Multi-Modal Learning Application - Support Language Learners with NLP Techniques and Eye-Tracking

Robert Geislinger, Ali Ebrahimi Pourasad, Deniz Gül, Daniel Djahangir, Seid Muhie Yimam, Steffen Remus, Chris Biemann
Language Technology Group, Universität Hamburg, Germany
firstname.lastname@uni-hamburg.de

Abstract

This paper presents a framework consisting of an iPad application and an NLP pipeline, designed to assist non-native speakers in learning English as a second language. The application provides beginner-level texts, which are augmented by contextual images to facilitate natural learning. The multi-modal iOS application can be fully controlled by employing eye-tracking components, aiming to enhance the reading experience by highlighting relevant parts of an image when the user naturally focuses a particular and potentially complex word. Moreover, this eye-tracking feature offers accessibility for individuals with physical disabilities.

1 Introduction

In our interconnected world, learning a new language is increasingly necessary for social, professional, or political purposes. Language acquisition is challenging, even though various supporting methods are available. For infants, parents often associate object names through pointing. Self-study of a language can involve using educational applications or engaging with media in the language. For instance, learning through activities like reading subtitles while watching films can be easier than solely relying on reading educational texts (Danan, 1992). This technique of learning, where individuals are presented with multiple representations, e.g., text and image, is known as multimodal learning. It has been shown in studies that this technique enhances learning comprehension (Wang et al., 2022).

This project offers a multi-modal learning application, which can be managed by tracking the users eye movement. It facilitates natural language learning by combining suitable sentences with related images. The target users of the application are beginners and individuals with motor difficulties, making it challenging for them to use touch-based applications. The application can be used independently by individuals or provided by organizations and educational institutions. The machine learning models used for the identifying a word and highlighting the respective object in an image are trained on English, but are easily exchangeable for other languages. The following user flow serves as an example of how the application can be used:

Upon launching the application, the user is presented with a selection of topics to choose from. After selecting a topic, a sentence is presented with a contextually fitting image. This could be a sentence about motorsports, accompanied by an image of a Formula One car that is relevant to the chosen topic. The NLP pipeline has previously identified potentially complex words which might be hard to learn or understand. While the user is reading the text, the eye-tracking component tracks the eye movement. If the user looks at a complex word, it is highlighted within the text and the image. E.g., if 'wheel' is identified as a complex word in the sentence, the wheels of the car in the image will be highlighted when the user looks at the word.

2 Related Work

The term Mobile Assisted Language Learning (MALL) was coined by Chinnery (2006) and describes the learning of languages with mobile devices. MALL applications can encompass a multimodal approach including face-to-face communication (Vigliocco et al., 2014) and the use of images and texts (Schneider et al., 2021). The popularity of MALL applications is evident, as seen in platforms like Duolingo[1], which has over 300 million users (Shortt et al., 2021). Language learning applications can support learners and enhance their speaking and critical thinking skills (Kusmaryani et al., 2019).

Eye-tracking is a method that tracks eye position

[1] https://www.duolingo.com

Proceedings of the 1st Workshop on Linguistic Insights from and for Multimodal Language Processing, pages 6–11

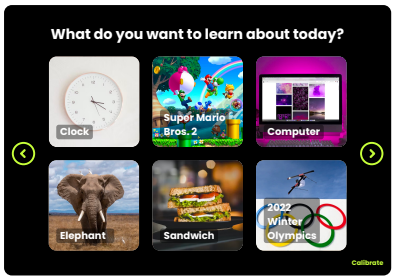

Figure 1: Menu View: Topic overview. Selection by touch and eye-tracking.

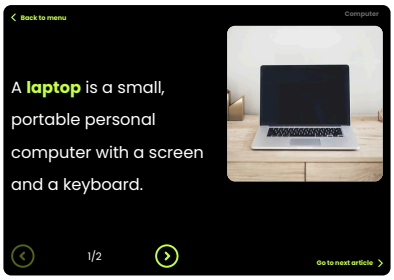

Figure 2: Reading View: A text about 'laptop' with the focus word laptop and a image about laptops.

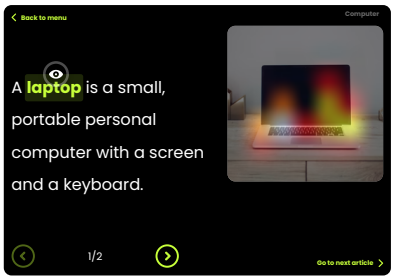

Figure 3: Reading View: The user looks at the word 'laptop which is highlighted in the image and text.

to identify an individual's gaze, such as images on a computer screen or real-world traffic signs. It has applications in psychology (Rahal and Fiedler, 2019; Li and Pollatsek, 2020), medicine (Harezlak and Kasprowski, 2018), and advertising (Lohse and Wu, 2001). Several eye-tracking solutions that differ in their accuracy and expense. Specialized eye-tracking hardware is often costly and used in laboratory environments. These devices are head-mounted (Cognolato et al., 2018) or use a fixed, steady camera in front of the user (Sharaev et al., 2021). Accessible eye-tracking for the masses as in the presented work can be achieved by utilizing inexpensive and commonly used consumer hardware, such as webcams or mobile devices (Papoutsaki, 2015). The main difference is that consumer hardware is generally less accurate, although the accuracy is improving with the evolution of consumer hardware such as mobile phones (Krafka et al., 2016). Technologies such as eye-tracking mostly benefits impaired people, but not exclusively (Elliott et al., 2019; Milde et al., 2021).

State-of-the-art computer vision models can predict unfamiliar concepts alongside predefined object categories by learning on datasets comprising ofnumerous images and their corresponding textual descriptions (Radford et al., 2021). By extracting visual and textual features from the input data and comparing them using a similarity metric, such models can determine the degree to which a given text input is related to a particular image. Using the approach, one can find the best matching image to a given text from a database of images (Salvador et al., 2017). Models for finding and highlighting parts of the image depending on a query are also available (Schneider and Biemann, 2022). New possibilities arise, like forecasting image content, but these models demand substantial computational resources for training and prediction, as well as extensive datasets to attain reasonable results (Radford et al., 2021).

3 System design

The system architecture consists of two main components: the frontend and the backend. The frontend is an iOS application that processes touch and eye-tracking inputs, while also displaying the pre-processed texts and images. The backend is used to process text-aligned image datasets and to extract important meta information, which is then used to present the user.

3.1 Frontend

An iOS application for the iPad was chosen as the frontend for the project due to access to Apple's augmented reality library, RealityKit[2], which provides eye and facial tracking capabilities and can generate screen coordinates of the user's current focus. The generated coordinates were found to be imprecise for accurate tracking, possibly because the library's coordinate system lacks calibration based on the user's distance and orientation to the device. A common practice to calibrate eye-tracking systems is to show calibration dots for the user to look at (Gunawardena et al., 2022). When the user opens the app, a custom calibration process starts to calculate more precise coordinates. The user gazes at four corner circles displayed on the screen to establish reference points. This step enhances the library's coordinate system, improving the accuracy of tracking the user's eye gaze. The user's viewpoint is represented by an eye pictogram within a circle, which is controlled by the user's eye movement, similar to a mouse pointer. This can be seen in Figure 3, where the eye pictogram is positioned above the word 'laptop'.

After calibration, the Menu View displays options to select a topic, as shown in Figure 1. Once a

[2] https://developer.apple.com/augmented-reality/realitykit/

topic is selected, the user is directed to the Reading View. Figure 2 illustrates the Reading View with a sentence about laptops, accompanied by an image that appropriately visualizes the sentence and its context. While reading, the application highlights the word 'laptop' in bold letters. Whenever the user's eye focuses on the word, it gets highlighted both in the sentence and in the image. This allows the user to learn the word intuitively without having to look up its definition. Figure 3 provides an example of this. After completing a sentence, the user can either learn more sentences within the same topic or move on to a different topic.

In order to create a functional eye-tracking system, several factors need to be taken into account. This is essential for accurately tracking the user's eyes and ultimately influencing their interaction with the application. The system utilizes the iPad's front camera, which has lower image quality than the rear camera. This introduces uncertainty due to lower resolution and issues related to low-light conditions. To overcome this uncertainty, it is necessary to optimize and mitigate other aspects of the eye-tracking system. When the iPad is in landscape mode, the camera is positioned on the side instead of the center, leading to more accurate eye-tracking on the side facing the camera. To get feedback on the tracking, five volunteers were asked to test the application on an iPad Mini 6th generation and iPad Pro 5 generation in a small pilot study. The different technical details of the devices, such as screen size, camera and processor, made it possible to look at various aspects. The users have reported problems with tracking on both devices and orientations. However, tracking consistently worked better when the elements were placed on the side closer to the camera and larger elements could be focused better than smaller elements. To address this issue, precise tracking elements, such as educational texts in Reading View, are positioned on the side facing the camera. In addition, elements as buttons and texts, are enlarged to help avoid collisions with the user's focused eye position during tracking. User head movement can significantly reduce eye-tracking system accuracy.

The system recalibrates the coordinate system if the predicted viewport is close to a button. It is assumed that the eye-tracking mechanism is imprecise, and the user is fixating at the center of the button. The offset between the button's center and the tracked point is calculated to adjust the coordinate system. To prevent accidental button activation, the user must gaze at the button for three seconds. The recalibration process is performed multiple times within the three-second period, with the ring around the pointer acting as a progress bar. After this period, the button's command is executed. This mechanism is also used to initiate the recalibration process when the user begins reading a text. In this application, users read texts from the top left corner to the bottom right. When the user's eyes are tracked near the first word, the recalibration is applied.

3.2 Backend

Figure 4 shows an overview of the preprocessing NLP pipeline, which filters the text documents and enriches them with corresponding descriptive images. The pipeline is based on Wang et al. (2022). The text dataset is a collection of documents from the Simple English Wikipedia[3]. This dataset covers a wide range of topics, including animals, food, cities and other subjects, making it diverse. The pipeline tokenizes the documents into sentences and processes them independently. First, the pipeline identifies complex words in a sentence, primarily those that exceed the language classification level B1 according to the Common European Framework of Reference for Languages (CEFR). For example, the word 'minute' in the context of time is classified as A1 (Beginner) and 'a minute amount of fuel' as quantification as C2 (Proficiency English). This classification is performed using the complex word identification algorithm developed by Srivastava (2022). The algorithm utilizes a sequential model developed by Rei (2017) that incorporates hand-engineered features, along with word embeddings, to classify complex words based on their context.

In addition to identifying complex words, depictable words are also identified. A word is considered highly depictable if it can be easily visualized, such as the word 'dog,' which represents a physical entity. On the other hand, the word 'creativity' is difficult to visualize because it represents an abstract concept without a concrete visual representation (Hessel et al., 2018).

First, the annotated image dataset, MSCOCO (Lin et al., 2014), is used to initialize the algorithm. Then the algorithm calculates the concreteness scores for the words in the annotations. If the

[3] https://simple.wikipedia.org

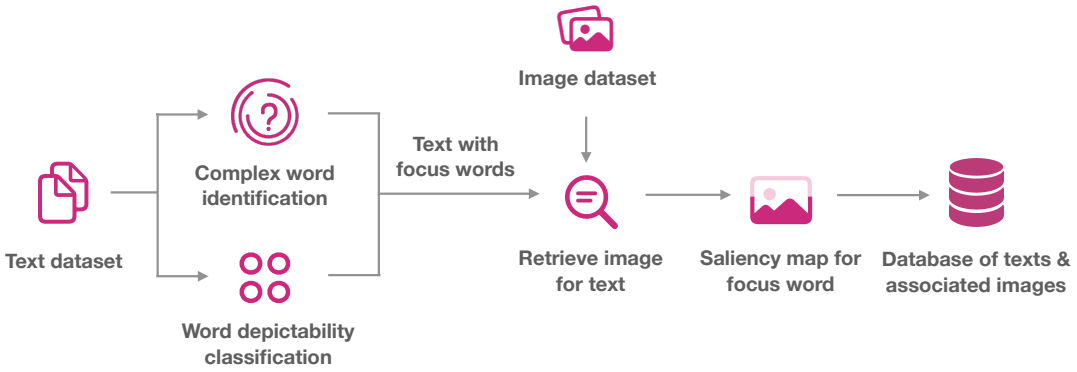

Figure 4: The preprocessing NLP pipeline, which enriches text with context-fitting images.

score exceeds a threshold of 50, as tested by Hessel et al. (2018), and the corresponding word is a noun, it is considered depictable. Once the depictable items are identified from the image dataset annotations, they are mapped to their corresponding words in the textual dataset, if those words exist.

After the complexity and representability classification, only those words that satisfy both criteria are considered. These complex and depictable words are referred to as 'focus words' (Wang et al., 2022) which require explanation and can be represented visually to facilitate learning. In the end, only sentences that contain at least one focus word are retained in the text dataset.

Next, each sentence in the filtered text dataset is matched with a relevant image that showcases the contextualized focus words. This step is crucial as words can possess multiple meanings based on their context. For example, the word 'bank' can refer to a shore in a river or a financial institution. To find relevant images, the CLIP model[4] (Radford et al., 2021) calculates the cosine similarity between images and words or sentences. The image dataset used is MSCOCO[5], which is also utilized for the word depictability classification. An image is considered similar to a sentence or word, if the similarity value calculated by CLIP exceeds a threshold of 4.0. Following the approach by Wang et al. (2022), sentences are only processed further, if there are five similar candidate images. The most similar candidate image to the focus words is selected as the associated image for a sentence. If none of the five candidate images show similarity to any focus word in the sentence, the sentence is excluded.

To highlight the focus words in the image, miniCLIP[6] is utilized for visualization. The generated saliency maps are superimposed on the original image shown in Figure 2 and Figure 3.

Once all sentences in a text document are processed throughout the processing steps, the text document and its retrieved images are stored in the database. The frontend can then access all the topics, sentences, and accompanying images from the database through a REST API.

4 Conclusion

The goal of this project was to support novice language learners by developing an educational iPad application. The application designed combines modern NLP techniques and eye-tracking technology enabling a multi-modal learning experience with beginner-friendly texts and accompanying images that help illustrate the content. The integrated eye-tracker analyzes the users' reading behavior and enhances their reading experience by highlighting relevant parts of images. Furthermore, eye-tracking enables individuals with physical disabilities to access the application. One of the major challenges encountered was implementing eye tracking on the iPad. Despite the efforts, improving the accuracy and stability of the eye-tracking system is necessary for it to be considered user-friendly. The main issues are the low image quality of the iPad's camera and ensuring the users's head stability during use. To address these challenges, one could explore alternative eye-tracking algorithms or contemplate integrating an external camera in the future to improve image quality.

As an alternative to relying solely on datasets, one could leverage AI generation tools such as Stable Diffusion (Rombach et al., 2022) or GPT-4 (OpenAI, 2023), which have the ability to create

[4] https://github.com/openai/CLIP
[5] https://cocodataset.org
[6] https://github.com/HendrikStrobelt/miniClip

images based on input descriptions.

The next step in this research should involve conducting user studies with language learners to quantitatively evaluate the effectiveness of using eye-tracking technology to highlight objects in contextual images during the learning process. How much users benefit from contextual images compared to users without this support would be part of an evaluation study. Also a usability study should be carried out with the aim of adapting the application to the needs of the users in the best possible way.

The project is openly available under a permissive Apache v2 License[7].

5 Acknowledgments

Funded by the Federal Ministry of Education and Research (BMBF) and the Free and Hanseatic City of Hamburg under the Excellence Strategy of the Federal Government and the Länder.

6 Limitations

The models utilized in the NLP pipeline have been specifically trained for the English language. While the pipeline can potentially be adapted to other languages with appropriate datasets, the availability of such datasets remains a challenge. The hard filtering process employed during dataset creation limits the languages for which fitting datasets are readily accessible. This restriction poses a barrier to deploying the pipeline for languages with small available datasets, as it would require significant efforts to collect and curate appropriate data for training.

The application relies on a server for its functionality, which poses a limitation in terms of scalability and availability. Running the application without a server connection is currently not possible, hindering its use in offline environments. Future improvements could explore alternative approaches, such as client-side implementations or optimizing server dependencies to minimize their impact on the application's usability.

Another important consideration is the computational power required to preprocess the data using the pipeline. The image and text data need to be processed beforehand to achieve satisfying results, which necessitates a server with sufficient computational capabilities.

[7] https://github.com/Alienmaster/MultimodalLearningIOSApp

7 Ethical Aspects

The ethical aspects of a language learning application with eye-tracking for disabled people revolve around ensuring inclusivity and equal opportunities for individuals with disabilities. The application prioritize user privacy and data security, ensuring that the eye-tracking data is not collected, shared or exploited. Even though the application was developed with a focus on eye-tracking, it is also fully usable with touch to give the user a choice. By providing the complete software and source code, including all models and data sets, users and developers can trace the use of the data within the application. If eye-tracking data is later used for optimisation purposes, sufficient safeguards must be in place to protect the security of the users' data. Due to the complete open-source approach, there are still no costs for either the user or the developer.

References

George M Chinnery. 2006. Going to the MALL: Mobile assisted language learning. *Language learning & technology*, 10(1):9–16.

Matteo Cognolato, Manfredo Atzori, and Henning Müller. 2018. Head-mounted eye gaze tracking devices: An overview of modern devices and recent advances. *Journal of Rehabilitation and Assistive Technologies Engineering*, 5(13):1–13.

Martine Danan. 1992. Reversed Subtitling and Dual Coding Theory: New Directions for Foreign Language Instruction. *Language Learning*, 42(4):497–527.

Michael A Elliott, Henrique Malvar, Lindsey L Maassel, Jon Campbell, Harish Kulkarni, Irina Spiridonova, Noelle Sophy, Jay Beavers, Ann Paradiso, Chuck Needham, et al. 2019. Eye-controlled, power wheelchair performs well for ALS patients. *Muscle & nerve*, 60(5):513–519.

Nishan Gunawardena, Jeewani Anupama Ginige, and Bahman Javadi. 2022. Eye-tracking Technologies in Mobile Devices Using Edge Computing: A Systematic Review. *ACM Computing Surveys*, 55(8):1–33.

Katarzyna Harezlak and Pawel Kasprowski. 2018. Application of eye tracking in medicine: A survey, research issues and challenges. *Computerized Medical Imaging and Graphics*, 65:176–190. Advances in Biomedical Image Processing.

Jack Hessel, David Mimno, and Lillian Lee. 2018. Quantifying the visual concreteness of words and topics in multimodal datasets. In *Proceedings of*

the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 2194–2205, New Orleans, Louisiana. Association for Computational Linguistics.

Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. 2016. Eye Tracking for Everyone. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2176–2184, Las Vegas, USA.

W Kusmaryani, B Musthafa, and Pupung Purnawarman. 2019. The influence of mobile applications on students' speaking skill and critical thinking in English language learning. In *Journal of Physics: Conference Series*, volume 1193, pages 1–6, Bogor, Indonesia.

Xingshan Li and Alexander Pollatsek. 2020. An Integrated Model of Word Processing and Eye-Movement Control During Chinese Reading. *Psychological Review*, 127(6):1139–1162.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *In Proceedings of European conference on computer vision (ECCV)*, pages 740–755, Zurich, Switzerland.

Gerald L Lohse and DJ Wu. 2001. Eye Movement Patterns on Chinese Yellow Pages Advertising. *Electronic Markets*, 11(2):87–96.

Benjamin Milde, Robert Geislinger, Irina Lindt, and Timo Baumann. 2021. Open Source Automatic Lecture Subtitling. In *Proceedings of Electronic Speech Signal Processing 2021 (ESSV)*, pages 128–134, Berlin, Germany.

OpenAI. 2023. GPT-4 Technical Report.

Alexandra Papoutsaki. 2015. Scalable Webcam Eye Tracking by Learning from User Interactions. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems 2015 (CHI)*, pages 219–222, Seoul, Republic of Korea.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *In Proceedings of International Conference on Machine Learning (PMLR)*, volume 139, pages 8748–8763, online.

Rima-Maria Rahal and Susann Fiedler. 2019. Understanding cognitive and affective mechanisms in social psychology through eye-tracking. *Journal of Experimental Social Psychology*, 85:1–14.

Marek Rei. 2017. Semi-supervised Multitask Learning for Sequence Labeling. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2121–2130, Vancouver, BC, Canada.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, New Orleans, LO, USA. IEEE Computer Society.

Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, and Antonio Torralba. 2017. Learning cross-modal embeddings for cooking recipes and food images. In *Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3020–3028, Honolulu, HI, USA.

Florian Schneider, Özge Alaçam, Xintong Wang, and Chris Biemann. 2021. Towards multi-modal text-image retrieval to improve human reading. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, Online. Association for Computational Linguistics.

Florian Schneider and Chris Biemann. 2022. Golden Retriever: A Real-Time Multi-Modal Text-Image Retrieval System with the Ability to Focus. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3245–3250, New York, NY, United States.

Maxim Sharaev, Svetlana Sushchinskaya, Valentina Bachurina, George Taranov, Evgeny Burnaev, and Marie Arsalidou. 2021. Machine learning, eye movements and mathematical problem solving. *Journal of Vision (jov)*, 21(9):2397–2397.

Mitchell Shortt, Shantanu Tilak, Irina Kuznetcova, Bethany Martens, and Babatunde Akinkuolie. 2021. Gamification in mobile-assisted language learning: A systematic review of Duolingo literature from public release of 2012 to early 2020. *Computer Assisted Language Learning*, pages 1–38.

Ankit Srivastava. 2022. Complex Word Identification for Language Learners. Master's thesis, Universität Hamburg, Hamburg, Germany.

Gabriella Vigliocco, Pamela Perniss, and David Vinson. 2014. Language as a multimodal phenomenon: implications for language learning, processing and evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1651):1–7.

Xintong Wang, Florian Schneider, Özge Alaçam, Prateek Chaudhury, and Chris Biemann. 2022. MOTIF: Contextualized Images for Complex Words to Improve Human Reading. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)*, pages 2468–2477, Marseille, France.