Speech and eye tracking features for L2 acquisition: a multimodal experiment

Sofiya Kobylyanskaya

Speech and eye tracking features for L2 acquisition: a multimodal experiment

Sofiya Kobylyanskaya[1,2]

[1] LISN-CNRS, France
[2] Paris-Saclay University, France

Abstract. Spoken language variation analysis is increasingly considered in multimodal settings combining knowledge from computer, human and social sciences. This work focuses on second language (L2) acquisition via the study of linguistic variation combined with eye-tracking measures. Its goal is to model L2 pronunciation, to understand and to predict through AI techniques the related metacognitive information concerning reading strategies, text comprehension and L2 level. We present an experimental protocol involving a reading aloud setup, as well as first data collection to gather L2 speech with associated eye-tracking measures.

Keywords: L2 acquisition · Speech variation · Eye-tracking · Multimodality · Education

1 Introduction

"LeCycle" is a trilateral project (France, Japan, Germany) aiming to improve knowledge transfer in various domains of education. As part of this project, this PhD work focuses on second language (L2) acquisition and evaluation using a multimodal approach that combines eye tracking and speech. These metrics will be integrated into an AI-based system to provide a comprehensive analysis of speakers' reading strategies and to predict their challenges in L2 text processing and pronunciation during a reading aloud setup. Additionally, we aim to find reliable influential strategies (nudges) permitting to reinforce speakers' L2 skills by improving their learning behavior. This paper presents the state of the art on the combination of eye-tracking, speech and nudges applied to L2 learning, the experimental protocol and the platform implemented to obtain the first dataset from 40 participants. Finally, we summarize further research directions and challenges.

2 State of the art

Multimodal teaching methods Nowadays the application of CALL (computer assisted language learning) is widely spread as it can be beneficial at several levels, *e.g.* it can stimulate the discussion among students [6], facilitate access to learning material, allow more flexibility in terms of study place

and rhythm, provide instant feedback about the student's performance. CALL systems rely on a variety of automatic measures and AI techniques such as: facial recognition to identify the student's emotional state, attention and comprehension level [15], body temperature recognition for attention and emotion recognition [3], eye-tracking analysis for L2 level prediction [2], speech recognition to estimate pronunciation errors, etc. However, CALL systems have some disadvantages, such as for instance the lack of personalization and the poor error recognition accuracy [6]. By combining eye-tracking and speech measures, this work aims to contribute to the improvement of CALL systems.

L2 pronunciation and speech-based metrics Previous work shows the interest of measuring speech features to assess the level of L2 mastering. For example, the verbal level can reflect the specificities of L2 pronunciation due to the speaker's L1 [8]. Hence, we can analyze realizations such as the voicelessness of consonants, the duration and the vowels' formants [19], etc. As for the paraverbal level, it can also reveal details about the level of comprehension, engagement, stress and other metacognitive states [26]. At this level, we can consider disfluencies such as pauses, hesitations and latencies that help the speaker to guide the interaction process [29]. They can also provide relevant information about L1-vs-L2 text processing strategies while reading aloud [14].

Speech and eye-tracking for L2 teaching and evaluation Eye-tracking information can complement speech features. For example, [22] shows a correlation between eye movement and accented syllables in speech perception. Studies on object naming also highlight the correlation between speech planning and eye movement [13] and the correlation between word length and time spent on the acquisition of its phonological form [13]. According to [21], about one third of the words are skipped during silent reading, especially function words (usually shorter) that occur more frequently and are more predictable than content words [24]. Rare words require more time to be processed than frequent ones, therefore fixations on them are longer [23] [24].

Eye-tracking in education and L2 learning Combining eye-tracking with machine learning can be used to understand students' mental state and motivation and aid in improving their learning achievements. For example, eye-tracking data can be used to classify emotional valence [16], predict co-occurring emotions [17], detect confusion [25] and predict educational goals while interacting with a pedagogical agent [16]. Eye-tracking can also be used in language learning to detect the language proficiency level [4] [2] and to understand the mechanism of syntactic processing when reading in L2 [9]. To our knowledge, most studies on eye-tracking in L2 learning were conducted in a silent reading experimental setup. The present data represent a first attempt to combine spoken and eye movement information, which can be a promising direction as it permits to capture both conscious and unconscious processes [10].

[3] https://www.techlearning.com/buying-guides/best-thermal-imaging-cameras-for-schools

Nudges in education During the education process, it is crucial not only to understand learner's strategies and L2 acquisition challenges, but also to contribute to their facilitation. One possible solution may be the use of nudges. The term nudge, coming from economy theory, is defined as an influential tactic that modifies consumer's behavior in a discrete and indirect manner relying on their affective system [28]. It can also be used in the education sphere, but according to [27] only 4% of nudges are related to education. For example, social comparison nudges can contribute to grades' increase [3] [11], those using extrinsic information such as rewards are efficient for younger children [18] [12], and nudges relying on deadlines can improve self-discipline [30]. These strategies can also be found in L2 acquisition and pronunciation remediation, *e.g.* using facilitating contexts [5]. One of our goals is to highlight difficulties in L2 speaking and pronunciation and to apply appropriate nudging strategies to facilitate phonetics and phonology acquisition.

3 Experimental protocol and first results

We collected speech and eye-tracking data from 40 French native speakers. We used "Eye Got it" [7], a platform developed for the project that permits to record both eye-tracking and audio, while associating a forced aligner for speech.

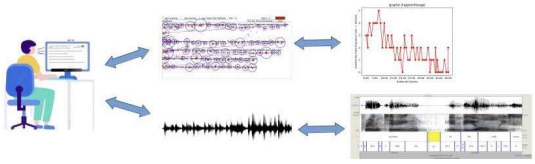

Fig. 1. Experiment process: eye+speech recordings followed by eye-voice span calculation and forced alignment of speech+transcription

Data was recorded in natural indoor conditions in a silent room. We used Tobii Nano Pro for eye tracking and the microphone AKG Perception Wireless 45 Sports Set Band-A 500-865 MHz for speech (Fig. 1). Total duration per subject is around 30 min.

In the following sections we describe the experimental setup from preparing the volunteer to recording their post-experimental feedback.

Pre-experiment All the participants are French native speakers, mostly students at different Parisian universities (>18 y.o.) and have at least a beginner level of English. They were asked to complete a survey concerning their linguistic background and vision issues (e.g. glasses). They were also invited to sign a consent form about the use of (anonymized) personal data.

Four texts are proposed, one in L1 French and 3 in L2 English: beginner, intermediate and advanced levels. The L1 text serves as a baseline for native pronunciation features. It contains declarative and interrogative sentences and

is 456 words long. As for L2, texts were selected from the website “English For Everyone” devoted to English learning. Texts are written by professional English teachers, are adapted to different L2 levels and include multiple choice questions for text comprehension. The following criteria were taken into account for text selection: levels from mid-beginner to mid-advanced; number of words; format ("short stories"); types of sentences and readability measures.

As in [20], we computed lexical and syntactic complexity using [1] for L2 texts. A correlation between text level and lexical complexity is observed, as well as a relation between some texts and their syntactic complexity.

Prior to recording, participants are familiarized with the equipment and the eye-tracker is calibrated with "Eye Got it". The volunteer sits at 60 cm from the screen and is encouraged to maintain this position during the reading phase to avoid recalibration.

Experiment: reading aloud All the participants read the same texts at their natural pace and volume and are free to use disfluencies. However, they are not allowed to look through the texts before the recording, in order to avoid pre-familiarization with potential unknown words, unexpected syntactic structures or any other lexical combinations in L2.

Post-experiment: pronunciation/comprehension feedback After reading each text, participants are asked to choose the words that were difficult to pronounce and/or to understand. The aim is to detect potential causes of non-canonical pronunciations and/or to correlate challenging words with disfluencies. Then, participants are invited to answer multiple choice questions about each text in L2. This task is aimed to combine the text comprehension level with the information provided in the survey in order to define the actual L2 level of the participants as labels for our future classification system. Note that the voice and the eye movement are not recorded during the post-experiment phase and the participants can take the time needed for the tasks.

Results of first-step data collection During the first stage of the experiments in February 2022, we collected data from 40 participants. Although we plan to extend the procedure to various socio-professional groups, the current volunteers are mainly academics and other staff members from different Parisian institutions.

Ages range from 18 to 35 and the participants have at least B1 level (according to their personal evaluation or to the score obtained at tests of English as foreign language). More than half of them wear glasses and have some vision problems. All the participants are native French speakers, around 14% of them are bilingual and >60% have exposure to other languages. Most of them (>60%) started learning English at the age of 6 y.o.-10 y.o.and around 30% of them have lived in an English-speaking country for at least several weeks.

4 Conclusion and further research

This paper focuses on an ongoing PhD work in the framework of the project "LeCycl". An experimental protocol has been built to gather eye-tracking and speech recordings for L2 acquisition have been described and here we describe the first results. Following work will focus on the contribution of the two modalities: machine learning algorithms will be applied to model L2 pronunciation, and nudging strategies will be added in order to facilitate L2 pronunciation acquisition. This innovative project involves many practical challenges, e.g. from eye-tracker calibration to L2 forced alignment and combination with eye-tracking measures. Ultimately, the most important challenge will concern appropriate machine learning techniques to efficiently combine speech and eye-tracking features.

References

1. Ai, H., Lu, X.: A web-based system for automatic measurement of lexical complexity. In: ALICO-10. pp. 8–12 (June 2010)
2. Augereau, O., Fujiyoshi, H., Kise, K.: Towards an automated estimation of english skill via toeic score based on reading analysis. In: ICPR'16. pp. 1285–1290 (2016)
3. Azmat, G., Iriberri, N.: 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)
4. Berzak, Y., Katz, B., Levy, R.: Assessing language proficiency from eye movements in reading. In: NAACL'18: HLT. vol. 1, pp. 1986–1996. ACL, New Orleans, Louisiana (Jun 2018)
5. Billières, M.: Méthode verbo tonale : diagnostic des erreurs sur l'axe clair/sombre. https://www.verbotonale-phonetique.com/methode-verbo-tonale-diagnostic-erreurs-axe-clair-sombre/
6. Derakhshan, A., Salehi, D., Rahimzadeh, M.: Computer-assisted language learning (call): Pedagogical pros and cons. International Journal of English Language and Literature Studies **4**, 111–120 (09 2015)
7. El Baha, M., Augereau, O., Kobylyanskaya, S., Vasilescu, I., Laurence, D.: Eye got it: a system for automatic calculation of the eye-voice span. 15th IAPR DAS (2022)
8. Flege, J.: Second language speech learning: Theory, findings and problems, pp. 229–273 (01 1995)
9. Frenck-Mestre, C.: 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** (04 2005)
10. Godfroid, A., Winke, P., Conklin, K.: Exploring the depths of second language processing with eye tracking: An introduction. Second Language Research **36**(3), 243–255 (2020)
11. Goulas, S., Megalokonomou, R.: Knowing who you are: The effect of feedback information on short and long term outcomes. Economic rese, University of Warwick - Department of Economics (2015)
12. Guryan, J., Kim, J.S., Park, K.H.: Motivation and incentives in education: Evidence from a summer reading experiment. Economics of Education Review **55**, 1–20 (2016)

13. Huettig, F., Rommers, J., Meyer, A.: Using the visual world paradigm to study language processing: A review and critical evaluation. Acta psychologica **137**, 151–71 (06 2011)
14. Kang, S.: Exploring l2 english learners' articulatory problems using a read-aloud task (2020)
15. Kumar, M.: Advanced educational technology. Sankalp Publication (2020)
16. Lallé, S., Conati, C., Azevedo, R.: Prediction of student achievement goals and emotion valence during interaction with pedagogical agents. In: AAMAS'18. p. 1222–1231 (2018)
17. Lallé, S., Murali, R., Conati, C., Azevedo, R.: Predicting co-occurring emotions from eye-tracking and interaction data in metatutor. In: Roll, I., McNamara, D., Sosnovsky, S., Luckin, R., Dimitrova, V. (eds.) Artificial Intelligence in Education. pp. 241–254. Springer International Publishing, Cham (2021)
18. Levitt, S.D., List, J.A., Neckermann, S., Sadoff, S.: The behavioralist goes to school: Leveraging behavioral economics to improve educational performance. American Economic Journal: Economic Policy **8**(4), 183–219 (November 2016)
19. Boula de Mareuil, P., Vieru-Dimulescu, B., Woehrling, C., Adda-Decker, M.: Accents étrangers et régionaux en francais: Caractérisation et identification. Traitement automatique des langues **49**(3), 135–163 (2008)
20. Novikova, J., Balagopalan, A., Shkaruta, K., Rudzicz, F.: Lexical features are more vulnerable, syntactic features have more predictive power. CoRR **abs/1910.00065** (2019)
21. Rayner, K.: The 35th sir frederick bartlett lecture: Eye movements and attention in reading, scene perception, and visual search. Quarterly Journal of Experimental Psychology **62**(8), 1457–1506 (2009)
22. Reinisch, E., Jesse, A., M. McQueen, J.: Early use of phonetic information in spoken word recognition: Lexical stress drives eye movements immediately. Quarterly Journal of Experimental Psychology **63**(4), 772–783 (2010)
23. Roberts, L., Siyanova-Chanturia, A.: Using eye-tracking to investigate topics in l2 acquisition and l2 processing. Studies in Second Language Acquisition **35** (06 2013)
24. Schotter, E., Fennell, A.: Readers can identify the meanings of words without looking at them: Evidence from regressive eye movements. Psychonomic Bulletin & Review **26** (09 2019)
25. Sims, S.D., Conati, C.: A Neural Architecture for Detecting User Confusion in Eye-Tracking Data, p. 15–23. Association for Computing Machinery, New York, NY, USA (2020)
26. Stolcke, A., Shriberg, E., Bates, R., Ostendorf, M., Hakkani-Tur, D., Plauche, M., Tur, G., Lu, Y.: Automatic detection of sentence boundaries and disfluencies based on recognized words. (01 1998)
27. Szaszi, B., Palinkas, A., Palfi, B., Szollosi, A., Aczel, B.: A systematic scoping review of the choice architecture movement: Toward understanding when and why nudges work. Journal of Behavioral Decision Making **31**(3), 355–366 (2018)
28. Thaler, R., Sunstein, C.: Nudge: Improving decisions about health, wealth, and happiness. Yale University Press (2008)
29. Vasilescu, I., Adda-Decker, M.: Language, gender, speaking style and language proficiency as factors influencing the autonomous vocalic filler production in spontaneous speech (09 2006)
30. Weijers, R.J., de Koning, B.B., Paas, F.: Nudging in education: From theory towards guidelines for successful implementation. European Journal of Psychology of Education **36**(3), 883–902 (2021)