E-002

A Design of Multimodal Learner Corpus for Practical Foreign-language Communication

Katsunori Kotani† Takehiko Yoshimi‡

1. Introduction

In this era of social globalization, there has been a growing demand for foreign language learners to act as mediators between those who do not share a common language (Council of Europe 2010). As a response to this requirement, the goal of foreign language learning should include the mediation as community interpreters (De Florio-Hansen 2013). Community interpreting is a bidirectional, face-to-face interpreting between local and nonlocal language speakers in a community (Hale 2011). This type of nonprofessional interpreting is collecting the attention owing to the social globalization, as it can facilitate the achievement of linguistic diversity for the protection of language rights, especially in areas where linguistic diversity has not yet been achieved, such as some parts of Asia.

Given this background of foreign language teaching, learner corpus research has targeted practical communication domains such as legal (Franceschi 2018) and tourism communication (Bonsignori & Cappelli 2022). Previous learner corpus research (Franceschi 2018, Bonsignori & Cappelli 2022) has utilized multimodal learner corpora to analyze learners' verbal and nonverbal behaviors in the practical communication settings. In the face-to-face interactive communication settings, learners are required to demonstrate appropriate verbal and nonverbal behaviors, as "we speak with our vocal organs, but we converse with our entire bodies" (Abercrombie 1968).

Recent learner corpus research using a multimodal learner corpus has explored the potentials of analysis of communication in a foreign/second language from the perspectives of learners' verbal and nonverbal behaviors (Franceschi 2018, Bonsignori & Cappelli 2022, Ishikawa 2022). A multimodal learner corpus utilizes video data for analyzing learners' verbal and nonverbal behaviors with linguistic and gestural properties. As research around second language acquisition has demonstrated the significant role of learner corpora for investigating the learning process or developing evaluation tools, a multimodal learner corpus is expected to provide further conceptual and empirical contribution. The previous studies demonstrated the usefulness of a multimodal learner corpus for analyzing learners' practical communication such as tourism or legal discourse. The multimodal learner corpus can be taken as the third generation of a learner corpus, following the first generation of learner corpus focusing on analyzing learners' writing skills through text data, and the second generation of a learner corpus incorporating speech-sound data for analyzing learners' speaking skills. As the learner corpus data have evolved from text to video, the focus of research interesting has shifted towards the multimodal analysis of learners' practical communication.

The present study aims to create a multimodal learner corpus, which can be used to investigate both verbal and nonverbal interpreting skills within interpreting learner corpus research. Additionally, this corpus can also be used to study interpreting as a target language use within the context of multimodal learner corpus research. Thus, this paper presents a design of a multimodal learner corpus as a resource for foreign language learning/teaching. This study also seeks to explore the usefulness of incorporating verbal and nonverbal data in an interpreting learner corpus for analyzing the language learning process, particularly, the significant effects of nonverbal behaviors to communication in a foreign language.

2. Previous multimodal learner corpus

Franceschi (2018) provided insights into the advantage of a multimodal learner corpus as a resource for legal English education. The multimodal learner corpus (Franceschi 2018) consisted of 4,599 tokens in total on 17 videoclips collected from 34 university students whose proficiency levels ranged between the B1 and B2 levels in the Common European Framework of Reference (CEFR) (Council of Europe 2001). The learners conducted lawyer-client interviews in pairs. The legal discourses were manually transcribed and annotated for verbal and nonverbal behaviors. The verbal annotation information captured the use of legal terms and the error information on vocabulary, grammar, and pronunciation. The nonverbal annotation information described the types and functions of gestures. The transcription and annotation tiers were accompanied with sill-frame images of nonverbal cues at the prominent moments judged by annotators. Figure 1 presents scenes where a lawyer-part student spoke to a client-part student, using metaphorically iconic hand movements which replicated the semantic contents. The advantage of a multimodal analysis can be seen in Figure 2 in which the multimodal analysis appropriately identified the communication skill of a lawyer-part student who succeeded in constructing a rapport

† Kansai Gaidai University
‡ Ryukoku University

relation with a client-part student by using nonverbal cues without verbal communication.

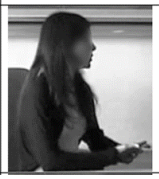
	Image frame	Verbal text	Non-verbal behaviour & interpretation
1		L: *Well, I'm here to* ***help you****.*	Intersecting hands and fingers with palms up and moving them towards client, as if pretending to give something to her, showing willingness to help

Figure 1: Illustrative data (i) (Franceschi 2018), where the author of this paper obscured the face.

	Image frame	Verbal text	Non-verbal behaviour & interpretation
4		C: *At the first I threatened to write as many mailing lists as possible to tell people not to buy computers from Carmecom, but they didn't take me seriously. So now I want a sort of revenge. I decided to hurt Carmecom […] its sellers to take me back my money.*	Looking at the client, holding hands together with fingers intertwined, showing interest/willingness to listen

Figure 2: illustrative data (ii) (Franceschi 2018), where the author of this paper obscured the face.

Bonsignori & Cappelli (2022) presented the importance of using multimodal corpus data for tourism English education. The multimodal learner corpus (Bonsignori & Cappelli 2022) consisted of 20,897 tokens in total for 2 hours and 58 minutes on 98 videoclips compiled from 49 university students at the B1 CEFR level. Each student engaged in two tourism discourses. The corpus had annotation tiers for verbal and nonverbal behaviors in L2 learners' tourism discourse. The verbal annotation tiers described lexical, syntactic, semantic, and phonological features, while the nonverbal annotation tiers explained the types and functions of gestures. As the goal of a multimodal analysis was to capture verbal and nonverbal behaviors comprehensively, this corpus represented the speech sounds fully with sound waves and orthographic transcriptions with an annotation tool, ELAN (Lausberg & Sloetjes 2009), as shown in Figure 3. However, nonverbal behaviors might have been unawarded at 00:00:06:965, although the learner was clasping the hands. This misidentification could be considered as a procedural problem regarding manual annotation due to unobtrusive observation (Kip et al. 2007). Here, the multimodal analysis should have identified the hand-clasping as an ineffective nonverbal cue. Although the learner started a discourse, the hand-clasping posture could be misinterpreted as a sign for the end of a proposition or topic sequence (Simmons-Mackie & Damico 2010).

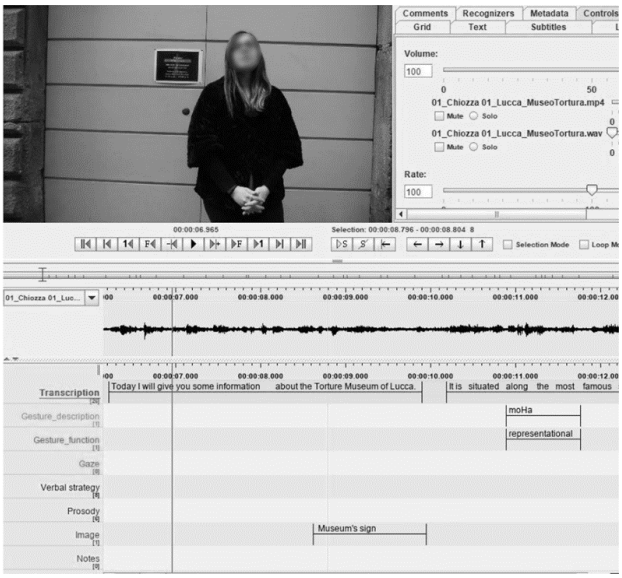


Figure 3: Illustrative data (Bonsignori & Cappelli 2022), where the author of this paper obscured the face.

Ishikawa (2022) demonstrated the potential benefits of incorporating gesture into language instruction through a study based on a multimodal learner corpus. The multimodal learner corpus (Ishikawa 2022) consisted of 1.6 million tokens, comprising 4,250 video clips of interview tests including conversations and picture descriptions. The corpus data were compiled from 405 learners from A1 to B2 CEFR levels. The corpus was annotated for the frequency of head and hand gestures (0; 1; 2; 3<), with head gestures including tilting or lifting the head, and hand gestures including touching the head, moving the hand, and pointing to the picture, as shown in Figure 4.

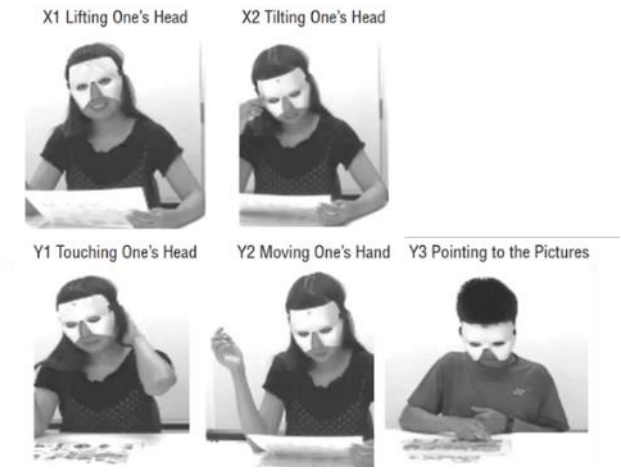


Figure 4: Head/hand movement (Ishikawa 2022).

3. Multimodal learner corpus

This study aims to develop a multimodal learner corpus as a resource for foreign language learning/teaching researchers to analyze learners' verbal and nonverbal skills in practical communication. Therefore, this study presents the design of a multimodal learner corpus in Section 3.1, the method for compiling data in Section 3.2, the annotation method in Section 3.3, and a case study on the analysis of nonverbal behaviors that were parsed with a gesture recognition tool in Section 3.4.

3.1 Corpus design

As a multimodal learner corpus serves as a resource for exploring the learning process and developing evaluation tools, multimodal learner corpus data have the potential to expand the empirical and conceptual scope by examining both verbal and nonverbal behaviors in learners' language use. The analysis of verbal behaviors can reveal aspects such as pronunciation, vocabulary usage, and grammar, whereas the analysis of nonverbal behaviors can uncover learners' gestures, facial expressions, and postures. Previous studies on a multimodal learner corpus (Franceschi 2018, Bonsignori & Cappelli 2022, Ishikawa 2022) found pedagogical advantages because the corpus data could provide insights into learners' communication strategies and enable learners to become more aware of the learning process.

Additionally, a multimodal learner corpus can provide significant features in evaluating proficiency levels (Ishikawa 2022). Parsing results of nonverbal behaviors can comprehensively represent learners' nonverbal behaviors in terms of gesture phases such as preparation, stroke, and hold phases (Bressem 2011) by extracting distinctive features such as ± movement, ± constant, and ± increase. The parsing results can further offer a solution for the privacy concerns in a multimodal corpus, which anonymizes learners' face but retains facial features (Guichon 2017).

To meet the expectations outlined above, a multimodal learner corpus should satisfy the following conditions. First, the corpus should include data on learner communication in a foreign language (Condition I for learner). Second, the corpus should include learners' bidirectional interpreting, allowing for analysis of their translation, listening, and speaking skills (Condition II for interpreting). Third, the corpus should comprise video data, enabling researchers to analyze both verbal and nonverbal communication skills (Condition III for multimodality). Finally, the corpus should track learners' verbal and nonverbal behaviors during interpreting to facilitate retrieval of annotation information (Condition IV for annotation).

The current learner corpora partially met conditions I-IV. The previous multimodal learner corpora (Franceschi 2018, Bonsignori & Cappelli 2022, and Ishikawa 2022) captured learners' verbal and nonverbal behaviors and satisfied the conditions I and III. However, these multimodal learner corpora failed to meet Condition II and IV. These corpora did not include bidirectional interpreting or track nonverbal behaviors throughout the discourse process because of the manual identification of nonverbal behaviors.

3.2 Data collection method

This study targets college students who are learning English as a foreign language (henceforth, EFL) and are expected to achieve a certain level of proficiency in order to mediate between local and nonlocal language speakers in social communication, specifically in community interpreting. Our focus is on EFL learners whose first language is Japanese.

The EFL learners in this study carry out two types of tasks. The first task is a unidirectional community interpreting task in the form of sight translation and consecutive interpreting. The goal of this task is to determine a learner's baseline for the nonverbal and verbal behaviors. In the sight-translation task, learners verbally translate linguistic landscapes such as public and commercial signs, while in the consecutive interpreting task, they interpret soundscapes such as public-place announcements. The second type of task involves bidirectional community interpreting for communication between English and Japanese language speakers. In this task, EFL learners listen to either an English or Japanese language speaker and interpret into the other language.

EFL learners complete interpreting tasks with an in-house data collection tool, which will be publicly available in the future. This data collection tool displays interpreting sources such as linguistic landscapes, soundscapes, and dialogues between Japanese and English language speakers. The tool also records EFL learners' interpreting on MP4 files as raw data for a multimodal interpreting learner corpus.

3.3 Data annotation method

The raw corpus data undergo the following preprocessing annotation steps. In the first step, speech sounds of the EFL learner are transcribed with a transcription tool such as Microsoft's Azure or Google's Speech-to-text. In the second step, nonverbal behaviors of an EFL learner are parsed with a gesture recognition tool such as Google's Mediapipe (Grishchenko & Bazarevsky 2020). In the final step, the retrieved verbal and nonverbal information is aligned with the video and speech sound with an annotation tool such as Max Planck's ELAN (Lausberg & Sloetjes 2009). Figure 5 presents sample annotation data on ELAN (Lausberg & Sloetjes 2009). The top tier displays the video where an EFL learner's face is superimposed with face meshes connecting the facial landmarks. The second tier represents the EFL learner's nonverbal behaviors with a spectrogram of the movements of body-part landmarks at the face, hands, and body. The third and fourth tiers exhibit the speech sound with a sound spectrogram and a sound wave, respectively. The annotation tiers show the land-mark positions. The annotation information also includes metadata such as an EFL learner's age, the English learning period, and English proficiency level as well as other relevant observations during the interpreting processes.

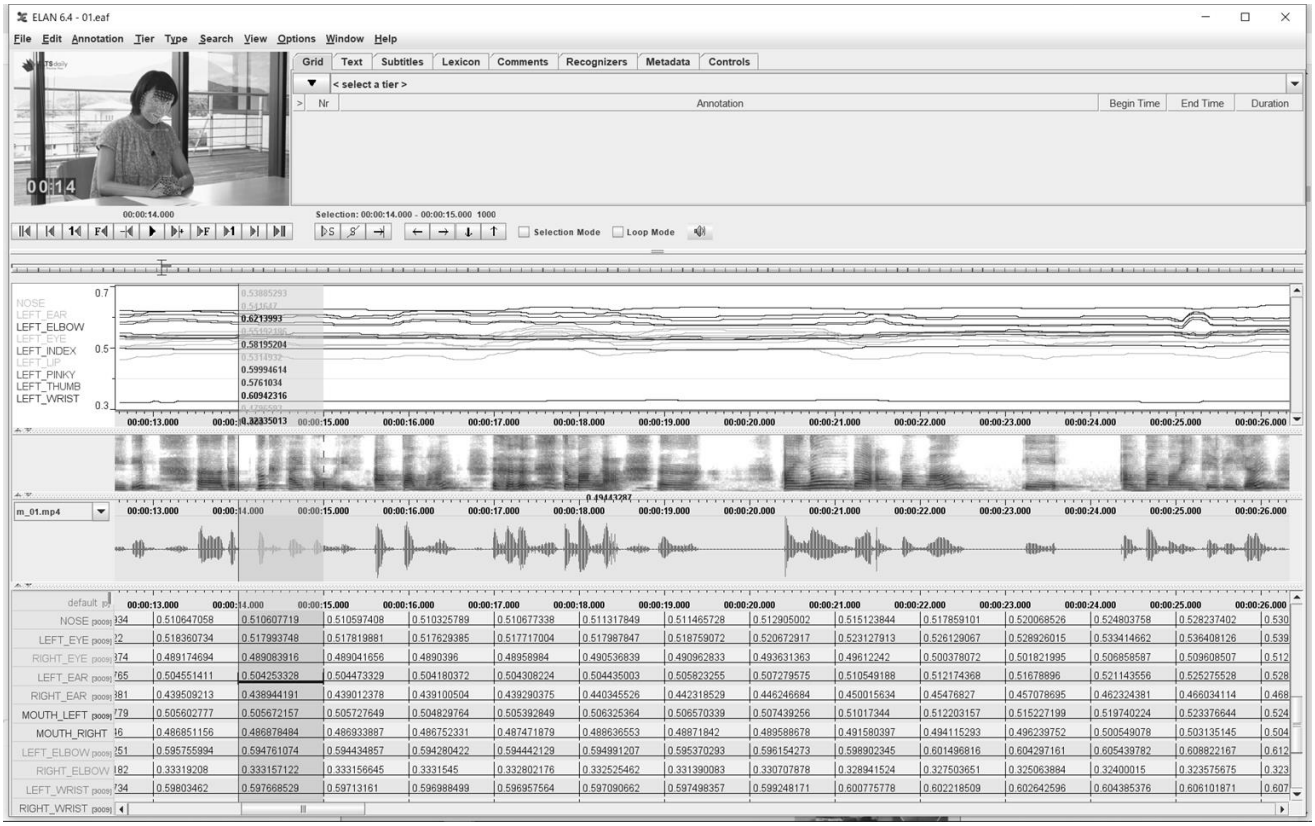


Figure 5: Sample data parsed for nonverbal behaviors ([Christopher G.]. (22, November 8). IELTS SPEAKING Mock Exam - Ami from Japan jp [Video]. IELTS Daily. https://www.youtube.com/watch?v=hvUa9vXE0bw).

The preprocessed corpus data undergo manual annotation steps, depending on the goals of the corpus analysis. For example, the nonverbal behavioral spectrogram at the second tier facilitates the annotation of the type and frequency of learners' facial expressions, gestures, and postures. Additionally, manual annotation for the nonverbal behaviors will identify gesture phases of preparation, stroke, and hold based on landmark positions. On the other hand, the speech-sound spectrogram and sound waves are used to annotate learners' pronunciation and the presence, absence, and duration of speech pauses.

3.4 Corpus evaluation

The validity of the multimodal interpreting learner corpus will be assessed based on the data collection method, annotation method, and data analysis. These three perspectives ensure the validity and reliability of the corpus for future research.

When developing a multimodal corpus, it is important to collect corpus data considering ethical issues such as obtaining consent from participants and protecting their privacy. One privacy protection method is anonymization of speakers' faces in a corpus. This study uses a face mesh technique to anonymize the speakers' faces by superimposing the facial landmark meshes onto the original facial images. This technique creates a new face image that obscures the speaker's face while still preserving the facial landmarks, as shown in Figures 5 and 6.

The annotation results will be assessed specifically for the validity of automatic preprocessing and the reliability of manual annotation processing. Automatic preprocessing will be evaluated the accuracy for the transcription process, the speech sound representation, and the gestural recognition. The manual annotation results will be evaluated for the inter-annotator reliability.

Last, the multimodal learner corpus will be evaluated for its contributions to the second language research. This corpus should provide evidence of the second language learning process for both verbal and nonverbal skills in practical communication. The corpus should also demonstrate linguistic and nonlinguistic properties that vary along with the learners' proficiency levels. These contributions facilitate to develop evaluation tools for learners' pronunciation, word/grammar usages, as well as their facial expressions, gestures, and postures.

3.5 Case study with parsed nonverbal behaviors

This study confirmed the validity of nonverbal behavioral information using preliminary corpus data. The goal of this preliminary experiment was to confirm whether the nonverbal behavioral spectrum could effectively reflect the gesture differences. Specifically, this experiment analyzed self-reference gestures of English speakers and Japanese speakers. Typically, speakers of Japanese use a nose-pointing gesture to indicate first-person reference. However, this behavior presents a cultural difference between the English and Japanese languages, as speakers of English may interpret the nose-pointing gesture as a sign of confidentiality or warning.

Figure 6 illustrates the two types of self-reference gestures at the uppermost layer, where the left side presents the self-reference gesture by English speakers, while the right side shows the one by Japanese speakers. The second layer shows the nonverbal behavioral spectrogram derived by Mediapipe (Grishchenko & Bazarevsky 2020)'s *y*-axis for the vertical positions of

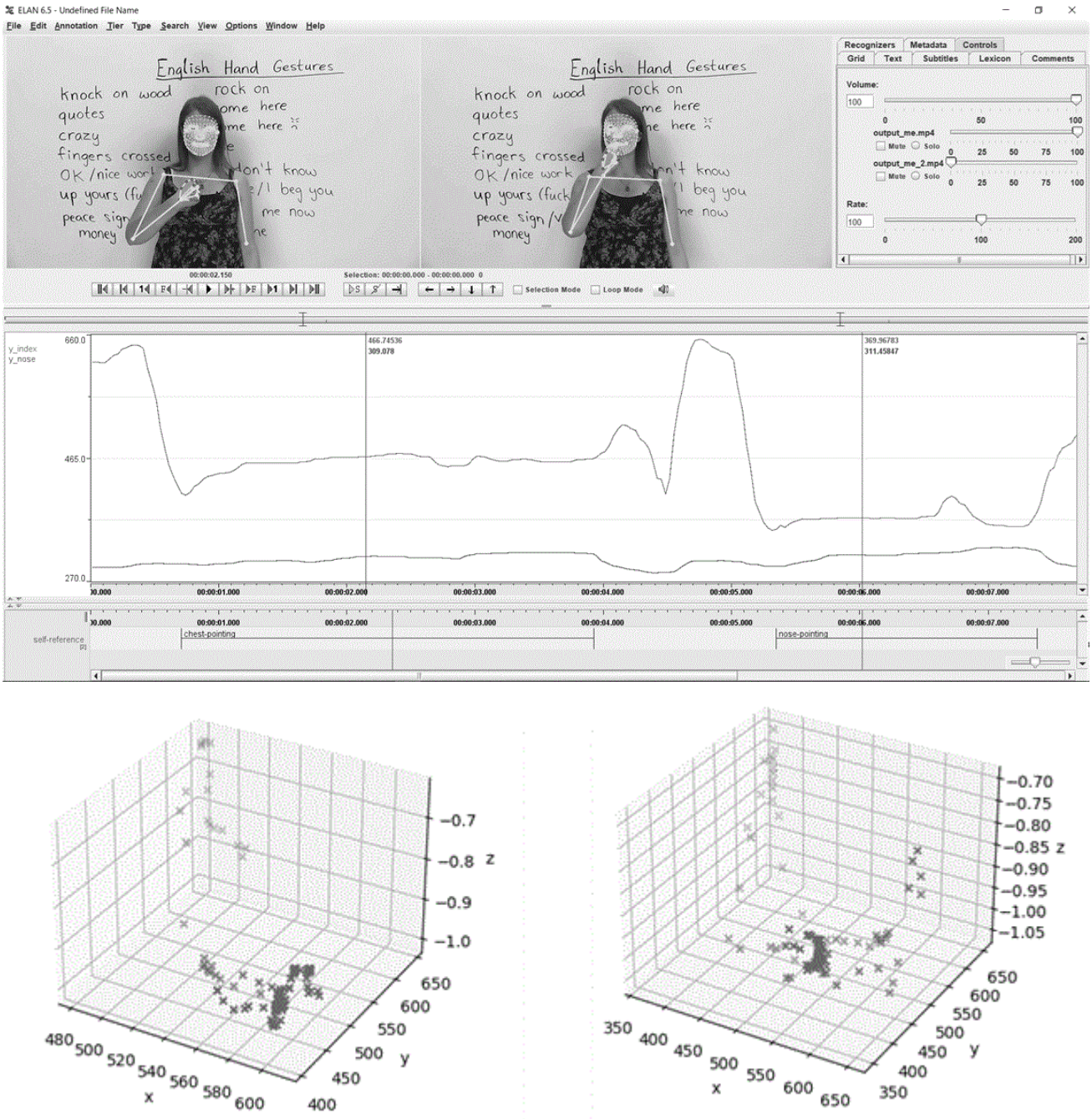


Figure 6: First-pronoun gestures in English (the left-side picture) and in Japanese (the right-side picture) ((22, November 8). English Conversation: The Meaning of Hand Gestures [Video]. English with Emma engVid. https://www.youtube.com/watch?v=xY_xiGadcgk).

an index finger and a nose in the English-type chest-pointing gesture (the left side) and the Japanese-type nose-pointing gesture (the right side). The spectrogram demonstrates the different distances between these two types of self-reference gestures. The distance between an index finger and a nose is greater in an English-type, chest-pointing gesture at the left side of the spectrogram than a Japanese-type, nose-pointing gesture at the right side. The lowest layer represents the three-dimensional scatter plots of *x*-, *y*-, and *z*-axes taken by Mediapipe (Grishchenko & Bazarevsky 2020), which shows the landmarks for a nose-pointing gesture appeared at higher positions than those for the chest-pointing gesture.

4. Conclusion

This study presented findings from the initial stages of development of a multimodal learner corpus. The corpus design considered learner characteristics, communication styles, corpus data, and annotations. We emphasized providing annotation information for both verbal and nonverbal behaviors throughout the discourse process because nonverbal behaviors were only partially identified owing to the limitation of manual annotation. We then introduced a nonverbal behavioral spectrogram obtained through gesture recognition. This spectrogram demonstrated its effectiveness in classifying gesture types. Moreover, the incorporation of gesture recognition facilitated facial expression analysis while anonymizing the speaker's face, which could be considered an additional benefit to multimodal corpus research. These results conceptually suggest that learners' nonverbal behaviors should be examined throughout the discourse process.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers, 22300299, 15H02940, 17K18679, and 21K18373.

References

[1] Abercrombie, D. 1968. Paralanguage. International Journal of Language & Communication Disorders, 3:55-59. https://doi.org/10.3109/13682826809011441

[2] Bonsignori, V. and Cappelli, G. 2022. Developing strategies for conceptual accessibility through multimodal literacy in the English for tourism classroom. Multimodal Communication, 11(1):31-47. https://doi.org/10.1515/mc-2020-0026

[3] Bressem, J. and Ladewig, H. S. 2011. Rethinking gesture phases: Articulatory features of gestural movement? Semiotica, 184:53-91. https://doi.org/10.1515/semi.2011.022

[4] De Florio-Hansen, I. 2013. Translation competence in foreign language learning: Can language methodology benefit from translation Studies? Journal of Linguistics and Language Teaching, 4(2): 39-68.

[5] Council of Europe. 2010. Guide for the Development and Implementation of Curricula for Plurilingual and Intercultural Education. Language Policy Unit. Strasbourg: Council of Europe.

[6] Franceschi, D. 2018. Simulating a lawyer-client interview: the case of Italian EFL University learners. ESP Across Cultures, 15:41-59.

[7] Grishchenko, I. and Bazarevsky, R. E. V. 2020. MediaPipe Holistic–Simultaneous Face, Hand and Pose Prediction, on Device. Retrieved March 9, 2023.

[8] Guichon, N. 2017. Sharing a multimodal corpus to study webcam-mediated language teaching. Language, Learning and Technology. 21: 56-75.

[9] Hale, S. 2011.Public service interpreting. In Kirsten Malmkj?r and Kevin Windle (eds.). The Oxford Handbook of Translation Studies. Oxford Academic, pages 343-356. https://doi.org/10.1093/oxfordhb/9780199239306.013.0024

[10] Ishikawa, S. 2022. L2 English learners' gesture use in a picture description task: A study based on the ICNALE Dialogue Module. In Raúl Ruiz Cecilia and Juan Ramón Guijarro Ojeda (eds.). Investigación e Innovación en Lengua Extranjera: Una Perspectiva Global, pages 721-750.

[11] Lausberg, H. and Sloetjes, H. 2009. Coding gestural behavior with the NEUROGES-ELAN system. Behavior Research Methods, 41(3):841-849. https://doi.org/10.3758/BRM.41.3.841.