure 6.1: Percentages of pauses of Japanese speakers reading manga in original vs textual form

Japanese speakers

The Figures 6.1 show that the percentage of pauses realized by Japanese speakers when reading manga in its original form is higher than that when reading manga in its textual form. In addition, for the original manga, we observe longer phonation time (Figure 6.2), which can be due to the presence of oral disfluencies (truncations, hesitations, and repetitions), and lower articulation rate (Figure 6.3).

We hypothesize that these results can be due to several factors that require further exploration, such as:

- the need to turn the pages more often in manga, as less text is fit into the pages than in basic texts

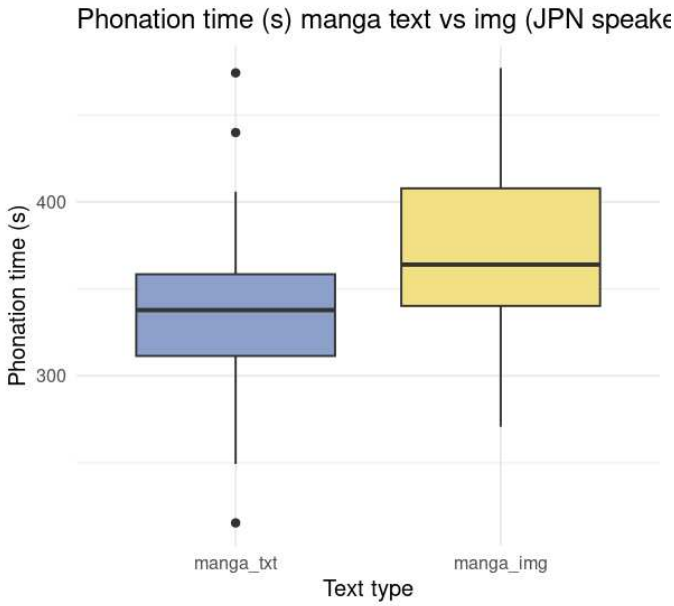

Figure 6.2: Phonation duration (s) of Japanese speakers reading manga in original vs textual form

- longer time needed to navigate between speech balloons compared to text lines
- the images can distract the attention from the text and driven to the images, and therefore increase pauses duration
- the combination of both visual and textual cues elicit higher cognitive load

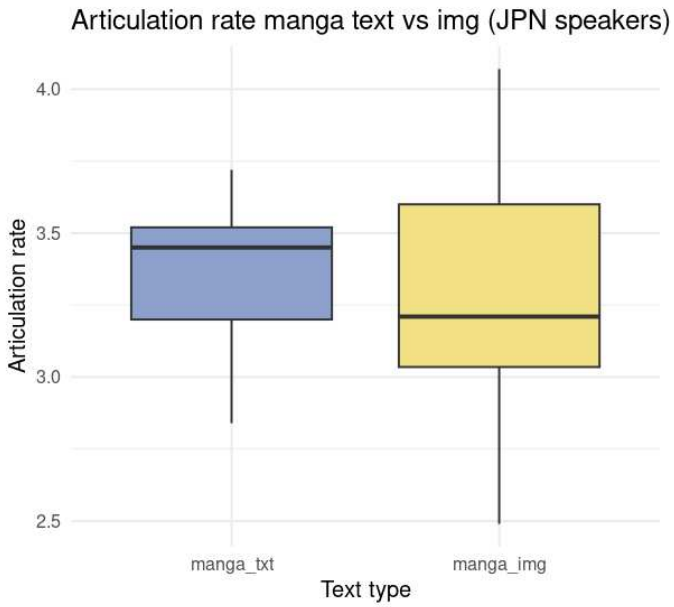

Figure 6.3: Articulation rate of Japanese speakers reading manga in original vs textual form

French speakers

For the manga read by French speakers, the number of oral disfluencies (truncations, hesitations and repetitions) were manually calculated and their percentage has been compared to those used in the texts from the first experiment. The average percentage of disfluencies used when reading manga remains in the limits of those used in the texts (Figure 6.4). In the texts, it increased with text complexity, and in manga its increase can be due to the number of infrequent words, and/or to the manga length. For example, the highest mean disfluenciy rate is used in the manga labeled as "mng_2", which is characterized by higher number of infrequent words compared to the other manga, as well as words relative to Japanese culture. However, it is similar to the disfluency rate observed when reading the advanced text.

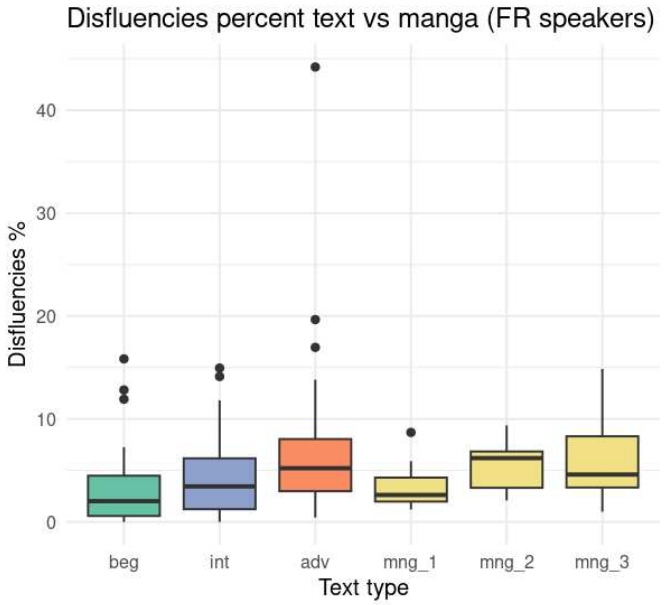

Figure 6.4: Percentage of disfluencies used by French speakers reading manga vs texts

The pause rate produced in manga also remains within the boundaries of those observed during text reading (Figure 6.5). We hypothesize that an increased pause rate in the first manga can be due to the familiarization phase of the experiment and the need to adjust to the experimental setup. As for the final manga "mng_3", its higher pause rate can be related to fatigue due to the experimental duration. Indeed, the experiment required about an hour of oral reading, which is unusual for French participants who got used to taking part in shorter experiments. The experimental setup, including its duration, is another parameter where we can observe cultural differences.

As for the articulation rate, its values are close or lower than those of the advanced text (Figure 6.6). It can be due to a new experimental setup, to a lack of experience of reading manga (among about an half of the French participants), to the attention distraction due to the presence of visual cues, and other factors that need further investigation.

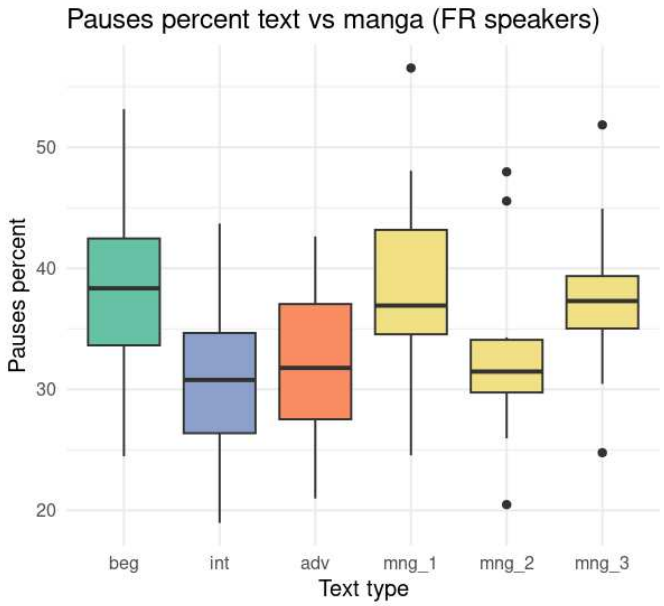

Figure 6.5: Percentages of pauses used by French speakers reading manga vs texts

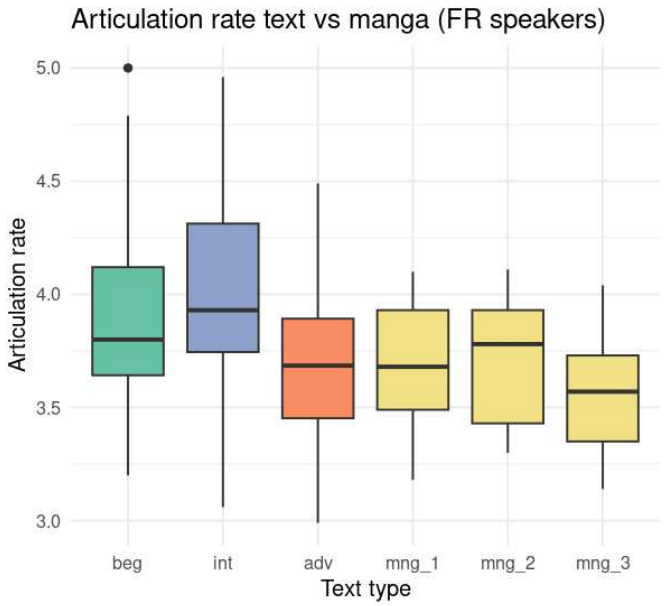

Figure 6.6: Articulation rate used by French speakers reading manga vs texts

6.3 Limitations and future work

In conclusion, we conducted a pilot study in order to investigate the potential of using manga in L2 learning context and its transfer to other cultures. Here, manga are implemented as nudging strategies, that is as medium potentially helping decoding the text through associated images, aiming at enhancing oral production through increasing motivation and text comprehension.

The preliminary results show no or little improvement in oral reading among Japanese speakers who are nev-

ertheless acquainted with reading manga, neither among French speakers, some of whom are less familiar with manga in general and with the topics proposed in our experiment in order to be able to compare with Japanese as L1 productions. For both speaker categories, we observe lower speech rate and more pausing time in manga compared to texts. The other observed differences features require more investigation.

Our study shows the challenges to conducting this kind of experiment in an intercultural and international context, as well as multimodal data processing challenges. In particular, several points need to be taken into account when designing the next experiment:

- select shorter manga extract to be read in their original and textual form by representatives of different cultures. Shorter texts are more suitable for our experimental goals, as they provoque less fatigue in the speakers, and the obtained data required less processing time, especially in case where manual processing is needed
- use the same recording equipment for all the experimental setups
- verify content comprehension to account for the role of images in comprehension

Chapter 7

Conclusions and perspectives

7.1 Conclusions

In this document we proposed an **interdisciplinary** and **cross-cultural** take of challenges in assessing performances in learning a new language in a connected and rapidly changing world where cross-cultural factors are increasingly present in all areas of daily life, including education. The thesis has been conducted in the framework of the trilateral project The Learning Cyclotron which put forward the notion of an ecosystem of knowledge circulation. Among the three functions identified as central to the project, that is Perceive, Master and Transfer, we mainly considered the perception aspects, that is the assessment of L2 learning level in our case, but also dimensions of the transfer as we questioned the easiness of transferring learning habits from one culture to another.

In line with LeCycl project objectives, we developed the **interdisciplinary** approach by considering challenges in assessing L2 English learning with cues from two modalities: we thus have focused on speech and eye tracking, two modalities mastered by the two teams out of three involved in the project and specifically in the supervision of my research, that is LISN CNRS (speech, Ioana Vasilescu and Laurence Devillers) and Osaka Metropolitan University (eye tracking, Koichi Kise, Motoi Iwata and Olivier Augereau).
As for the **cross-cultural** dimension, we considered such challenges through the specificity brought by two cultures, French and Japanese, as well through the cross-cultural preliminary experiment during which the specific Japanese learning medium, manga, was used during a reading aloud experiment with French subjects.

More specifically in this thesis we have proposed a multimodal protocol during which L1 speakers of French and Japanese completed several reading tasks that allow to assess their level in L2 English. We collected speech and eye tracking data and we compared them with native speakers of English. We also performed a cross-comparison to identify the most efficient speech and eye movement cues that correlate with French vs Japanese specificity, or are

language independent. Our research questions concerned the impact of multimodality, the relation between efficient cues and cross-cultural features, the machine learning techniques as tools to predict L2 specificity and finally, the interest of using cross-culturally manga as learning medium allowing nudging implementation for education. In line with the research questions and the review of the literature, we have built hypotheses that underline that all considered factors (modality, culture, nudging) may have different weight according to the experimental set up. In order to circumscribe such weight and pave the way for more in depth studies we collected corresponding data and took advantage of our partners experience and data. In particular, part of our analyses are conducted on the data and corresponding annotations, provided by Professors Koichi Kise and Motoi Iwata from Osaka Metropolitan University. We then semi-automatically processed the data and we conducted linguistics and eye tracking analyses in order to highlight features that we modeled by using Random Forest algorithms in order to predict L2 challenges as function of the modality and cultural patterns.

Our contributions are as follows:

1. **Concerning the spoken modality:**

   We provided an analysis of speech features extracted at different levels including: text level (prosodic features and disfluencies), word level (prosody and disfluencies) and phone level (formant realization) by using visualization and LME modeling. The LME models highlighted the difficulty of assessing L2 level of speakers having different L1s: the difference between beginner and advanced Japanese speakers is less salients than the difference between beginner and advanced French speakers. Indeed, speakers can show different progress with L2 level improvement depending on their L1 and the linguistic distance (in a broad sense genetic and typological) with the target language. We hypothesize that Japanese speakers have different challenges in mastering English L2 compared to French speakers with respect to their mother tongue. Such challenges may translate in different salient features in modeling the L2 acquisition and suggest that L2 assessment may be L1 dependent.

   After the features were analyzed, they were incorporated into a L2 prediction model separately for French and Japanese speakers. The results show that features related to prosody and pronunciation specificities contribute to the model decision-making. However, the most important features differ depending on the speaker's L1: for French speakers, the model relied mostly on articulation rate, speech rate, number of pauses and vowel's /u/ F2; for Japanese speakers, the most important features included those related to /u/ duration and F1, pauses duration, intensity and pitch, while speech rate was less important.

2. **Concerning the eye tracking modality:**

   A similar approach was implemented as for speech features. Eye tracking features were extracted and analyzed at text and word level. Then, they were incorporated into the L2 prediction model and contributed to the model

performance for both French and Japanese speakers. Interestingly, while fixation duration contributed to the models for both French and Japanese speakers, features related to forward saccades were more representative of French speakers' L2 level, and features related to backward saccades, or regressions, were more salient predictors of the Japanese speakers' L2 level. These results are obtained using the features importance analysis after the model was fit.

Similar to speech features, the LME showed that the distinction between beginner and advanced Japanese speakers is less salient than for French speakers.

3. **Concerning the combined features:**

   The combination of eye tracking with speech features in equal proportion allowed to predict L2 level for both French and Japanese speakers, although better performance was obtained on French speakers, for whom more data was available. The feature importance analysis shows that when combining an equal number of speech and eye-tracking features, the model relies mostly on speech features, however both feature categories contribute to the model decision. The best performance was obtained, however, using solely speech features for French speakers, and only eye-tracking features for Japanese speakers. The model and the feature selection can be improved with future work discussed below.

4. **Concerning linguistic analysis vs prediction:**

   On the contrary to our expectations, not all the features that showed their correlation with L2 level in LME models were actually efficient in RF modeling. Particularly, features related to pitch and intensity contributed to the L2 level prediction for Japanese speakers while no significant correlation was observed in LME models. The LME modeling showed that vowel duration was correlated with French speakers' L2 level, but not with that of Japanese speakers, however, the RF modeling showed inverse results.

5. **Concerning the influence of L1 on linguistic and eye-tracking patterns:**

   The results summarized above suggest the necessity to take into account speaker's L1 when assessing L2 proficiency, because speaker's performance can be reflected at different degrees depending on the difficulty of the acquisition of the target language due to the distance between L1 and L2. Indeed, the study of Chiswick and Miller (2005) that calculated the linguistic distance from a variety of L1s to L2 English highlights varying challenges in acquisition of L2 English depending on the linguistic distance between L1 and L2 English. The linguistic distance can be one of the reasons for the differences observed in the spoken and eye movement behavior of French and Japanese speakers in the reading aloud experiment despite the similar reported L2 level.

6. **Concerning the nudges:**

   We conducted a preliminary experiment aiming at accounting for the role of images in oral reading production and at evaluating the possibility of transferring culturally adapted nudging strategies (manga) to another

culture. The results showed no improvement from the point of view of speech production: both Japanese speakers, who are acquainted with manga reading, and French speakers, some of whom are less familiar with manga, used longer pauses and slower speech rate when reading manga compared to plain text. We hypothesize that manga reading can be more challenging for both cultures, as it combines both visual and textual cues to be processed. For French speakers, who are less acquainted with manga, the task represents more difficulties due to the reading direction and the topics that they cover.

Although, no improvement was observed for both cultures, this experiment allowed us to question the implemented experimental setup and to propose improvement directions discussed below.

7.2 Future work

With the above results in mind, future work should focus on the following directions:

1. **Data:** Data collection and processing are the foundation of such interdisciplinary work and we underlined the difficulty in identifying relevant and shared database in the scientific community. Collect more data, especially for Japanese and beginner speakers in order to test our results on a larger and more balanced dataset will be a future necessary step.

2. **Features:** Extract and evaluate other features in order to get a deeper insight into reading aloud challenges of the representatives of both L1s. For example, for speech features, we could include the analysis of other phones, although it is time-consuming and is difficult to implement in real life applications. For eye-tracking, we could add information about the blinks, pupil dilation, saccade speed, which can provide information about cognitive load during task completion.

3. **Automatic vs human assessement:** For a more robust automatic assessment, we should consider adding human assessment provided by L2 professionals in order to correlate it with the participants' self-reported L2 level.

4. **Refining L2 level definition:** It is important to distinguish overall L2 level from L2 fluency level and from pronunciation proficiency. This issue can also be addressed by providing human scores assessing oral fluency and pronunciation.

5. **What makes a word difficult:** For the analysis of difficult words, we propose to take into account the difference between challenges of word processing related to comprehension and pronunciation or both, as well as to consider words that were previously unknown by the speakers, but understood in context while reading.

6. **Machine learning modeling:** As for the L2 automatic assessment, we consider testing other models that may provide more prediction performance, although less explicable. For example, we consider implementing Long

Short-Term Memory (LSTM) neural network, which captures temporal dependencies and can serve to modelize spoken and eye movement behavior.

Other models can be considered, for example, Convolutional Neural Networks (CNN) which takes as an input a spectrogram or a visual representation of eye movement patterns.

Finally, to sum up all our proposals for future work, this thesis has highlighted the complexity, challenges and rewarding aspects of an interdisciplinary research, and also shown that it opens the way to numerous research paths that may be of interest to several disciplines. Future work may enrich **digital sciences** and in particular machine learning approaches when one considers heterogeneous and multimodal data. It may also enrich **linguistics** by providing a new take on language acquisition assessed with multimodal cues in a cross-cultural perspective, as well more largely **humanities and social sciences** if we focus on the latter dimension. Last but not least, **education and ethics** can take advantage of the work on research protocols built to assess modern language learning and the experimental protocols designed to get insight into learning challenges while respecting ethical aspects.

Appendix A

Texts for reading aloud experiments

A.1 Beginner text

It is Thursday. It is raining today. It is a rainy day. Anna is inside the house. Anna is watching TV. Anna is watching TV inside the house. Anna cannot go outside. It is raining outside. Anna cannot go outside because it is raining outside. Anna is bored. Wait! Anna hears someone at the door. Someone is at the door of her house. Anna opens the door. What does Anna see? Anna sees a dog. The dog is small. Anna sees a small dog. The dog is wet. The dog is wet from the rain. Awww! You are all wet! Anna says to the dog. You are very cute! Mom! Anna says. Yes, dear? Says Anna's mom. There is a dog here! There is a dog at the door! Anna says. What? Says Anna's mom. A dog? Anna's mom comes to the door. Anna's mom sees the wet dog. The wet dog looks cute. Can we keep it? Asks Anna. Yes, we can. Says Anna's mom. The dog is happy.

A.2 Intermediate text

Ben jumped from bed as soon as the first bit of sun peeped through his window. He grabbed his backpack from his closet and opened it on the floor. Ben put in some of his favorite things to take with him on his trip to visit Grandpa. He put in a book on building forts, a book on making go-carts, and a new book he had gotten from the library about a kid detective who creates his own spy gear. He also put in a model car kit and his stuffed bear. He was ready to go! Going to visit his grandfather for a week by himself was always Ben's favorite part of summer vacation. Grandpa would take him fishing and to baseball games. Grandpa also taught Ben how to fix things around the house. Last year, when he was eight years old, Ben had learned how to replace a broken doorknob and how to fix a leaky faucet. Grandpa was patient and did not mind taking many hours to show Ben how to use his tools. Ben's mom stuck her head in his bedroom door. Grandpa's here. She said with a smile. Ben grabbed his backpack and ran into the kitchen where Grandpa was waiting. Ready, big guy? Asked Grandpa. Or do you want to eat breakfast before we

leave? Ready. Said Ben. As he kissed his mother goodbye, he felt his stomach rumble. We can eat later!

A.3 Advanced text

My cell phone rings again. It is futile to ignore it anymore. Valerie is persistent. When Valerie wants something, she will continue to bedevil me until I acquiesce. Hello. I answer. State Fair, Bobbie? She asks in her singsong voice. When are we heading out? Only two more days left! I abhor the State Fair. The boisterous crowds, the insanely long lines and the impossibility of finding a clean restroom all combine to make this an event that I dread. For Valerie, my best friend since the angst of middle school, the State Fair is a sign that divine powers really do exist. Really, Bobbie, where else can you pet a cow, ride a horse, fall ten stories, see the world's smallest person and eat fried macaroni and cheese? Valerie asks gleefully. Hell? I guess. The fried food at the State Fair is a gastronomical nightmare on its own. I once tried a fried pickle at the fair and was sick to my stomach for hours. And a fried donut hamburger with bacon, cheese and a greasy egg? How could that not be deleterious? I have not seen Valerie for a good month. Our schedules are both so hectic. My hatred of the State Fair becomes inconsequential compared to my desire to hang with Val. Alas, I ignore my anti-fair bias for the umpteenth year. Pick me up at noon. I say and hang up the phone.

Appendix B

Manga extract in textual form

- Scorpion hot pot, hm? Can't say I agree with your method though. - Who are you!? - When you eat huge scorpions, always take off the pincers, head, legs, and tail. The tail will give you the runs. - It will, huh? The book said it would be fine. - Well, technically, yes. It just tastes nasty. - You should cut scores along the body. The heat travels better, broth gets out, and the whole dish is tastier. Easier to eat too. And take out the innards as well. They're bitter, and their texture isn't good. If you stew 'em and ferment 'em, they're good beer snacks, but that's a bit much for beginners. - With walking mushrooms, throw away the rump and three centimeters of the surface. The legs are tasty, so use all of 'em. - The legs are good? - That's right. - I almost threw them away. - They have a nice, unique aroma. See? - "Foot smell," huh? - This pot's too small. Let's use mine. - You're freakishly well prepared. - Just scorpion and mushrooms will be a bit bland. - Hm… - Put these in. Wait! Not those!! - Marcille. - No! No! I can't, i really can't! Listen! This is a graveyard! Let's say monsters are okay. But plants that put down roots are out! For religious reasons. - It's fine this way! It looks good enough with just scorpion and walking mushroom! - Let's have it like this, okay? - But, Marcille… - Who are you anyway!? What on earth is.. - Marcille! Above you! - Above me? Urk! Slime! - Don't move, Marcille! - Rats. Right over my face. Fire spell oh… That won't work. I can't chant. Ugk! Hey, a life review. Come to think of it, the first thing that killed me was a slime. - Marcille, are you okay? - I'm fine. - I got some up my nose though. Thaah! Hff! - Here. Blow. - You beat a slime with a knife? - It's easy if you know their makeup. They look amorphous, but actually, they're more uniform than your average human. Slimes' internal organs look like this. In human terms, it would be like turning your stomach inside out and surrounding your head and organs with gastric fluids. - They sense when their prey's exhaled and leap at them. That means yelling and shouting makes you an easier target. - It's really not edible as it is now, but wash it well in hot water and a little citrus juice. Either wipe the moisture off or knead salt into, then dry it fully in the sun, and you've got a real delicacy. However, if possible, it's better to starve it for about two weeks. Takes time to dry it. It's not really a food you can just snack on in the dungeon. - This is a portable slimedrying net i made. It dries while you're walking around with it. It takes time for it to be ready, but I have some finished stuff right here. Let's put this in today. - But it's a delicacy, isn't it? - I don't mind. Your interest

makes me happier than anything else. I've researched monster cooking in this dungeon for over a decade. - Ten years! - All right, sit tight a minute. It'll be ready soon. - So huge scorpion turns red when you boil it? - I wonder if it's really a scorpion… - Smells pretty good. - Seeing it is a lot different from reading about it. - Heating the flesh shrinks it down a bit. It'll come out of the shell real easy. - You're right. - It's good! - Ain't it though! - The flavor changes so much depending on how it's cooked. - It sure does, doesn't it? - Give me a bowl too! What's this? - Dried slime innards. - Wow! It's delicious! - So this is how you eat slimes… - They're tasty any way you eat 'em. They're good if you soak 'em in fruit juice too. - These tree roots are nice and starchy. - They're not really roots. They're the stems of dungeon plants that grow upside down. - This algae is soft and yummy too. Is it another dungeon plant? - That stuff sprouts anywhere moist. It's plain old algae. - We're in the dungeon all the time, but I had no idea any of this was here. - Whew. - Man, I'm stuffed… - By the way...we skipped our introductions. - My name's Senshi. Means "seeker" in dwarfish. - I'm Laios. The magic user is Marcille. And that's Chilchuck, our picklock. - You seem to be journeying for a reason… - A monster ate one of our companions in the depths of the dungeon. We want to save her before she's digested. - A monster! What kind was it? - A dragon. One with brightred scales. - Brightred scales... The depths... The red dragon, hm!? - Well, I hope so, but… I hear dragons sleep most of the time to maintain those huge bodies. Its digestion should be far slower than other monsters'. - Listen, could I join you? - We'd love to have you along. You'd be a big help to us too. - Is that right! Say, thanks! - That's our line. - Cooking the red dragon's been a dream of mine for years! The red dragon... Should I go with classic steaks? Or would ground steak be better? Shabushabu's a good one too. Or if it has eggs, there's mother and child rice bowls...really, okay to eat that?

The thought did cross our minds...but nobody said anything. Dungeon food. After all...it's eat or be eaten. There's no hierarchy involved. Eating is, quite simply, the exclusive privilege of the living. Dungeon food. Ah, dungeon food.

Bibliography

H. Ai and X. Lu. A web-based system for automatic measurement of lexical complexity. In *27th Annual Symposium of the Computer-Assisted Language Consortium (CALICO-10)*, Amherst, MA, June 8-12 2010.

H. Ali Mehenni, S. Kobylyanskaya, I. Vasilescu, and L. Devillers. Nudges with Conversational Agents and Social Robots: A First Experiment with Children at a Primary School. In L. F. D'Haro, Z. Callejas, and S. Nakamura, editors, *Conversational Dialogue Systems for the Next Decade*, pages 257–270. Springer Singapore, Singapore, 2021a. ISBN 978-981-15-8395-7. doi: 10.1007/978-981-15-8395-7_19.

H. Ali Mehenni, S. Kobylyanskaya, I. Vasilescu, and L. Devillers. Children as candidates to verbal nudging in a human-robot experiment. ICMI '20 Companion, page 482–486, New York, NY, USA, 2021b. Association for Computing Machinery. ISBN 9781450380027. doi: 10.1145/3395035.3425224. URL https://doi.org/10.1145/3395035.3425224.

J. Anderson-Hsieh, R. Johnson, and K. Koehler. The Relationship Between Native Speaker Judgments of Nonnative Pronunciation and Deviance in Segmentais, Prosody, and Syllable Structure. *Language Learning*, 42(4):529–555, 1992. ISSN 1467-9922. doi: 10.1111/j.1467-1770.1992.tb01043.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-1770.1992.tb01043.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-1770.1992.tb01043.x.

D. Ariely and K. Wertenbroch. Procrastination, deadlines, and performance: Self-control by precommitment. *Psychological Science*, 13(3):219–224, 2002. doi: 10.1111/1467-9280.00441. PMID: 12009041.

M. Arlin and G. Roth. Pupils' use of time while reading comics and books. *American Educational Research Journal*, 15 (2):201–216, 1978. doi: 10.2307/1162460. URL https://doi.org/10.2307/1162460.

V. Aryadoust, S. Foo, and L. Y. Ng. What can gaze behaviors, neuroimaging data, and test scores tell us about test method effects and cognitive load in listening assessments? *Language Testing*, 39(1):56–89, Jan. 2022. ISSN 0265-5322. doi: 10.1177/02655322211026876. URL https://doi.org/10.1177/02655322211026876. Publisher: SAGE Publications Ltd.

O. Augereau, H. Fujiyoshi, and K. Kise. Towards an automated estimation of english skill via toeic score based on reading analysis. In *ICPR'16*, pages 1285–1290, 2016a.

O. Augereau, M. Matsubara, and K. Kise. Comic visualization on smartphones based on eye tracking. In *Proceedings of the 1st International Workshop on coMics ANalysis, Processing and Understanding*, MANPU '16, pages 1–4, New York, NY, USA, Dec. 2016b. Association for Computing Machinery. ISBN 978-1-4503-4784-6. doi: 10.1145/3011549.3011553. URL https://doi.org/10.1145/3011549.3011553.

O. Augereau, M. Iwata, and K. Kise. An Overview of Comics Research in Computer Science. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 03, pages 54–59, Nov. 2017. doi: 10.1109/ICDAR.2017.292. URL https://ieeexplore.ieee.org/abstract/document/8270237. ISSN: 2379-2140.

P. Avery and S. Ehrlich. *Teaching American English Pronunciation*. Oxford University Press, New York, 1992.

G. Azmat and N. Iriberri. The importance of relative performance feedback information: Evidence from a natural experiment using high school students. *Journal of Public Economics*, 94(7-8):435–452, 2010.

R. H. Baayen, R. Piepenbrock, and L. Gulikers. Celex2 ldc96l14. Web Download, 1995.

A. Baevski, H. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020. URL https://arxiv.org/abs/2006.11477.

N. Ballier, A. Méli, M. Amand, and J.-B. Yunès. Using Whisper LLM for Automatic Phonetic Diagnosis of L2 Speech: A Case Study with French Learners of English. In A. A. F. Mourad Abbas, editor, *6th International Conference on Natural Language and Speech Processing (ICNLSP 2023)*, volume 6 of *Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023)*, Trento (Italy), Italy, Dec. 2023. Mourad Abbas, Abed Alhakim Freihat. URL https://hal.science/hal-04547597. Issue: 282-292.

S. Bannò, K. M. Knill, M. Matassoni, V. Raina, and M. J. F. Gales. L2 proficiency assessment using self-supervised speech representations, Nov. 2022. URL http://arxiv.org/abs/2211.08849. arXiv:2211.08849.

S. Bax. The cognitive processing of candidates during reading tests: Evidence from eye-tracking. *Language Testing*, 30(4):441–465, Oct. 2013. ISSN 0265-5322. doi: 10.1177/0265532212473244. URL https://doi.org/10.1177/0265532212473244. Publisher: SAGE Publications Ltd.

T. Bent and A. Bradlow. The Interlanguage Speech Intelligibility Benefit. *The Journal of the Acoustical Society of America*, 114:1600–10, Oct. 2003. doi: 10.1121/1.1603234.

Y. Berzak, B. Katz, and R. Levy. Assessing language proficiency from eye movements in reading, 2018. URL https://arxiv.org/abs/1804.07329.

Y. Berzak, C. Nakamura, A. Smith, E. Weng, B. Katz, S. Flynn, and R. Levy. CELER: A 365-Participant Corpus of Eye Movements in L1 and L2 English Reading. *Open Mind*, 6:41–50, 07 2022. ISSN 2470-2986. doi: 10.1162/opmi_a_00054. URL https://doi.org/10.1162/opmi_a_00054.

P. Blignaut and D. Wium. Eye-tracking data quality as affected by ethnicity and experimental design. *Behavior Research Methods*, 46(1):67–80, 2014. doi: 10.3758/s13428-013-0343-0. URL https://doi.org/10.3758/s13428-013-0343-0.

P. Boersma and D. Weenink. Praat: doing phonetics by computer [computer program], 1992–2022. Version 6.2.06, retrieved 23 January 2022 from https://www.praat.org.

O.-S. Bohn and J. E. Flege. Interlingual identification and the role of foreign language experience in l2 vowel perception. *Applied Psycholinguistics*, 11(3):303–328, 1990.

H. R. Bosker, H. Quené, T. Sanders, and N. H. de Jong. The perception of fluency in native and nonnative speech. *Language Learning*, 64(3):579–614, 2014. doi: https://doi.org/10.1111/lang.12067. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/lang.12067.

P. Boula de Mareüil and B. Vieru-Dimulescu. The contribution of prosody to the perception of foreign accent. *Phonetica*, 63(4):247–267, 2006. URL https://karger.com/pho/article-abstract/63/4/247/274278. Publisher: S. Karger AG Basel, Switzerland.

L. Bovens. Nudges and Cultural Variance: a Note on Selinger and Whyte. *Knowledge, Technology & Policy*, 23, Dec. 2010. doi: 10.1007/s12130-010-9128-2.

M. G. Busà, M. Urbani, et al. A cross linguistic analysis of pitch range in english l1 and l2. In *ICPhS*, pages 380–383, 2011.

G. Buscher, A. Dengel, and L. van Elst. Eye movements as implicit relevance feedback. In *CHI '08 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '08, page 2991–2996, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605580128. doi: 10.1145/1358628.1358796. URL https://doi.org/10.1145/1358628.1358796.

Y. Cao, L. Zhou, S. Lee, L. Cabello, M. Chen, and D. Hershcovich. Assessing cross-cultural alignment between chatgpt and human societies: An empirical study, 2023. URL https://arxiv.org/abs/2303.17466.

M. Capliez. Typologie des erreurs de production d'anglais des francophones : segments vs. suprasegments. *Recherche et pratiques pédagogiques en langues. Cahiers de l'Apliut*, (Vol. XXX N° 3):44–60, Oct. 2011. ISSN 2257-5405. doi: 10.4000/apliut.1645. URL https://journals.openedition.org/apliut/1645. Number: Vol. XXX N° 3 Publisher: Association des Professeurs de Langues des IUT (APLIUT).

M. Capliez. *Acquisition and learning of English phonology by French speakers: on the roles of segments and suprasegments*. PhD thesis, Université Lille III – CHARLES DE GAULLE, École Doctorale Sciences de l'Homme et de la Société, 09 2016.

B. L. Castleman and L. C. Page. Summer nudging: Can personalized text messages and peer mentor outreach increase college going among low-income high school graduates? *Journal of Economic Behavior & Organization*, 115: 144–160, 2015. ISSN 0167-2681. Behavioral Economics of Education.

D. Chang and C. Weng. Late esl learners' difficulties of distinction between lax and tense vowels. In J. Levis and K. LeVelle, editors, *Proceedings of the 4th Pronunciation in Second Language Learning and Teaching Conference*, pages 129–139, Ames, IA, 2013. Iowa State University. Conference held in Aug. 2012.

C.-H. Chen, C.-S. Koong, and C. Liao. Influences of Integrating Dynamic Assessment into a Speech Recognition Learning Design to Support Students' English Speaking Skills, Learning Anxiety and Cognitive Load. *Educational Technology & Society*, 25(1):1–14, 2022. ISSN 1176-3647. URL https://www.jstor.org/stable/48647026. Publisher: International Forum of Educational Technology & Society.

B. R. Chiswick and P. W. Miller. Linguistic distance: A quantitative measure of the distance between english and other languages. *Journal of Multilingual and Multicultural Development*, 26:1–11, 2005.

K. Conklin and A. Pellicer-Sánchez. Using eye-tracking in applied linguistics and second language research. *Second Language Research*, 32(3):453–467, 2016. doi: 10.1177/0267658316637401. URL https://doi.org/10.1177/0267658316637401.

K. Conklin, A. Pellicer-Sánchez, and G. Carrol. *Eye-tracking: A guide for applied linguistics research*. Cambridge University Press, 2018.

L. Copeland, T. Gedeon, and B. Mendis. Predicting reading comprehension scores from eye movements using artificial neural networks and fuzzy output error. *Artificial Intelligence Research*, 3, 07 2014. doi: 10.5430/air.v3n3p35.

I. Cunnings. Interference in native and non-native sentence processing. *Bilingualism: Language and Cognition*, 20(4): 712–721, 2017. doi: 10.1017/S1366728916001243.

M. T. Damgaard and H. Nielsen. Nudging in education. *Economics of Education Review*, 64(C):313–342, 2018. URL https://EconPapers.repec.org/RePEc:eee:ecoedu:v:64:y:2018:i:c:p:313-342.

A. Davies. *The Native Speaker: Myth and Reality*. Multilingual Matters, Bristol, Blue Ridge Summit, 2003. ISBN 9781853596247. doi: doi:10.21832/9781853596247. URL https://doi.org/10.21832/9781853596247.

D. Davis, I. Jivet, R. F. Kizilcec, G. Chen, C. Hauff, and G.-J. Houben. Follow the successful crowd: raising mooc completion rates through social comparison at scale. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, LAK '17, page 454–463, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450348706. doi: 10.1145/3027385.3027411. URL https://doi.org/10.1145/3027385.3027411.

V. De Fino, L. Fontan, J. Pinquier, I. Ferrané, and S. Detey. Prediction of l2 speech proficiency based on multi-level linguistic features. In *23rd INTERSPEECH Conference : Human and Humanizing Speech Technology (INTERSPEECH 2022)*, Incheon, South Korea, Sep 2022. The Acoustical Society of Korea. hal-03775950.

N. de Jong. Fluency in second language assessment. pages 203–218. Mar. 2016. ISBN 978-1-61451-382-7. doi: 10.1515/9781614513827-015.

N. H. de Jong and T. Wempe. Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, 41(2):385–390, May 2009. ISSN 1554-3528. doi: 10.3758/BRM.41.2.385. URL https://doi.org/10.3758/BRM.41.2.385.

N. H. de Jong, R. Groenhout, R. Schoonen, and J. H. Hulstijn. Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behavior. *Applied Psycholinguistics*, 36(2):223–243, Mar. 2015. ISSN 0142-7164, 1469-1817. doi: 10.1017/S0142716413000210. URL https://www.cambridge.org/core/journals/applied-psycholinguistics/article/abs/second-language-fluency-speaking-style-or-proficiency-correcting-measures-of-second-language-fluency-f87E4C697BFA1A80702518CB52F630121.

N. De Weers and M. Munro. The role of duration in japanese speakers' productions of english vowels. In J. Levis, editor, *Proceedings of the 9th Pronunciation in Second Language Learning and Teaching Conference*, pages 41–53, Ames, IA, 2018. Iowa State University. ISSN 2380-9566, University of Utah, September, 2017.

S. Detey. A longitudinal interphonological corpus of japanese learners of french, 2011–2019.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL https://arxiv.org/abs/1810.04805.

M. El Baha, O. Augereau, S. Kobylyanskaya, I. Vasilescu, and L. Devillers. Eye got it: A system for automatic calculation of the eye-voice span. In S. Uchida, E. Barney, and V. Eglin, editors, *Document Analysis Systems*, pages 713–725, Cham, 2022. Springer International Publishing. ISBN 978-3-031-06555-2.

A. Eneh and O. Eneh. Enhancing pupils' reading achievement by use of comics and cartoons in teaching reading. *Journal of Applied Science*, 11(3):8058–62, 2008.

D. Esteve, M. Perea, B. Angele, et al. Individual differences in word skipping during reading in english as l2. *Psychonomic Bulletin & Review*, 2024. doi: 10.3758/s13423-024-02529-w.

J. Fathi, M. Rahimi, and A. Derakhshan. Improving EFL learners' speaking skills and willingness to communicate via artificial intelligence-mediated interactions. *System*, 121:103254, Apr. 2024. ISSN 0346-251X. doi: 10.1016/j.system.2024.103254. URL https://www.sciencedirect.com/science/article/pii/S0346251X24000368.

G. Feng, K. Miller, H. Shu, and H. Zhang. Orthography and the development of reading processes: An eye-movement study of chinese and english. *Child Development*, 80(3):720–735, 2009. doi: https://doi.org/10.1111/j.1467-8624.2009.01293.x. URL https://srcd.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-8624.2009.01293.x.

M. Ferro, C. Marzi, A. Nadalini, L. Taxitari, A. Lento, and V. Pirrelli. ReadLet: A Dataset for Oral, Visual and Tactile Text Reading Data of Early and Mature Readers. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13595–13609, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.1188.

J. Field. Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly*, 39, 09 2005. doi: 10.2307/3588487.

J. Flege. Speech learning in a second language. In C. Ferguson, L. Menn, and C. Stoel-Gammon, editors, *Phonological development: Models, research, implications*, pages 565–604. York Press, Timonium, 1992.

J. Flege. *Second language speech learning: Theory, findings and problems*, pages 229–273. 01 1995.

C. Frenck-Mestre. Eye-movement recording as a tool for studying syntactic processing in a second language: a review of methodologies and experimental findings. *Second Language Research*, 21(2):175–198, 2005. doi: 10.1191/0267658305sr257oa. URL https://doi.org/10.1191/0267658305sr257oa.

U. Garain, O. Pandit, O. Augereau, A. Okoso, and K. Kise. *Identification of Reader Specific Difficult Words by Analyzing Eye Gaze and Document Content*. Nov. 2017. doi: 10.1109/ICDAR.2017.221. Pages: 1351.

A. Godfroid and B. Hui. Five common pitfalls in eye-tracking research. *Second Language Research*, 36(3):277–305, 2020. doi: 10.1177/0267658320921218. URL https://doi.org/10.1177/0267658320921218.

S. Goulas and R. Megalokonomou. Knowing who you are: The effect of feedback information on short and long term outcomes. Economic rese, University of Warwick - Department of Economics, 2015.

C. Graham and N. Roll. Evaluating OpenAI's Whisper ASR: Performance analysis across diverse accents and speaker traits. *JASA Express Letters*, 4(2):025206, Feb. 2024. ISSN 2691-1191. doi: 10.1121/10.0024876. URL https://doi.org/10.1121/10.0024876.

P. Grainger. The impact of cultural background on the choice of language learning strategies in the jfl context. *System*, 40(4):483–493, 2012.

R. Gretter, M. Matassoni, K. Allgaier, S. Tchistiakova, and D. Falavigna. Automatic Assessment of Spoken Language Proficiency of Non-native Children. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7435–7439, May 2019. doi: 10.1109/ICASSP.2019.8683268. URL https://ieeexplore.ieee.org/abstract/document/8683268. ISSN: 2379-190X.

L. D. Hahn. Primary Stress and Intelligibility: Research to Motivate the Teaching of Suprasegmentals. *TESOL Quarterly*, 38(2):201–223, 2004. ISSN 00398322. doi: 10.2307/3588378. URL http://www.jstor.org/stable/3588378. Publisher: [Wiley, Teachers of English to Speakers of Other Languages, Inc. (TESOL)].

K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, and J. Van de Weijer. *Eye tracking: A comprehensive guide to methods and measures*. oup Oxford, 2011.

C.-N. Hsieh, K. Zechner, and X. Xi. Features Measuring Fluency and Pronunciation. pages 101–122. Nov. 2019. ISBN 978-1-315-16510-3. doi: 10.4324/9781315165103-7.

X. Hu and V. Aryadoust. A systematic review of eye-tracking technology in second language research. *Languages*, 9 (4):141, 2024.

F. Huettig, J. Rommers, and A. Meyer. Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta psychologica*, 137:151–71, 06 2011. doi: 10.1016/j.actpsy.2010.11.003.

M. Hutin, S. Kobylyanskaya, and L. Devillers. Nudges in Technology-Mediated Knowledge Transfer: Two Experimental Designs. In *Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 267–273, Cambridge United Kingdom, Sept. 2022. ACM. ISBN 978-1-4503-9423-9. doi: 10.1145/3544793.3560379.

T. Isaacs. *Assessing Pronunciation*, chapter 8, pages 140–155. John Wiley & Sons, Ltd, 2013. ISBN 9781118411360. doi: https://doi.org/10.1002/9781118411360.wbcla012. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118411360.wbcla012.

A. Ito, M. J. Pickering, and M. Corley. Investigating the time-course of phonological prediction in native and non-native speakers of english: A visual world eye-tracking study. *Journal of Memory and Language*, 98:1–11, 2018. ISSN 0749-596X. doi: https://doi.org/10.1016/j.jml.2017.09.002. URL https://www.sciencedirect.com/science/article/pii/S0749596X17300633.

N. Iwashita, A. Brown, T. McNamara, and S. O'hagan. Assessed levels of second language speaking proficiency: How distinct? *Applied linguistics*, 29(1):24–49, 2008.

M. Iwata, A. Ito, and K. Kise. A study to achieve manga character retrieval method for manga images. In *2014 11th IAPR International Workshop on Document Analysis Systems*, pages 309–313, 2014. doi: 10.1109/DAS.2014.60.

E. Izumi, K. Uchimoto, and H. Isahara. The nict jle corpus: Exploiting the language learners' speech database for research and education. *International Journal of the Computer, the Internet and Management*, 12(2):119–125, 2004.

X. Jiang and M. Pell. Predicting confidence and doubt in accented speakers: Human perception and machine learning experiments. In *Speech Prosody 2018*, pages 269–273, 2018. doi: 10.21437/SpeechProsody.2018-55.

N. Kalashnikova, I. Vasilescu, and L. Devillers. Linguistic nudges and verbal interaction with robots, smart-speakers, and humans. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10555–10564, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.923.

H. Kang. Understanding online reading through the eyes of first and second language readers: An exploratory study. *Computers & Education*, 73:1–8, 2014. ISSN 0360-1315. doi: https://doi.org/10.1016/j.compedu.2013.12.005. URL https://www.sciencedirect.com/science/article/pii/S0360131513003291.

O. Kang. Relative salience of suprasegmental features on judgments of l2 comprehensibility and accentedness. *System*, 38:301–315, 06 2010. doi: 10.1016/j.system.2010.01.005.

O. Kang. Relative impact of pronunciation features on ratings of non-native speakersâ€™ oral proficiency. *Pronunciation in Second Language Learning and Teaching Proceedings*, 4(1), 2012.

O. Kang, D. Rubin, and L. Pickering. Suprasegmental measures of accentedness and judgments of language learner proficiency in oral english. *The Modern Language Journal*, 94(4):554–566, 2010. doi: https://doi.org/10.1111/j.1540-4781.2010.01091.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-4781.2010.01091.x.

M. C. Kelley and B. V. Tucker. A comparison of four vowel overlap measures. *The Journal of the Acoustical Society of America*, 147(1):137–145, Jan. 2020. ISSN 0001-4966. doi: 10.1121/10.0000494. URL https://doi.org/10.1121/10.0000494.

S. Kennedy and P. Trofimovich. Intelligibility, comprehensibility, and accentedness of l2 speech: The role of listener experience and semantic context. *The Canadian Modern Language Review*, 64(3):459–489, 2008. doi: 10.3138/cmlr.64.3.459. URL https://doi.org/10.3138/cmlr.64.3.459.

A. Kipp, L. L, J. Mariani, and F. Schiel. Translanguage English Database (TED) Speech. 2002. doi: 10.35111/D6H3-PF89. URL https://catalog.ldc.upenn.edu/LDC2002S04. Artwork Size: 2936012 KB Pages: 2936012 KB.

T. Kisler, U. Reichel, and F. Schiel. Multilingual processing of speech via web services. *Computer Speech & Language*, 45:326–347, Sept. 2017. ISSN 0885-2308. doi: 10.1016/j.csl.2017.01.005.

Y. Kobayashi and M. Abe. Automated scoring of l2 spoken english with random forests. *Journal of Pan-Pacific Association of Applied Linguistics*, 20(1):55–73, 2016.

S. Kobylyanskaya. Speech and eye tracking features for l2 acquisition: A multimodal experiment. In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners' and Doctoral Consortium: 23rd International Conference, AIED 2022, Durham, UK, July 27–31,*

2022, Proceedings, Part II, page 47–52, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-11646-9. doi: 10.1007/978-3-031-11647-6_8. URL https://doi.org/10.1007/978-3-031-11647-6_8.

S. Kobylyanskaya, I. Vasilescu, and L. Devillers. Vers la compréhension des difficultés de lecture en L2 à travers des paramètres acoustiques et de mouvement des yeux. 2023.

M. Kondo, H. Tsubaki, and Y. Sagisaka. Segmental variation of japanese speakers' english: Analysis of "the north wind and the sun" in aesop corpus. *Journal of the Phonetic Society of Japan*, 19(1):3–17, 2015. doi: 10.24467/onseikenkyu.19.1_3.

M. Kuhn, P. Schwanenflugel, and E. Meisinger. Aligning theory and assessment of reading fluency: Automaticity, prosody, and definitions of fluency. *Reading Research Quarterly*, 45:232–253, Jan. 2010.

K. Kunze, S. Sanchez, T. Dingler, O. Augereau, K. Kise, M. Inami, and T. Tsutomu. The augmented narrative: toward estimating reader engagement. In *Proceedings of the 6th Augmented Human International Conference*, AH '15, pages 163–164, New York, NY, USA, Mar. 2015. Association for Computing Machinery. ISBN 978-1-4503-3349-8. doi: 10.1145/2735711.2735814. URL https://doi.org/10.1145/2735711.2735814.

A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen. lmertest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13):1–26, 2017. doi: 10.18637/jss.v082.i13. URL https://www.jstatsoft.org/index.php/jss/article/view/v082i13.

K. Kyriakopoulos, K. Knill, and M. Gales. A deep learning approach to assessing non-native pronunciation of english using phone distances. In *Interspeech 2018*, pages 1626–1630, 2018. doi: 10.21437/Interspeech.2018-1087.

K. Kyriakopoulos, K. M. Knill, and M. J. Gales. A Deep Learning Approach to Automatic Characterisation of Rhythm in Non-Native English Speech. In *Interspeech 2019*, pages 1836–1840. ISCA, Sept. 2019. doi: 10.21437/Interspeech.2019-3186. URL https://www.isca-archive.org/interspeech_2019/kyriakopoulos19_interspeech.html.

S. Lallé, R. Murali, and C. Conati. Predicting co-occurring emotions from eye-tracking and interaction data in metatutor. In *Proceedings of the International Conference on Artificial Intelligence in Education (AIED)*, pages 241–254. Springer, 2021.

H. Lee and J. H. Lee. The effects of ai-guided individualized language learning: A meta-analysis. *Language, Learning and Technology*, 28:134–162, 06 2024.

X. Lu. The relationship of lexical richness to the quality of esl learners' oral narratives. *The Modern Language Journal*, 96(2):190–208, 2012.

A. Malinin, A. Ragni, K. Knill, and M. Gales. Incorporating uncertainty into deep learning for spoken language assessment. In R. Barzilay and M.-Y. Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational*

Linguistics (Volume 2: Short Papers), pages 45–50, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-2008. URL https://aclanthology.org/P17-2008.

M. Matsubara, O. Augereau, C. L. Sanches, and K. Kise. Emotional arousal estimation while reading comics based on physiological signal analysis. In *Proceedings of the 1st International Workshop on coMics ANalysis, Processing and Understanding*, MANPU '16, pages 1–4, New York, NY, USA, Dec. 2016. Association for Computing Machinery. ISBN 978-1-4503-4784-6. doi: 10.1145/3011549.3011556. URL https://doi.org/10.1145/3011549.3011556.

K. McDonough. Identifying the impact of negative feedback and learners' responses on esl question development. *Studies in Second Language Acquisition*, 27(1):79–103, 2005. doi: 10.1017/S0272263105050047.

M. Mirzaei and K. Meshgi. The use of machine learning in developing learner-adaptive tools for second language acquisition. In *CALL for all Languages - EUROCALL 2023 Short Papers*, University of Iceland, Reykjavik, August 15-18 2023. doi: 10.4995/EuroCALL2023.2023.16996.

M. A. Mohsen. Computer-mediated corrective feedback to improve l2 writing skills: A meta-analysis. *Journal of Educational Computing Research*, 60(5):1253–1276, 2022. doi: 10.1177/07356331211064066. URL https://doi.org/10.1177/07356331211064066.

M. J. Munro and T. M. Derwing. MODELING PERCEPTIONS OF THE ACCENTEDNESS AND COMPREHENSIBILITY OF L2 SPEECH The Role of Speaking Rate. *Studies in Second Language Acquisition*, 23 (4):451–468, Dec. 2001. ISSN 1470-1545, 0272-2631. doi: 10.1017/S0272263101004016. URL https://www.cambridge.org/core/journals/studies-in-second-language-acquisition/article/abs/modeling-perceptions-of-the-accentedness-and-comprehensibility-of-l2-speech-the-role-of-speaking-rate/E22C75E244694BF6D6E083149FBEFA51.

M. J. Munro and T. M. Derwing. The functional load principle in ESL pronunciation instruction: An exploratory study. *System*, 34(4):520–531, Dec. 2006. ISSN 0346-251X. doi: 10.1016/j.system.2006.09.004. URL https://www.sciencedirect.com/science/article/pii/S0346251X06000856.

S. Murakami and M. Bryce. Manga as an educational medium. *The International Journal of the Humanities*, 7:47–55, 01 2009. doi: 10.18848/1447-9508/CGP/v07i10/42761.

S. Nahatame, T. Ogiso, Y. Kimura, and Y. Ushiro. TECO: An Eye-tracking Corpus of Japanese L2 English Learners' Text Reading. *Research Methods in Applied Linguistics*, 3(2):100123, Aug. 2024. ISSN 2772-7661. doi: 10.1016/j.rmal.2024.100123. URL https://www.sciencedirect.com/science/article/pii/S2772766124000296.

T. Neuhaus. Nudging esl learning - improving second language learning by improving decision architectures. *International Journal of Education Humanities and Social Science*, 4(01):116, 2021. ISSN 2582-0745. URL http://ijehss.com/.

J. Novikova, A. Balagopalan, K. Shkaruta, and F. Rudzicz. Lexical features are more vulnerable, syntactic features have more predictive power. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 431–443, 2019.

J. Nycz and L. Hall-Lew. Best practices in measuring vowel merger. *Proceedings of Meetings on Acoustics*, 20(1):060008, Aug. 2014. ISSN 1939-800X. doi: 10.1121/1.4894063. URL https://doi.org/10.1121/1.4894063.

M. Nyström and K. Holmqvist. An adaptive algorithm for fixation, saccade, and glissade detection in eyetracking data. *Behavior Research Methods*, 42(1):188–204, 2010. doi: 10.3758/BRM.42.1.188. URL https://doi.org/10.3758/BRM.42.1.188.

OCDE. *Opportunities, guidelines and guardrails for effective and equitable use of AI in education*. OCDE, 2023. doi: doi:https://doi.org/10.1787/2b39e98b-en. URL https://www.oecd-ilibrary.org/content/component/2b39e98b-en.

C. of Europe. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Council of Europe, 2001.

OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, S. Balaji, V. Balcom, P. Baltescu, H. Bao, M. Bavarian, J. Belgum, I. Bello, J. Berdine, G. Bernadett-Shapiro, C. Berner, L. Bogdonoff, O. Boiko, M. Boyd, A.-L. Brakman, G. Brockman, T. Brooks, M. Brundage, K. Button, T. Cai, R. Campbell, A. Cann, B. Carey, C. Carlson, R. Carmichael, B. Chan, C. Chang, F. Chantzis, D. Chen, S. Chen, R. Chen, J. Chen, M. Chen, B. Chess, C. Cho, C. Chu, H. W. Chung, D. Cummings, J. Currier, Y. Dai, C. Decareaux, T. Degry, N. Deutsch, D. Deville, A. Dhar, D. Dohan, S. Dowling, S. Dunning, A. Ecoffet, A. Eleti, T. Eloundou, D. Farhi, L. Fedus, N. Felix, S. P. Fishman, J. Forte, I. Fulford, L. Gao, E. Georges, C. Gibson, V. Goel, T. Gogineni, G. Goh, R. Gontijo-Lopes, J. Gordon, M. Grafstein, S. Gray, R. Greene, J. Gross, S. S. Gu, Y. Guo, C. Hallacy, J. Han, J. Harris, Y. He, M. Heaton, J. Heidecke, C. Hesse, A. Hickey, W. Hickey, P. Hoeschele, B. Houghton, K. Hsu, S. Hu, X. Hu, J. Huizinga, S. Jain, S. Jain, J. Jang, A. Jiang, R. Jiang, H. Jin, D. Jin, S. Jomoto, B. Jonn, H. Jun, T. Kaftan, Łukasz Kaiser, A. Kamali, I. Kanitscheider, N. S. Keskar, T. Khan, L. Kilpatrick, J. W. Kim, C. Kim, Y. Kim, J. H. Kirchner, J. Kiros, M. Knight, D. Kokotajlo, Łukasz Kondraciuk, A. Kondrich, A. Konstantinidis, K. Kosic, G. Krueger, V. Kuo, M. Lampe, I. Lan, T. Lee, J. Leike, J. Leung, D. Levy, C. M. Li, R. Lim, M. Lin, S. Lin, M. Litwin, T. Lopez, R. Lowe, P. Lue, A. Makanju, K. Malfacini, S. Manning, T. Markov, Y. Markovski, B. Martin, K. Mayer, A. Mayne, B. McGrew, S. M. McKinney, C. McLeavey, P. McMillan, J. McNeil, D. Medina, A. Mehta, J. Menick, L. Metz, A. Mishchenko, P. Mishkin, V. Monaco, E. Morikawa, D. Mossing, T. Mu, M. Murati, O. Murk, D. Mély, A. Nair, R. Nakano, R. Nayak, A. Neelakantan, R. Ngo, H. Noh, L. Ouyang, C. O'Keefe, J. Pachocki, A. Paino, J. Palermo, A. Pantuliano, G. Parascandolo, J. Parish, E. Parparita, A. Passos, M. Pavlov, A. Peng, A. Perelman, F. de Avila Belbute Peres, M. Petrov, H. P. de Oliveira Pinto, Michael, Pokorny, M. Pokrass, V. H. Pong, T. Powell, A. Power, B. Power, E. Proehl, R. Puri, A. Rad-

ford, J. Rae, A. Ramesh, C. Raymond, F. Real, K. Rimbach, C. Ross, B. Rotsted, H. Roussez, N. Ryder, M. Saltarelli, T. Sanders, S. Santurkar, G. Sastry, H. Schmidt, D. Schnurr, J. Schulman, D. Selsam, K. Sheppard, T. Sherbakov, J. Shieh, S. Shoker, P. Shyam, S. Sidor, E. Sigler, M. Simens, J. Sitkin, K. Slama, I. Sohl, B. Sokolowsky, Y. Song, N. Staudacher, F. P. Such, N. Summers, I. Sutskever, J. Tang, N. Tezak, M. B. Thompson, P. Tillet, A. Tootoonchian, E. Tseng, P. Tuggle, N. Turley, J. Tworek, J. F. C. Uribe, A. Vallone, A. Vijayvergiya, C. Voss, C. Wainwright, J. J. Wang, A. Wang, B. Wang, J. Ward, J. Wei, C. Weinmann, A. Welihinda, P. Welinder, J. Weng, L. Weng, M. Wiethoff, D. Willner, C. Winter, S. Wolrich, H. Wong, L. Workman, S. Wu, J. Wu, M. Wu, K. Xiao, T. Xu, S. Yoo, K. Yu, Q. Yuan, W. Zaremba, R. Zellers, C. Zhang, M. Zhang, S. Zhao, T. Zheng, J. Zhuang, W. Zhuk, and B. Zoph. Gpt-4 technical report, 2024. URL `https://arxiv.org/abs/2303.08774`.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

A. Pellicer-Sánchez, L. Vilkaitė-Lozdienė, and A. Siyanova-Chanturia. *Examining L2 Learners' Confidence of Collocational Knowledge*, pages 117–136. Multilingual Matters, Bristol, Blue Ridge Summit, 2022. ISBN 9781788923750. doi: doi:10.21832/9781788923750-010. URL `https://doi.org/10.21832/9781788923750-010`.

R. Pérez-Ramón, M. Kondo, S. Detey, L. Fontan, M. Amand, and T. Kamiyama. Nativeness and Intelligibility of Japanese accented English Consonants by French Listeners. In *International Congress of Phonetic Sciences (ICPHS 2023)*, pages 2581–2585, Prague, Czech Republic, Aug. 2023. URL `https://hal.science/hal-04191719`.

L. Perkhofer and O. Lehner. Using Gaze Behavior to Measure Cognitive Load. In F. D. Davis, R. Riedl, J. vom Brocke, P.-M. Léger, and A. B. Randolph, editors, *Information Systems and Neuroscience*, pages 73–83, Cham, 2019. Springer International Publishing. ISBN 978-3-030-01087-4. doi: 10.1007/978-3-030-01087-4_9.

J. Peters, M. Frank, and M. Rohloff. Vocal fold vibratory patterns in bilingual speakers of low and high german. In R. Skarnitzl and J. Volín, editors, *Proceedings of the 20th International Congress of Phonetic Sciences – ICPhS 2023*. International Phonetic Association, 2023.

N. Phan, A. von Zansen, M. Kautonen, E. Voskoboinik, T. Grosz, R. Hilden, and M. Kurimo. Automated content assessment and feedback for Finnish L2 learners in a picture description speaking task. pages 317–321, 2024. doi: 10.21437/Interspeech.2024-1166. URL `https://www.isca-archive.org/interspeech_2024/phan24_interspeech.html`.

K. L. Pike. *The Intonation of American English*. Ann Arbor: University of Michigan Press, 1945.

K. C. S. Pillai. Some New Test Criteria in Multivariate Analysis. *The Annals of Mathematical Statistics*, 26(1):117–121, Mar. 1955. ISSN 0003-4851, 2168-8990. doi: 10.1214/aoms/1177728599. URL

https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-26/issue-1/Some-New-Test-Criteria-in-Multivariate-Analysis/10.1214/aoms/1177728599.full. Publisher: Institute of Mathematical Statistics.

L. Polyanskaya, M. Ordin, and M. G. Busà. Relative salience of speech rhythm and speech rate on perceived foreign accent in a second language. *Language and Speech*, 60(3):333–355, 2017. doi: 10.1177/0023830916648720. URL https://doi.org/10.1177/0023830916648720. PMID: 28915779.

A. Popescu, L. Lamel, I. Vasilescu, and L. Devillers. An investigation of syllable position /l/ allophony in L2 English learners usingWord Error Rate as an index of phonetic proficiency. In *13th International Seminar on Speech Production (ISSP2024)*, Autrans, France, May 2024. URL https://hal.science/hal-04451662.

C. Prator and B. Robinett. *Manual of American English Pronunciation*. Harcourt Brace Jovanovich Japan, Inc, Tokyo, 1986.

K. Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124 (3):372–422, 1998. ISSN 1939-1455. doi: 10.1037/0033-2909.124.3.372. Place: US Publisher: American Psychological Association.

K. Rayner, T. J. Slattery, D. Drieghe, and S. P. Liversedge. Eye movements and word skipping during reading: Effects of word length and predictability. *Journal of Experimental Psychology: Human Perception and Performance*, 37(2): 514–528, 2011. doi: 10.1037/a0020990.

M. E. Renwick, I. Vasilescu, C. Dutrey, L. Lamel, and B. Vieru. Marginal contrast among romanian vowels: Evidence from asr and functional load. In *Interspeech 2016*, volume 2016, pages 2433–2437, 2016.

A. Révész, M. Stainer, J. Jung, M. Lee, and M. Michel. Using eye-tracking as a tool to develop l2 lexical skills. *Language Learning and Technology*, 2022.

C. Rigaud, N. Le Thanh, J.-C. Burie, J.-M. Ogier, M. Iwata, E. Imazu, and K. Kise. Speech balloon and speaker association for comics and manga understanding. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 351–355. IEEE, 2015. URL https://ieeexplore.ieee.org/abstract/document/7333782/.

C. Rigaud, T.-N. Le, J.-C. Burie, J.-M. Ogier, S. Ishimaru, M. Iwata, and K. Kise. Semi-automatic Text and Graphics Extraction of Manga Using Eye Tracking Information. In *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, pages 120–125, Santorini, Greece, Apr. 2016. IEEE. ISBN 978-1-5090-1792-8. doi: 10.1109/DAS.2016.72. URL http://ieeexplore.ieee.org/document/7490104/.

T. Riney and J. Anderson-Hsieh. Japanese pronunciation of english. *JALT Journal*, 1993.

L. Roberts and A. Siyanova-Chanturia. Using eye-tracking to investigate topics in l2 acquisition and l2 processing. *Studies in Second Language Acquisition*, 35, 06 2013. doi: 10.1017/S0272263112000861.

R. Rose. Silent and filled pauses and speech planning in first and second language production. *Proceedings of DiSS*, 2017:49–52, 2017. URL https://www.ida.liu.se/~robek28/conferences/diss2017/DiSS2017_Rose.pdf.

K. Saito and Y. Akiyama. Linguistic correlates of comprehensibility in second language Japanese speech. *Journal of Second Language Pronunciation*, 3(2):199–217, Jan. 2017. ISSN 2215-1931, 2215-194X. doi: 10.1075/jslp.3.2.02sai. URL https://www.jbe-platform.com/content/journals/10.1075/jslp.3.2.02sai. Publisher: John Benjamins.

P. Sarada. Comics as a powerful tool to enhance english language usage. *IUP Journal of English Studies*, 11(1):60, 2016.

A. Schleicher. The impact of covid-19 on education: insights from education at a glance 2020. *OECD*, 2020.

G. Schwartz. Voice quality and l2 proficiency in the english tense-lax contrast. *Anglophonia*, 27, 2019. doi: 10.4000/anglophonia.2058. URL http://journals.openedition.org/anglophonia/2058. Online since 25 November 2019, connection on 16 October 2024.

E. Selinger and K. P. Whyte. Competence and trust in choice architecture. *Knowledge, Technology & Policy*, 23:461–482, 2010. doi: 10.1007/s12130-010-9127-3. URL https://doi.org/10.1007/s12130-010-9127-3.

N. Siegelman, S. Schroeder, C. Acarturk, H.-D. Ahn, S. Alexeeva, S. Amenta, R. Bertram, R. Bonandrini, M. Brysbaert, D. Chernova, S. Fonseca, N. Dirix, W. Duyck, A. Fella, R. Frost, C. Gattei, A. Kalaitzi, N. Kwon, K. Lõo, and V. Kuperman. Expanding horizons of cross-linguistic research on reading: The Multilingual Eye-movement Corpus (MECO). *Behavior Research Methods*, 54:1–21, Feb. 2022. doi: 10.3758/s13428-021-01772-6.

H. Singer, S. J. Samuels, and J. Spiroff. The effect of pictures and contextual conditions on learning responses to printed words. *Reading Research Quarterly*, 9(4):555–567, 1973-1974. doi: 10.2307/747002. URL https://doi.org/10.2307/747002.

M. Takahashi and S. Kiyokawa. Eye movement during silent and oral reading: How can we compensate the loss of multisensory process during silent reading? *i-Perception*, 2:842–842, 10 2011. doi: 10.1068/ic842.

T. Takaike, M. Iwata, and K. Kise. Estimation of unknown words using speech and eye gaze when reading aloud comics. In *Pattern Recognition, Computer Vision, and Image Processing. ICPR 2022 International Workshops and Challenges: Montreal, QC, Canada, August 21–25, 2022, Proceedings, Part II*, page 91–106, Berlin, Heidelberg, 2023. Springer-Verlag. ISBN 978-3-031-37741-9. doi: 10.1007/978-3-031-37742-6_7. URL https://doi.org/10.1007/978-3-031-37742-6_7.

P. Tavakoli. Pausing patterns: differences between L2 learners and native speakers. *ELT Journal*, 65(1):71–79, Jan. 2011. ISSN 0951-0893. doi: 10.1093/elt/ccq020. URL https://doi.org/10.1093/elt/ccq020.

P. Tavakoli, P. Skehan, and R. Ellis. Planning and task performance in a second language. *Strategic planningtask structureand performance testing*, 2005.

R. H. Thaler and C. R. Sunstein. *Nudge: Improving decisions about health, wealth, and happiness*. Nudge: Improving decisions about health, wealth, and happiness. Yale University Press, New Haven, CT, US, 2008. ISBN 978-0-300-12223-7. Pages: x, 293.

G. Unser-Schutz. Manga as a linguistic resource for learning. *JALT2010 Conference Proceedings*, 01 2011.

M. van Oldenbeek, T. J. Winkler, J. Buhl-Wiggers, and D. Hardt. Nudging in blended learning: Evaluation of email-based progress feedback in a flipped-classroom information systems course. In *Proceedings of the 27th European Conference on Information Systems (ECIS)*, June 8-14 2019. ISBN 978-1-7336325-0-8. URL https://aisel.aisnet.org/ecis2019_rp/186. Research Papers.

A. Vargo, M. Iwata, M. Hutin, S. Kobylyanskaya, I. Vasilescu, O. Augereau, K. Watanabe, S. Ishimaru, B. Tag, T. Dingler, K. Kise, L. Devillers, and A. Dengel. Learning cyclotron: An ecosystem of knowledge circulation. In *Adjunct Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2022 ACM International Symposium on Wearable Computers*, UbiComp/ISWC '22 Adjunct, page 308–312, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394239. doi: 10.1145/3544793.3560383. URL https://doi.org/10.1145/3544793.3560383.

I. Vasilescu and M. Adda-Decker. Language, gender, speaking style and language proficiency as factors influencing the autonomous vocalic filler production in spontaneous speech. 09 2006.

C. Vorstius, R. Radach, and C. Lonigan. Eye movements in developing readers: A comparison of silent and oral sentence reading. *Visual Cognition*, 22:458–485, Jan. 2014. doi: 10.1080/13506285.2014.881445.

K. Watanabe. Sentence stress perception by Japanese students. *Journal of Phonetics*, 16(2):181–186, 1988. ISSN 0095-4470. doi: https://doi.org/10.1016/S0095-4470(19)30485-1. URL https://www.sciencedirect.com/science/article/pii/S0095447019304851.

S. Williams, P. Foulkes, and V. Hughes. Analysis of forced aligner performance on L2 English speech. *Speech Communication*, 158:103042, Mar. 2024. ISSN 0167-6393. doi: 10.1016/j.specom.2024.103042. URL https://www.sciencedirect.com/science/article/pii/S0167639324000141.

X.-Y. Wu. Artificial Intelligence in L2 learning: A meta-analysis of contextual, instructional, and social-emotional moderators. *System*, 126:103498, Nov. 2024. ISSN 0346-251X. doi: 10.1016/j.system.2024.103498. URL https://www.sciencedirect.com/science/article/pii/S0346251X2400280X.

S.-Y. Yoon, M. Hasegawa-Johnson, and R. Sproat. Automated pronunciation scoring using confidence scoring and landmark-based svm. In *Tenth Annual Conference of the International Speech Communication Association*, 2009.

K. Zechner, D. Higgins, X. Xi, and D. M. Williamson. Automatic scoring of non-native spontaneous speech in tests of spoken english. *Speech Communication*, 51(10):883–895, 2009. ISSN 0167-6393. doi: https://doi.org/10.1016/j.specom.2009.04.009. URL https://www.sciencedirect.com/science/article/pii/S0167639309000703. Spoken Language Technology for Education.

F. Zimmerer, J. Jügler, B. Andreeva, B. Möbius, and J. Trouvain. Too cautious to vary more? a comparison of pitch variation in native and non-native productions of french and german speakers. 05 2014. doi: 10.21437/SpeechProsody.2014-196.

Titre: Vers l'évaluation multimodale du niveau L2: caractéristiques de la parole et du mouvement des yeux dans un cadre multimodal

Mots clés: évaluation automatique de la L2, multimodalité, parole, mouvement des yeux

Résumé: Ces dernières années, le monde de l'éducation a subi des changements importants, notamment avec la digitalisation massive du système en 2020 et l'avancement des technologies d'IA générative. Le projet LeCycl "Learning Cyclotron" (Vargo et al., 2023) s'inscrit dans cette tendance scientifique et vise à accélérer la circulation des connaissances; en prenant en compte trois processus principaux d'apprentissage : la perception, la maîtrise et le transfert. Cette thèse, faisant partie du projet LeCycl, exploite des stratégies de lecture orale en langue étrangère (L2) en analysant les difficultés rencontrées par des représentants de différentes cultures et leurs stratégies pour les surmonter. À cette fin, nous utilisons des indices multimodaux (la parole et le mouvement des yeux) ainsi qu'un protocole original introduisant des nudges (représentés par des bandes dessinées) pour l'adaptation culturelle (Hutin et al., 2023), Nous avons développé un protocole impliquant la collecte de données de lecture à voix haute des locuteurs français et japonais de l'anglais L2 et des locuteurs natifs de l'anglais (Kobylyanskaya, 2022). Nous avons analysé les performances des locuteurs à travers des mesures acoustiques et linguistiques (réalisation acoustique des phonèmes, prosodie et disfluences telles que les pauses, hésitations, troncations), ainsi que des mesures de mouvements oculaires (El Baha et al., 2022; Kobylyanskaya et al., 2023). Ensuite, nous avons utilisé des méthodes d'apprentissage automatique pour définir le niveau de L2 des locuteurs à partir de ces mesures. Enfin, nous avons évalué la contribution des images sur les performances de lecture orale des locuteurs. Les résultats montrent que les représentants de langues maternelles Français et Japonais sont confrontés à des défis spécifiques lorsqu'ils lisent des textes en langue étrangère et adoptent différentes stratégies pour les surmonter, ce qui se traduit à la fois au niveau verbal et oculaire. Nos résultats soulignent la nécessité de développer des outils d'apprentissage adaptés aux cultures ainsi que les défis associés à leur conception.

Title: Towards multimodal assessment of L2 level: speech and eye tracking features in a cross-cultural setting

Keywords: L2 assessment, multimodality, speech, eye-tracking

Abstract: In recent years, the world of education has undergone critical changes, especially with the system's massive digitalization in 2020, as well as the advancement of generative AI technologies. LeCycl "Learning Cyclotron" (Vargo et al., 2023) project is a part of this scientific trend and its aim is to accelerate the knowledge flow. It takes into consideration 3 main processes of learning: perception, mastering and transfer. This thesis, as part of the LeCycl project, focuses on exploring second language (L2) oral reading strategies and analyzing difficulties faced by representatives of different cultures and their techniques of coping with them. For this purpose, we are relying on multimodal cues including speech and eye tracking, as well as an original protocol that introduces nudges (represented by comic books) for the cultural adaptation (Hutin et al., 2023). For this purpose, we developed a protocol involving the collection of reading aloud data from both native and non-native English speakers (French and Japanese speakers) (Kobylyanskaya, 2022). We analyzed speakers' performance through acoustic and linguistic measures (phoneme realization, prosody and disfluencies such as pauses, hesitations, truncations), as well as eye movement measures (El Baha et al. 2022; Kobylyanskaya et al., 2023). Then, we used machine learning methods to define the speaker's L2 level based on the extracted measures. Finally, we evaluate the contribution of comic books images on speakers' oral reading performance. The results highlight that the representatives of different cultures face different challenges when reading in a foreign language and employ different strategies to overcome them, which are translated both at verbal and ocular levels. Our results underline the need for culturally adapted learning tools and the challenges involved in developing them.