Project MOSLA: Recording Every Moment of Second Language Acquisition

Masato Hagiwara
Octanove Labs & Earth Species Project
Seattle, WA, USA
masato@octanove.com

Joshua Tanner
Mantra Inc.
Tokyo, Japan
josh@mantra.co.jp

Abstract

Second language acquisition (SLA) is a complex and dynamic process. Many SLA studies that have attempted to record and analyze this process have typically focused on a single modality (e.g., textual output of learners), covered only a short period of time, and/or lacked control (e.g., failed to capture every aspect of the learning process). In Project MOSLA (Moments of Second Language Acquisition), we have created a longitudinal, multimodal, multilingual, and controlled dataset by inviting participants to learn one of three target languages (Arabic, Spanish, and Chinese) from scratch over a span of two years, exclusively through online instruction, and recording every lesson using Zoom. The dataset is semi-automatically annotated with speaker/language IDs and transcripts by both human annotators and fine-tuned state-of-the-art speech models. Our experiments reveal linguistic insights into learners' proficiency development over time, as well as the potential for automatically detecting the areas of focus on the screen purely from the unannotated multimodal data. Our dataset is freely available for research purposes and can serve as a valuable resource for a wide range of applications, including but not limited to SLA, proficiency assessment, language and speech processing, pedagogy, and multimodal learning analytics.

Keywords: second language acquisition, multimodal learning analytics, speech processing

1. Introduction

The acquisition of a second language is a complex and dynamic process characterized by various milestones and challenges that learners encounter along their journey. Many studies have attempted to record the learning process, although most studies are unimodal (e.g., capturing only the textual output of learners, Geertzen et al., 2014), cover only a short period (e.g., containing snapshots of learner's progress, Settles et al., 2018), and/or are limited in control (e.g., not capturing every aspect of the learning process, Stasaski et al., 2020). It has long been recognized that multimodal, longitudinal interaction is a crucial factor in SLA (Hampel and Stickler, 2012).

In order to shed light on the complex and dynamic nature of the SLA process, in Project MOSLA (Moments of Second Language Acquisition), we created a longitudinal, multimodal, multilingual, and controlled dataset by inviting participants to learn a new language from scratch solely through online instruction over a span of two years and documenting every lesson using Zoom. This dataset, comprising over 250 hours of recorded lessons, captures the rich and nuanced aspects of language learning, including verbal and non-verbal communication, the use of teaching materials, student-teacher interactions, and the evolving proficiency of learners. Notably, the MOSLA dataset encompasses a diverse set of target languages—Arabic, Spanish, and Chinese—including two languages that employ non-Latin alphabets, highlighting the dataset's unique cross-linguistic scope.

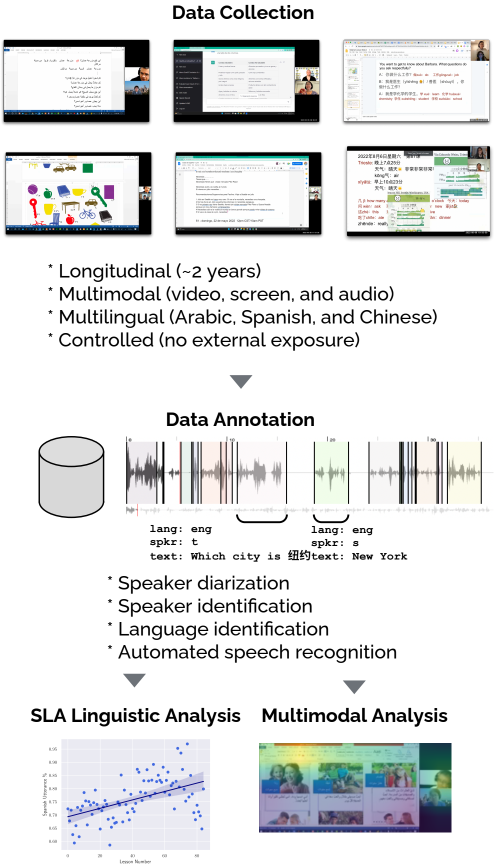


Figure 1: Overview of Project MOSLA

To enhance the dataset's utility, we semi-automatically annotated all the utterances in the

LREC-COLING 2024, pages 13092–13102
20-25 May, 2024.

recorded audio with start and end offsets, speaker and language IDs, and transcripts. This annotation was accomplished using human annotators and state-of-the-art machine learning models for speaker diarization, speaker and language identification, and automatic speech recognition. The resulting metadata offers valuable insights into the distribution of speech and speaker identities throughout the learning process, as well as transcriptions of spoken content.

In this paper, we provide an overview of the creation, annotation, analysis, and applications of the MOSLA dataset. We begin by describing our data collection method and then discuss the process of human and machine annotation. We empirically demonstrate that fine-tuning state-of-the-art speech models with a small amount of human-annotated data results in substantial improvements in speaker and language identification, as well as speech recognition performance. Additionally, we show that our data can reveal linguistic insights into the learners' acquisition process of the target language, such as the percentage of non-English utterances and lexical diversity. Furthermore, we demonstrate that, through the use of deep neural network models, we can determine where on the screen the teacher and the learner are focusing, solely from the unannotated multimodal video and audio data. The MOSLA dataset represents a significant contribution to the field of SLA research, providing a rich source of data for investigating the factors influencing language learning outcomes, the role of multimodal cues in the acquisition process, and the development of innovative educational tools.

The MOSLA dataset is freely available for research and non-commercial purposes, ensuring that it can benefit the broader academic community and contribute to advancements in the field of second language acquisition. It can be accessed here: https://www.octanove.com/mosla.html.

2. Related Work

In the field of SLA, there have been many studies that aimed to record and analyze the learning process, providing valuable insights into language learning. However, many of these studies have limited temporal coverage, typically spanning only several months (Vercellotti, 2015; Saito and Akiyama, 2017). Duolingo publishes the Second Language Acquisition Modeling (SLAM) dataset (Settles et al., 2018), which contains learner production in their target language. However, the data covers only a 30-day period, offering a relatively short-term perspective on language acquisition. The CIMA dataset (Stasaski et al., 2020) contains tutor-learner interaction data during language learning, but it lacks multimodal and longitudinal characteristics. Similarly, the Teacher-Student Chatroom Corpus (Caines et al., 2020) collected textual interactions between teachers and students during online English teaching but also lacks multimodal and longitudinal aspects.

In the realm of grammatical error correction (GEC), there is a substantial body of research (Bryant et al., 2023) but relatively few GEC corpora focus on longitudinal learning. One noteworthy exception is the EFCamDat corpus (Geertzen et al., 2014), one of the largest GEC corpora, with a collection period spanning a few years. However, only a few of its users participated over the entire duration, with many starting or ending their learning outside the collection period.

In other domains of learning analytics, Kubat et al. (2007) collected two years' worth of multimodal data on first language development through the Human Speechome Project (HSP) (Roy et al., 2006), primarily focusing on first language acquisition and involving data from a single individual. Demszky and Hill (2023) collected and analyzed transcripts of teacher-student discourse in elementary math classrooms.

MOSLA is closely related to the field of multimodal learning analytics (MMLA) (Mu et al., 2020) and web-based language learning (WBLL) (Cong-Lem, 2018). For example, Donnelly et al. (2017) analyzed classroom audio recordings, and Monkaresi et al. (2017) examined facial expressions as part of the learning analytics process.

The MOSLA dataset holds the potential to be valuable for various applications, including assessment (Settles et al., 2020), proficiency estimation (Vajjala and Rama, 2018), knowledge tracing (Piech et al., 2015), grammatical error correction (Bryant et al., 2023), automated assessment of speaking proficiency (Fan and Yan, 2020), and optimization of pedagogical approaches (Lepper and Woolverton, 2002), among others. Its longitudinal, multimodal nature makes it a unique resource for studying the complexities of the SLA process.

3. Data Collection

Data collection took place between February 2021 and February 2023. The teacher and the learner had weekly language instruction over zoom. Specifically,

- A learner (a complete beginner) and a teacher have a private lesson per week online (e.g., on Zoom) for at least two years.
- Every lesson is recorded (video, audio, and screen share).
- The learner is not allowed to learn the target language outside of these lessons.

	# Videos	Total duration
Arabic	95	102 hrs
Spanish	85	84 hrs
Chinese	84	84 hrs

Table 1: Raw Statistics of Collected Data by Course

- All the materials the learner is exposed to are recorded (e.g., via screen share).

All the learners in this study were already proficient in two or more languages (their L1 and L2, typically English) before the study started and are generally highly motivated individuals. Below is additional information on the individual courses:

- `ara`: Arabic (Modern Standard Arabic)
  - Teacher L1: Levantine Arabic
  - Learner L1: Japanese
  - Learner L2s: English, Mandarin Chinese
  - Learner Age: 35-44
  - Learner Gender: Male
- `spa`: Spanish (Latin American)
  - Teacher L1: Spanish (Latin American)
  - Learner L1: Mandarin Chinese
  - Learner L2s: English, Japanese
  - Learner Age: 35-44
  - Learner Gender: Female
- `zho`: Mandarin Chinese
  - Teacher L1: Mandarin Chinese
  - Learner L1: Spanish (Latin American)
  - Learner L2s: English, Italian, German
  - Learner Age: 25-34
  - Learner Gender: Female

All the teachers have a minimum of five years of professional experience teaching the target language. Additionally, all the participants are fluent in English, and the teaching instructions were conducted in English, at least initially. The study did not impose restrictions on the teaching methods employed by the instructors; they were free to use their preferred approaches. However, instructors were advised not to use copyrighted materials, such as textbooks and online courses, as-is, unless used in a supplementary capacity. As mentioned earlier, learners were prohibited from learning the target language outside of this study and were not assigned any explicit tasks beyond the classroom. Nevertheless, they were encouraged to review the recorded lesson videos for self-assessment purposes.

All the teaching was conducted via Zoom, and the video and audio were recorded using its standard recording functionality under the default settings. Participants used their own preferred devices for recording audio and video, which means that there was no quality control in regard to the devices.

4. Data Annotation

In addition to the video data for each lesson, MOSLA includes two sets of annotations containing information about the speech of the student and teacher: a smaller human-annotated set, and a complete machine-annotated set. Data from the human-annotated set is used to train machine learning models as shown in Figure 3, which generate a complete set of data for all lessons. We release all models trained this way for use in future research.

4.1. Human Annotation

We employ a bilingual annotator for each language pair, such that the annotator speaks both English and the language being learned. Annotation is done on five minute samples, which are selected as follows: we perform an independent random trial with a 5% chance to succeed for each possible sample, and keep up to one sample per lesson[1]. The first and last segment of each file are excluded from possible selection, as these often consist of technical setup or greetings instead of language education content.

We use Hachiue (Hayashibe, 2021) for annotation, as it provides an easy to use web interface which allows annotators to mark arbitrary sections of the file as utterances and attach data to them. Annotators were instructed to create segments for distinct utterances from each speaker which include a speaker label (teacher, student or other), a label for the dominant language of the utterance (there are a number of code-switched utterances containing multiple languages), and a literal transcription of the speech. An example of segment annotations is shown in Figure 2.

4.2. Machine Annotation

Using the human annotation data, we train and evaluate machine learning models to perform each task necessary for annotation: diarization, speaker and language classification, and automatic speech recognition. These models are then combined into a machine annotation pipeline (Figure 3) that we

[1] As an exception to this, a small number of Chinese lessons are slightly over-annotated.

Language	Source	Total Duration	Utterance Duration	Utterance Count	Target Language	Student Utterance
Arabic	Human	3.0 hrs	2.6 hrs	2,330	82%	50%
	Machine	101.5 hrs	73.9 hrs	80,441	81%	57%
Spanish	Human	2.5 hrs	2.2 hrs	1,006	85%	52%
	Machine	83.7 hrs	61.1 hrs	62,980	82%	50%
Chinese	Human	4.0 hrs	3.3 hrs	4,375	66%	24%
	Machine	84.4 hrs	65.6 hrs	58,917	72%	33%

Table 2: Annotation Data

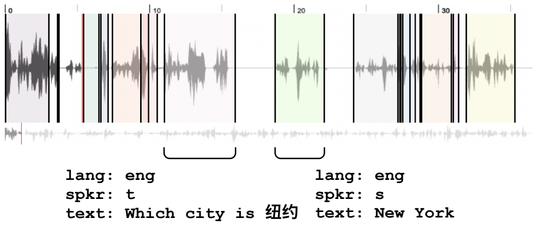


Figure 2: Example Annotation

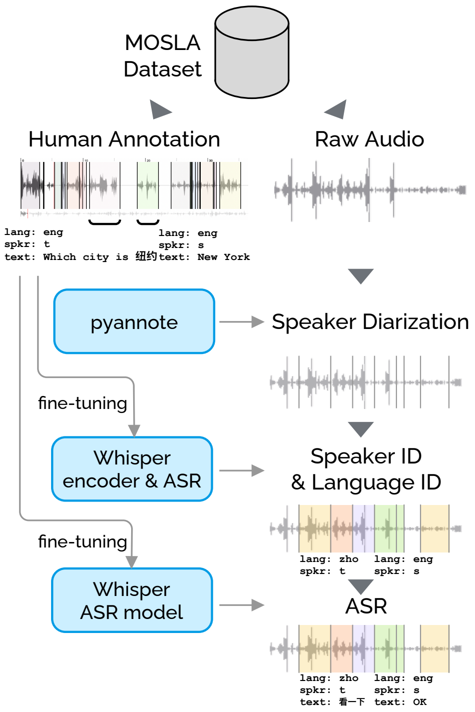


Figure 3: Overview of the annotation pipeline

use to generate a set of annotations for the full duration of every lesson. The structure of the machine annotations are identical to the human annotations.

We randomly select 20% of annotated segments for each language to use as evaluation data, and use the remainder of the annotated data as training data. Summary statistics about both sets of annotation data can be seen in Table 2.

4.2.1. Diarization

We use Pyannote (Bredin, 2023; Plaquet and Bredin, 2023) for speaker diarization. We experiment with both fine-tuning the speaker embeddings and segmentation model and with fine-tuning the diarization hyperparameters, but ultimately find the default pipeline to be the most performant. Note that because we perform speaker classification as a separate supervised task, we do not actually use the speaker clustering labels from the diarization pipeline, and care only about speech segmentation. This means that we could perform only voice activity detection (VAD) instead of full speaker diarization, but we found that diarization outperformed plain VAD for speech segmentation. Diarization also has the benefit of allowing us to process overlapping utterances from different speakers.

Segmentation F1-score for our system can be seen in Table 3. While Pyannote's diarization model performed well enough to produce useful results, it is also the weakest component of our pipeline, likely due to background noise and fluctuating audio quality in the lesson recordings. We experimented with other diarization models such as those from the NeMo toolkit (Harper et al., 2019), but were unable to find any which outperformed Pyannote.

	Diarization	VAD Only
Arabic	79.6	79.0
Spanish	69.4	66.6
Chinese	86.1	84.6

Table 3: Segmentation F1 score, computed as the harmonic mean of purity and coverage[2]

[2] For details on purity, coverage and Segmentation F1 computation, see the Pyannote metrics documentation.

4.2.2. Utterance Classification

We treat language and speaker identification as supervised classification tasks, where the input is the audio of a single utterance and the output is a label representing the dominant language in the utterance or the speaker of the utterance, respectively.

We primarily experiment with Whisper (Radford et al., 2022) for these tasks, as it is known to perform well not only on automated speech recognition (ASR) but also on related speech and sound detection tasks (Gong et al., 2023). We found that Whisper (the `whisper-large-v2` model) performs quite well when fine-tuned on our data. For speaker classification, we remove Whisper's autoregressive decoder component and replace it with a simple linear classification head with one hidden layer, such that the output of Whisper's encoder is fed directly into the classifier. As can be seen in Table 4 and the classification row for Table 5, this configuration performs fairly well on our data.

	Arabic	Spanish	Chinese
Classification	90%	92%	95%

Table 4: Speaker identification accuracy

We try the same classifier configuration for language identification, but find that our best performance comes from using the standard Whisper architecture, including decoder, fine-tuned on our data. That is, we use our ASR model for language identification by taking a single decoding step and selecting the most likely language token representing either English or the target language. We hypothesize that this performance gap exists because Whisper is already trained to output language tokens, and consequently has learned how to perform language identification using parameters in its decoder.

	Arabic	Spanish	Chinese
whisper-large-v2	46%	59%	76%
ASR fine-tuned	95%	95%	92%
Classification	95%	89%	90%

Table 5: Language identification accuracy

4.2.3. Automatic Speech Recognition

We also use Whisper for automatic speech recognition (ASR), finding once again that fine-tuning on our annotated data substantially improves performance. For both training and evaluation, the input in all cases is a single utterance as annotated by our human annotators, with the annotated speech as gold output labels. Note that we also provide the language of the utterance to the model by forcing the first decoded token to be the language token representing the utterance's dominant language. We use human-annotated gold language labels when training and evaluating our ASR models. Both classification and ASR models were fine-tuned for three epochs with a batch size of eight and a learning rate of 1×10^{-6}, using the cross-entropy loss. We measure ASR performance with character error rate (CER), in part because there is no standard way to calculate word error rate (WER) for languages without spaces like Chinese. Character error rate can be thought of as a measurement of the edit distance between the output of the model and the reference transcription. That is, given a reference of length N characters and model output which can be transformed into this reference with S substitutions, D deletions and I deletions, CER is computed as:

$$CER = \frac{S + D + I}{N} \tag{1}$$

ASR model performance can be seen in Table 6. Fine-tuning the model improves performance on all languages, but most dramatically for Arabic and Chinese, where the error rates after fine-tuning are nearly half of the original. We speculate that Whisper may have benefited more from fine-tuning in these two languages because it was weaker in them to begin with: Whisper's reported ASR performance on Arabic and Chinese was substantially worse than Spanish in the original work (Radford et al., 2022).

	Arabic	Spanish	Chinese
whisper-large-v2	60%	33%	32%
ASR fine-tuned	25%	28%	17%

Table 6: CER on each language for ASR models. Punctuation and Arabic diacritics are excluded for all CER computation.

4.2.4. Pipeline Scoring & Error Propagation

Scores in the previous sections are computed by comparing model outputs to human outputs for each human-annotated utterance. However, when running the machine annotation pipeline there is no guarantee that output from diarization or other steps will be correct, and consequently we can expect some degree of error propagation to later tasks in the pipeline. In particular, errors in speech segmentation are potentially damaging to all other tasks, and errors in language classification could lead to worse ASR output because the utterance language is used to bias the ASR model's output.

Because diarization output will not line up perfectly with human annotated utterances, we compute metrics per five minute human-annotated segment instead of per utterance in order to accurately gauge the performance of our pipeline. CER is computed by concatenating the speech in all utterances output by the pipeline and comparing it to the concatenation of all human-annotated utterance text. For speaker and language identification, we compute the identification error rate (IER) for each. IER can be thought of as a measurement of the percentage of the total duration that is classified incorrectly in some way, and is calculated as:

$$IER = \frac{f + m + c}{t} \tag{2}$$

Where f is the duration of false positives (non-speech incorrectly identified as speech), m is the duration of missed speech (speech incorrectly identified as non-speech), c is the duration of correctly identified speech assigned the wrong classification label, and t is the total duration. Note that because f and m both depend exclusively on the performance of the model identifying speech, IER is particularly sensitive to diarization performance.

		Arabic	Spanish	Chinese
Spk ID	Gold Seg	5%	10%	4%
	Pipeline	24%	27%	17%
Lang ID	Gold Seg	4%	3%	5%
	Pipeline	24%	23%	21%
ASR	Gold Seg	23%	27%	16%
	−Gold Lang	28%	27%	17%
	Pipeline	34%	33%	31%

Table 7: Error rates for pipeline components: CER for ASR and IER for classification

In Table 7, we present the performance of our pipeline components using human-annotated gold speech segmentation, and pipeline diarization. We also include ASR with gold speech segmentation but pipeline language identification. As we can see from these results, errors in diarization have a substantial effect on the performance of downstream tasks. Precise start and end times for utterances are arguably not necessary for downstream analysis focused on speech content, suggesting that the increase in IER for classification tasks may not matter in some cases, but errors in diarization also lead to an average increase in CER of approximately 10% for ASR. We leave speech segmentation approaches which are more resilient to issues such as variable audio quality to future work.

5. Experiments

5.1. Linguistic Analysis

To demonstrate the kind of analysis that our data can be used for, we compute summary statistics to track changes in the learner and teacher's speech over time. We use a mix of human and machine-annotated data for this, using human data where available and machine-annotated data otherwise.

We begin by examining the percentage of utterances made in the target language by both the teacher and student in each lesson. This is important both because listening and speaking practice are critical to language acquisition, and because for students the degree of target language use in a learning context has been linked to proficiency in that language (Turnbull and Dailey-O'Cain, 2009; Carranza, 1995). We find that the percentage of target language utterances consistently increases over time for both the student and teacher: Spearman's ρ for the correlation between lesson number and % of target language utterances ranged from 0.32 to 0.73, with all p values < 0.01. Data for the Spanish student can be seen in Figure 5. All figures presented in this section are linear regressions with 95% confidence intervals.

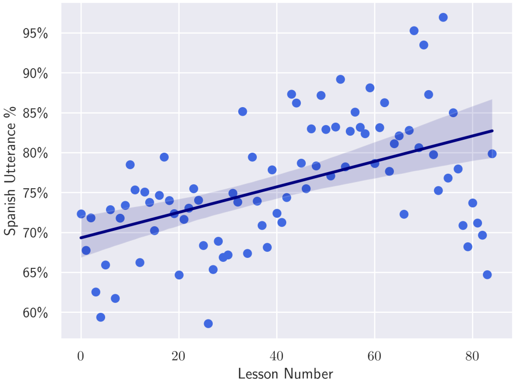
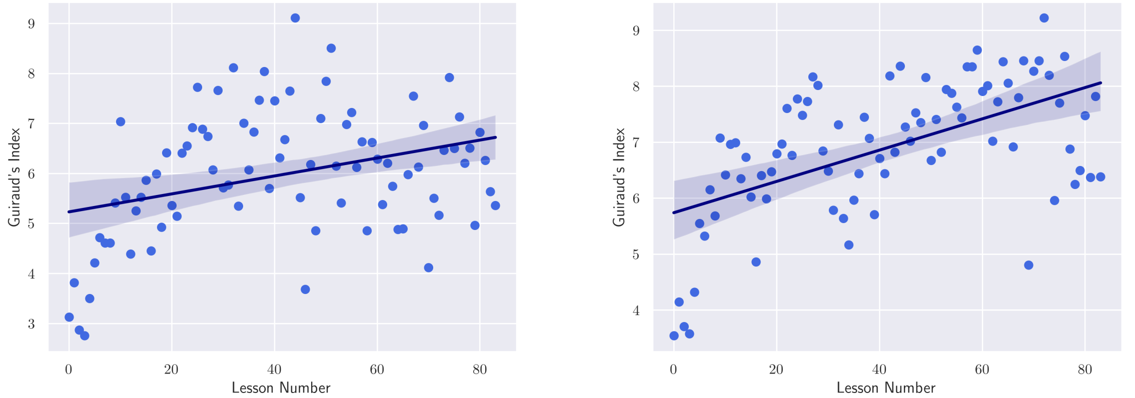


Figure 5: Student Spanish utterance %

Next, we look at metrics designed to measure lexical diversity in speech, as they have been shown to correlate with assessments of learner ability (Engber, 1995) and can grow over time for language learners (Hsieh, 2016). For computing these metrics, we tokenize Spanish and Chinese using spaCy (which internally uses pkuseg for Chinese) (Honnibal et al., 2020; Luo et al., 2019), and Arabic with CAMeL tools (Obeid et al., 2020). Token data is then cleaned by removing tokens consisting of punctuation, numbers, whitespace and stop words.

Token-type ratio (TTR) is one measure of lexical diversity which is commonly used in linguistics research (Thomas, 2005). TTR is calculated as



(a) Student

(b) Teacher

Figure 4: Guiraud's index for the Chinese student and teacher

the number of unique tokens (words) divided by the number of total tokens, and ideally would be expected to increase over time as the learner's vocabulary expands and the teacher moves on to using more complex language. However, TTR has been shown to be unstable in some circumstances, such as when there is substantial variance in the total number of tokens (Van Hout and Vermeer, 2007). This instability is mirrored in our results: while TTR grows over time for some students and teachers in some languages, correlations are often weak or have p values substantially higher than 0.05 suggesting no correlation at all.

Some alternatives to TTR have been proposed to address its shortcomings. In particular, we look at Guiraud's index (Guiraud, 1954), which mitigates the influence of total number of tokens by using the square root of the total token count as the denominator. We present standard TTR and Guiraud's index below as TTR and $TTR_{guiraud}$ below, where N is the total number of tokens and V is the number of *unique* tokens.

$$TTR = \frac{V}{N} \qquad TTR_{guiraud} = \frac{V}{\sqrt{N}} \tag{3}$$

Metric		Arabic	Spanish	Chinese
Target Lang %	Student	0.58	0.48	0.46
	Teacher	0.72	0.32	0.73
Guiraud's Index	Student	0.32	0.38	0.30
	Teacher	0.37	0.55	0.53

Table 8: Spearman's ρ for correlation between summary statistics and lesson number. All correlations have $p < 0.01$.

We find that Guiraud's, like % of target language utterances, consistently increases over time (i.e. correlates with lesson number) for both students and teachers as can be seen in Table 8. Spearman's ρ ranges from 0.30 to 0.55 with all p values < 0.01. Interestingly, this effect is measurably stronger for teachers, who had a mean ρ of 0.48 as opposed to students' 0.33. A comparison of change in Guiraud's index for the Chinese student and teacher can be seen in Figure 4. Assuming that students made measurable progress over the course of their lessons and that teachers gradually increased the difficulty of lesson content, these results show that this progression is reflected in our data, and also speak to the suitability of Guiraud's index as a metric.

5.2. Multimodal Analysis

We also illustrate how the rich multimodal data in the MOSLA dataset can be harnessed to gain insights into teacher and student behaviors using modern machine learning techniques.

Our objective here is to use machine learning techniques to determine the area of focus for both the teacher and the student on the screen, based solely on unannotated raw audio and video data.

Specifically, we use the Matchmap method, as described in (Harwath et al., 2018), to align the raw audio and the image in an unsupervised manner. The underlying principle of this method is that when parts of the input image and the audio co-occur frequently, it results in a high similarity score for that combination. The Matchmap method, as shown in Figure 6, encodes an image and an audio clip using separate encoders, producing a grid or sequence of latent representations for each modality. Let $a_{t,u}$ be the u-th element of the audio representation vector $\mathbf{a}_t$ at time t, and $i_{x,y,u}$ be the u-th element of the image representation vector $\mathrm{i}_{x,y}$ at position (x, y). After applying a linear projection layer (f_a and f_i,

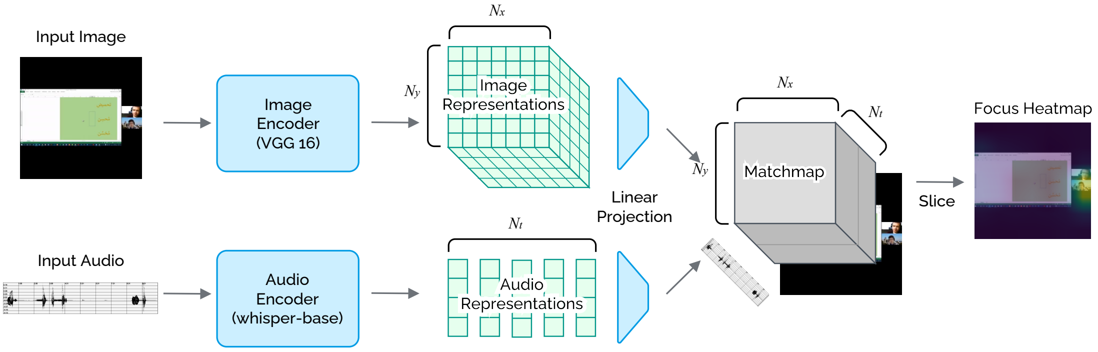


Figure 6: Overview of the Matchmap method

respectively) to each modality, the method computes a three-dimensional matrix called Matchmap as:

$$M_{x,y,t} = f_i(\mathbf{i})^T_{x,y} f_a(\mathbf{a})_t, \tag{4}$$

which quantifies the degree of "compatibility" between the image at position (x, y) and the audio at time t. Finally, the Matchmap matrix is aggregated to determine the overall similarity (referred to as SISA—Sum Image, Sum Audio) between a given image I and audio A instances using a simple arithmetic mean:

$$S(I, A) = \frac{1}{N_x N_y N_t} \sum_{x,y,t} M_{x,y,t} \tag{5}$$

where N_x, N_y, N_t denote the width and height of the encoded image, and the length of the encoded audio sequence, respectively.

To learn the Matchmap matrix without the need for labels, we adopt a contrastive learning approach. This approach maximizes the similarity between true image-audio pairs (I_i, A_i) while minimizing the similarity between randomly chosen "imposter" images I_i^{imp} and audio A_i^{imp}. Specifically, the Matchmap method uses the following loss function as the learning objective:

$$L = \sum_{i=1}^{N_b} \Big(\max(0, S(I_i, A_i^{imp}) - S(I_i, A_i) + \eta) + \max(0, S(I_i^{imp}, A^i) - S(I_i, A_i) + \eta) \Big) \tag{6}$$

where N is the number of instance per batch. Imposter images and audio were created by randomly permutating the instances within each batch. We set $\eta = 1$ in our experiments.

We initially extracted 100 random 10-second chunks from each Arabic lesson video. Images were generated by calculating the average of all the frames within each chunk. In this experiment, we used the pretrained VGG16 model (Simonyan and Zisserman, 2015) before the last pooling layer for encoding images and the Whisper (Radford et al., 2022) encoder (base model) for audio. Audio representations were downsampled by averaging every 10-frame window, resulting in 7×7 feature maps for images and 50 frames for audio. Both image and audio representations were then transformed using a 512-dimensional linear layer before computing the matchmap. The entire network, including the encoders and linear layers, was optimized using Adam with a learning rate of 1.0×10^{-4} for 30 epochs, with a batch size of 32.

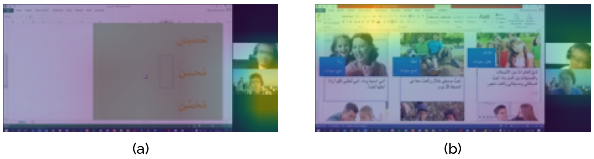


Figure 7: Example visualization of Matchmap

Figure 7 displays some examples of visualized matchmaps. These images were generated by slicing the computed matchmap at time t when discourse is taking place, whether initiated by the teacher or student, and then overlaying it as a heatmap onto the original image. As can be seen in the figure, the matchmap highlights relevant parts of the input image, such as the speaker (a) and/or the learning content (b). While we have not conducted a formal evaluation of this model, these results suggest that similar multimodal analytics approaches may prove effective for tasks such as speaker diarization, automated speech recognition, and facial expression analysis.

6. Conclusion

In Project MOSLA (Moments of Second Language Acquisition), we address the complexity of SLA by creating a longitudinal, multimodal, multilingual, and controlled dataset that captures every moment of SLA learners' experiences through online instruction. With human and machine annotations generated using state-of-the-art speech models, the MOSLA dataset provides insights into the distribution of spoken language, speaker identities, and the content of spoken discourse. Our experiments highlight the potential of this resource in revealing target language usage and lexical development, as well as in identifying the areas of focus for both learners and educators during interactions. By offering open access to the MOSLA dataset for research and non-commercial purposes, we hope to inspire a wide array of studies, fostering a deeper understanding of the multifaceted nature of SLA and facilitating the development of more effective pedagogical approaches for second language learners.

7. Ethical Considerations

As it is difficult to imagine possible harms as a result of further research or technology built on a dataset about language acquisition, our primary ethical concerns relate to the fairness of compensation and exposure to risk for participants in the study. In regards to compensation: all participants—students, teachers, and annotators—were paid well above the minimum hourly wage in the country in which this research was conducted.

We view risk to participants as consisting broadly of two categories: possible exposure of personally identifiable information (PII) relating to teachers or students, and possible appropriation of teaching material. Our primary mitigation against these risks is that access to MOSLA data will require consenting to a terms of use document which explicitly prohibits attempts to extract PII, appropriate teaching materials, redistribute the data, or otherwise use it for anything other than research. There is no explicit PII included anywhere in the data; our concern is only preventing the possibility of PII being inferred from conversation content in lessons. Furthermore, all participants knew from before their first lesson that they were being recorded with the intent of eventually publishing the data, had the option to withdraw at any point, and had and continue to have the right to request removal of any data, at any time, for any reason.

One other possible area of concern is copyrighted materials. In order to address this, teachers were asked to refrain from using copyrighted materials except in a supplementary capacity, and we are confident that any such usage included in the MOSLA lessons falls under fair use for teaching and research.

Finally, in place of an IRB or equivalent institutional review board which we did not have access to, we had a third-party ethics review conducted by an external resesarcher with an extensive background in AI and data ethics.

8. Acknowledgements

We would like to thank Paulino Brener and Abdulwahed Danou for their participation as teachers and Xiaoling Mo for her participation as both teacher and learner. Participants are thanked in the acknowledgements at their request, and had the option to remain anonymous.

We would also like to thank Keisuke Sakaguchi and Kasumi Yamazaki for their valuable advice, which informed this research. Finally, we would like to thank Marius Miron for providing a third-party ethics review.

Bibliographical References

Hervé Bredin. 2023. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Proc. INTERSPEECH 2023*.

Christopher Bryant, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe. 2023. Grammatical Error Correction: A Survey of the State of the Art. *Computational Linguistics*, pages 1–59.

Andrew Caines, Helen Yannakoudakis, Helena Edmondson, Helen Allen, Pascual Pérez-Paredes, Bill Byrne, and Paula Buttery. 2020. The teacher-student chatroom corpus. In *Proceedings of the 9th Workshop on NLP for Computer Assisted Language Learning*, pages 10–20, Gothenburg, Sweden. LiU Electronic Press.

Isolda Carranza. 1995. Multilevel analysis of two-way immersion classroom discourse. *Georgetown University round table on languages and linguistics*, pages 169–187.

Ngo Cong-Lem. 2018. Web-based language learning (wbll) for enhancing l2 speaking performance: A review. *Advances in Language and Literary Studies*.

Dorottya Demszky and Heather Hill. 2023. The NCTE transcripts: A dataset of elementary math classroom transcripts. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 528–538, Toronto, Canada. Association for Computational Linguistics.

Patrick J. Donnelly, Nathaniel Blanchard, Andrew M. Olney, Sean Kelly, Martin Nystrand, and Sidney K. D'Mello. 2017. Words matter: Automatic detection of teacher questions in live classroom discourse using linguistics, acoustics, and context. In *Proceedings of the Seventh International*

Learning Analytics & Knowledge Conference, LAK '17, page 218–227, New York, NY, USA. Association for Computing Machinery.

Cheryl A Engber. 1995. The relationship of lexical proficiency to the quality of esl compositions. *Journal of second language writing*, 4(2):139–155.

Jason Fan and Xun Yan. 2020. Assessing speaking proficiency: A narrative review of speaking assessment research within the argument-based validation framework. *Frontiers in Psychology*, 11.

Jeroen Geertzen, Theodora Alexopoulou, and Anna Korhonen. 2014. Automatic linguistic annotation oflarge scale l2 databases: The efcambridge open language database(efcamdat).

Yuan Gong, Sameer Khurana, Leonid Karlinsky, and James Glass. 2023. Whisper-AT: Noise-Robust Automatic Speech Recognizers are Also Strong General Audio Event Taggers. In *Proc. INTERSPEECH 2023*, pages 2798–2802.

Pierre Guiraud. 1954. Les caractères statistiques du vocabulaire: essai de méthodologie. *(No Title)*.

Regine Hampel and Ursula Stickler. 2012. The use of videoconferencing to support multimodal interaction in an online language classroom. *ReCALL*, 24(2):116–137.

Eric Harper, Somshubra Majumdar, Oleksii Kuchaiev, Li Jason, Yang Zhang, Evelina Bakhturina, Vahid Noroozi, Sandeep Subramanian, Koluguri Nithin, Huang Jocelyn, Fei Jia, Jagadeesh Balam, Xuesong Yang, Micha Livne, Yi Dong, Sean Naren, and Boris Ginsburg. 2019. NeMo: a toolkit for Conversational AI and Large Language Models.

David Harwath, Adrià Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass. 2018. Jointly discovering visual objects and spoken words from raw sensory input. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part VI*, page 659–677, Berlin, Heidelberg. Springer-Verlag.

Yuta Hayashibe. 2021. Hachiue. `https://github.com/koniwa/hachiue`.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Yufen Hsieh. 2016. An exploratory study on singaporean primary school students' development in chinese writing. *The Asia-Pacific Education Researcher*, 25:541–548.

Rony Kubat, Philip DeCamp, and Brandon Roy. 2007. Totalrecall: Visualization and semi-automatic annotation of very large audio-visual corpora. In *Proceedings of the 9th International Conference on Multimodal Interfaces*, ICMI '07, page 208–215, New York, NY, USA. Association for Computing Machinery.

Mark R. Lepper and Maria Woolverton. 2002. Chapter 7 - the wisdom of practice: Lessons learned from the study of highly effective tutors. In Joshua Aronson, editor, *Improving Academic Achievement*, Educational Psychology, pages 135–158. Academic Press, San Diego.

Ruixuan Luo, Jingjing Xu, Yi Zhang, Zhiyuan Zhang, Xuancheng Ren, and Xu Sun. 2019. Pkuseg: A toolkit for multi-domain chinese word segmentation. *CoRR*, abs/1906.11455.

Hamed Monkaresi, Nigel Bosch, Rafael A. Calvo, and Sidney K. D'Mello. 2017. Automated detection of engagement using video-based estimation of facial expressions and heart rate. *IEEE Transactions on Affective Computing*, 8(1):15–28.

Su Mu, Meng Cui, and Xiaodi Huang. 2020. Multimodal data fusion in learning analytics: A systematic review. *Sensors*, 20(23).

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. CAMeL tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France. European Language Resources Association.

Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. 2015. Deep knowledge tracing. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Alexis Plaquet and Hervé Bredin. 2023. Powerset multi-class cross entropy loss for neural speaker diarization. In *Proc. INTERSPEECH 2023*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Deb Roy, Rupal Patel, Philip DeCamp, Rony Kubat, Michael Fleischman, Brandon Roy, Nikolaos Mavridis, Stefanie Tellex, Alexia Salata, Jethran Guinness, Michael Levit, and Peter Gorniak. 2006. The human speechome project. In *Symbol Grounding and Beyond*, pages 192–196, Berlin, Heidelberg. Springer Berlin Heidelberg.

Kazuya Saito and Yuka Akiyama. 2017. Video-based interaction, negotiation for comprehensibility, and second language speech learning: A longitudinal study. *Language Learning*, 67(1):43–74.

Burr Settles, Chris Brust, Erin Gustafson, Masato Hagiwara, and Nitin Madnani. 2018. Second language acquisition modeling. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 56–65, New Orleans, Louisiana. Association for Computational Linguistics.

Burr Settles, Geoffrey T. LaFlair, and Masato Hagiwara. 2020. Machine Learning–Driven Language Assessment. *Transactions of the Association for Computational Linguistics*, 8:247–263.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.

Katherine Stasaski, Kimberly Kao, and Marti A. Hearst. 2020. CIMA: A large open access dialogue dataset for tutoring. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–64, Seattle, WA, USA → Online. Association for Computational Linguistics.

Dax Thomas. 2005. Type-token ratios in one teacher's classroom talk: An investigation of lexical complexity.

Miles Turnbull and Jennifer Dailey-O'Cain. 2009. *First language use in second and foreign language learning*. Multilingual Matters.

Sowmya Vajjala and Taraka Rama. 2018. Experiments with universal CEFR classification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 147–153, New Orleans, Louisiana. Association for Computational Linguistics.

Roeland Van Hout and Anne Vermeer. 2007. Comparing measures of lexical richness. *Modelling and assessing vocabulary knowledge*, 93:115.

Mary Lou Vercellotti. 2015. The Development of Complexity, Accuracy, and Fluency in Second Language Performance: A Longitudinal Study. *Applied Linguistics*, 38(1):90–111.