# Introduction

We know that bilingually raised children tend to pick up on learning a new language more quickly and efficiently than monolingual children. Studies suggest that bilingualism enhances cognitive flexibility, phonological awareness, and general language-learning ability. (Kovács & Mehler, 2009)

Word segmentation is the process of recognizing where one word ends, and another begins. This is an important skill for learning a language, whether as a baby hearing speech for the first time or an adult learning a new language. People use different clues to figure out word boundaries, such as the rhythm of speech, common sound patterns, and how often certain sounds appear together (Swingley & Algayres, 2024).

If a child is not raised with two languages but gets frequent exposure to a foreign language, would they also do a better job at understanding and learning a new language than a child that gets exposed to only one language?

Since testing this on real children is difficult for ethical and practical reasons, I will use Automatic Speech Recognition (ASR) models to simulate language learning. Research shows that ASR models help us understand how people process and learn languages (Dupoux, 2018). By training these models with different language experiences, we can see how well they recognize word boundaries in a new language. This could give us clues about whether hearing multiple languages helps with learning a new one.  
This is the basis of what I will be doing to answer the research questions; “Can a computational model with regular exposure to a second language do a better job at word splicing compared to a computational model that did not get this exposure?”

# Methode

For this we will use self-supervising ASR models, as children also just learn by listening to adults, rather than getting labelled data & regular feedback. The model I will be using is wav2vec 2.0, which is an open-source ASR model by Meta AI that can both be pretrained as well as be trained from scratch with either audio or transcripts data in the language that I desire. This way, if translating the audio data so we can evaluate the results turns out to be too complicated, we have textual input to fall back on. Wav2vec 2.0 uses Connectionist Temporal Classification (CTC) loss, which makes it particularly well-suited for segmentation tasks (Inaguma, Dalmia, Yan, Watanabe, 2021). It also provides built-in methods for extracting embeddings from raw waveforms, making direct speech processing easier as it doesn’t require prior phonetic transcriptions. (Baevski, Hsu, Conneau, Auli, 2021)

We need a lot of speech data to train the model (Synnaeve et al, 2020). The performance of the model will be based on the language, the quality of the data & the variability of the speech (Beech & Swingley, 2023). The best open-source data that fits all these criteria is Common Voice by Mozilla, as they have a high speaker diversity, over 50 available languages, clear isolated voice sounds and an option for textual data to fall back on if the evaluation for sound proves too difficult.

We will start by training the model on the English language. This will be our base model; we will then test it on a different language from a similar language family, German in the case, as this should help our model with the segmentation task (Caines, Altmann-Richer, Buttery, 2019). Since we are working with limited data, going for a phonetically, syntactically & segmentation wise very different language might limit the results. Once we have an idea of what this amount of training does for our model, we will train the model again, with the same amount of English data, but add a yet undecided amount of data in another foreign Language. Then test this model again to see if this helped the model perform better on the segmentation task.   
It would be particularly interesting to see if adding a language in the same language family as the previous (Dutch, for example) or something less closely related (French) will yield different results. More importantly however, we will have a look at how much time of the second language will result in significant improvements. Is sitting your child in front of Dora for an hour a day, going to help their Spanish grades later in life?

To evaluate the word boundaries found by our model in the test data. We will be using a CTC-based approach (Kürzinger , Winkelbauer, Li, Watzel, & Rigoll, 2020), that uses forward propagation, backtracking and the CTC score given by our wav2vec model to give us a confidence score on how accurate the model has sliced the words of the foreign language. For this we will also be using an open-source python package by the same name (Ludwig Kürzinger, CTC - segmentation, 2022). The final evaluation will include word boundary accuracy, CTC alignment scores, the Levenshtein distance and Phoneme Error Rate as done in the research by Swingley and Algayres, in their paper on Segmentation of Sentence Stimuli (Swingley & Algayres, 2024).

# Relevance to Artificial Intelligence

This research is relevant for AI as it explores how exposure to multiple languages influences learning in self-supervised speech models. Data is hard and expensive to come by. So not having to fully train a model on multiple language data sets while still getting sufficient results could save both time and money.

# References

Kovacs, A. M., & Mehler, J. (2009, April 13). *Cognitive gains in 7-month-old bilingual infants.* Proceedings of the National Academy of Sciences, 106(16), 6556–6560. https://doi.org/10.1073/pnas.0811323106

Baevski, A., Hsu, W. N., Conneau, A., Auli, M. (2021, October 25). Unsupervised Speech Recognition: *Wav2vec.* San Diego, CA, NeurIPS https://ai.meta.com/research/publications/unsupervised-speech-recognition/

Beech, C., & Swingley, D. (2023, June). Consequences of phonological variation for algorithmic word segmentation. Computational Modeling of the Segmentation of Sentence Stimuli From an Infant Word-Finding Study. Cognitive Science: Volume 48, Issue 3, e13427. https://onlinelibrary.wiley.com/doi/full/10.1111/cogs.13427

Swingley, D., Algayres, R., (2024, March 25). *Computational Modeling of the Segmentation of Sentence Stimuli From an Infant Word-Finding Study.* Cognitive Science: Volume 48, Issue 3, e13427. https://onlinelibrary.wiley.com/doi/full/10.1111/cogs.13427

Dupoux, E. (2018, February 14). *Cognitive Science in the era of Artificial Intelligence: A roadmap for reverse-engineering the infant language-learner.* From Arxiv, Cornell University: https://arxiv.org/abs/1607.08723

Synnaeve, G., Xu, Q, X., Kahn, J., Likhomanenko, T., Grave, E., Pratap, V., Sriram, A., Liptchinsky, V., & Collobert, R., (2020, July 15). *End-to-end ASR: from Supervised to Semi-Supervised Learning with Modern Architectures.* From Arxiv, Cornell University: https://arxiv.org/abs/1911.08460

Inaguma, H., Dalmia, S., Yan, B., Wantanabe, S., (2021, september 27). FAST-MD: FAST MULTI-DECODER END-TO-END SPEECH TRANSLATION WITH NON-AUTOREGRESSIVE HIDDEN INTERMEDIATES. Kyoto University, Japan. https://arxiv.org/pdf/2109.12804v1

Kürzinger, L., Winkelbauer, D., Li, L., Watzel, T., & Rigoll, G., (2020, september 29). *CTC-Segmentation of Large Corpora for German End-to-End Speech Recognition.* From Springer Nature Link: https://link.springer.com/chapter/10.1007/978-3-030-60276-5\_27

Kürzinger, L., Winkelbauer, D., Li, L., Watzel, T., Kamo, N. & Rigoll, G., (2022, february 2). *CTC - segmentation*. Github: https://github.com/lumaku/ctc-segmentation

Caines, A., Altmann-Richer, E., & Buttery, P., (2019, october 11). The cross-linguistic performance of word segmentation models over time. Journal of Child Language: Volume 46, Issue 6. Cambridge, United Kingdom, Cambridge University Press. https://www.cambridge.org/core/journals/journal-of-child-language/article/abs/crosslinguistic-performance-of-word-segmentation-models-over-time/8225EE295600DDB56F56F3C953C4F652?utm\_source=chatgpt.com