2M-BELEBELE: Highly Multilingual Speech and American Sign Language Comprehension Dataset

Anonymous ACL submission

Abstract

We introduce the first highly multilingual speech and American Sign Language (ASL) comprehension dataset by extending BELEBELE. Our dataset covers 74 spoken languages at the intersection of BELEBELE and FLEURS, and one sign language (ASL). We evaluate 2M-BELEBELE dataset for both 5-shot and zero-shot settings and across languages, the speech comprehension accuracy is $\approx$ 10% average lower compared to reading comprehension.

1 Introduction

From an AI perspective, text understanding and generation services are used globally in more than a hundred languages, but the scarcity of labeled data poses a significant challenge to developing functional systems in most languages. Although natural language processing (NLP) datasets with extensive language coverage, such as FLORES-200 (NLLBTeam, 2024), are available, they mainly concentrate on machine translation (MT). Multilingual evaluation benchmarks such as those for multilingual question answering (Lewis et al., 2020; Clark et al., 2020), natural language inference (Conneau et al., 2018), summarization (Hasan et al., 2021; Ladhak et al., 2020), and reasoning datasets (Ponti et al., 2020; Lin et al., 2021) collectively cover only about 30 languages. Furthermore, the extension of such datasets to speech or American Sign Language (ASL) is lacking, with the exception of FLEURS (Conneau et al., 2022; Tanzer, 2024), which is based on FLORES-200.

The recent BELEBELE benchmark is the first corpus that addresses text reading comprehension for a large number of languages following a multi-way parallel approach (Bandarkar et al., 2023). The entire BELEBELE text statistics are summarized in Table 3 in Appendix A. Currently, there are no highly multilingual evaluation datasets for natural language understanding that cover either both speech and text, or ASL.

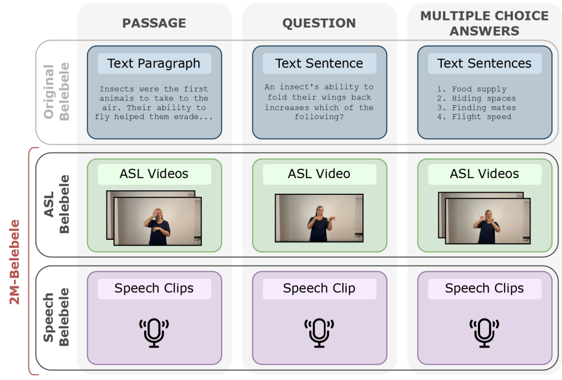

Figure 1: 2M-BELEBELE entry: beyond passage, question and multiple choice answers in text from BELEBELE, we extend to ASL and 74 speech languages.

In this work, we extend the BELEBELE dataset to speech and sign (Section 3). By doing so, we create the first highly multilingual speech and sign comprehension dataset: 2M-BELEBELE, which is composed of human speech recordings covering 74 languages and human sign recordings for ASL.

As a by-product of 2M-BELEBELE, we also extend the FLEURS dataset (which is widely used to benchmark language identification and ASR) by providing recordings for more FLORES-200 sentences than were previously available and adding sign language, creating a new 2M-FLORES. This 2M-FLORES extends FLEURS by 20%.

Finally, we provide a very basic set of experiments that evaluate 2M-BELEBELE and provide some reference results. We use direct and/or cascaded systems to evaluate 2M-BELEBELE dataset (Section 4). We also list several further experimentation that 2M-BELEBELE unblocks. Note that the main contribution of this paper is the creation of the first highly multilingual speech and sign comprehension dataset. The complete set of experiments is out of the scope of this paper (Limitations). By open-sourcing our dataset, we encourage the scientific community to pursue such experimentation.

2 Related Work

Speech Comprehension The outstanding performance of some MT and text-to-speech (TTS) models has enabled a rise in the number of works using synthetically generated training data. Furthermore, some recent works propose to also use synthetic data for evaluation; e.g., (Üstün et al., 2024; SEAMLESSCommunicationTeam, 2025; Nguyen et al., 2024; Nachmani et al., 2023). This strategy allows researchers to extend datasets to low-resource languages and to other modalities, such as speech. However, we prove that using synthetic data for evaluation does not provide comparable conclusions as relying on human speech for the particular task of automatic speech recognition (ASR) and the FLEURS domain (Appendix E). The evaluation dataset that is closest to the speech comprehension evaluation dataset presented in this paper is the generative QA dataset proposed in (Nachmani et al., 2023). The dataset covers 300 questions in English.

ASL Comprehension Compared to spoken languages, sign languages are considered low-resource languages for natural language processing (Yin et al., 2021). Most popular datasets cover small domains of discourse; e.g., weather broadcasts (Camgoz et al., 2018), which has limited real world applications. There have been previous releases of large scale open domain sign language datasets; e.g., (Albanie et al., 2021; Shi et al., 2022; Uthus et al., 2024). However, the results and challenges on such datasets suggest that computational sign language research still requires additional datasets to reach the performance of their spoken language counterparts (Müller et al., 2022, 2023). With the release of the ASL extension of the BELEBELE dataset, we aim to provide additional, high quality sign language data with gloss annotations to underpin further computational sign language research. Furthermore, due to the paragraph-level nature of the BELEBELE dataset, we enable paragraph-context sign language translation, which has been reported to improve translation performance (Sincan et al., 2023).

3 2M-BELEBELE

FLEURS and BELEBELE passage alignment. Since BELEBELE uses passages constructed from sentences in the FLORES-200 dataset, and FLEURS (Conneau et al., 2022) is a human speech version of FLORES-200 for a subset of its languages, we create a speech version of BELEBELE by aligning its passages with the speech segments available in FLEURS. This extension can be done without extra human annotation, just by computing the alignment between FLEURS and BELEBELE passages. However, such alignment does not cover the entire BELEBELE corpus because FLEURS does not cover the entirety of FLORES-200. There are 74 languages shared between FLEURS and BELEBELE. FLEURS does not cover the same passages as BELEBELE in all those 74 languages, which means that some languages have more speech passages than others. In general, we are able to match almost ≈ 80% of the passages. Figure 2 shows the number of FLEURS paragraphs we can match, thus obtaining the number of paragraphs that must be recorded in order to cover all passages BELEBELE.

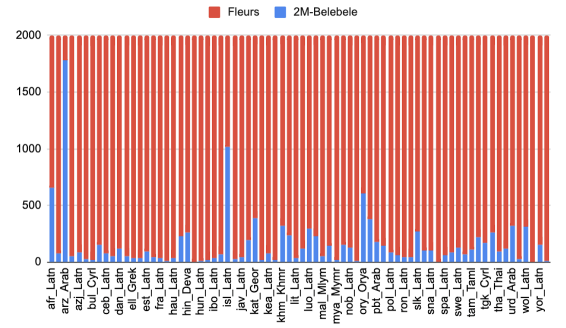

Figure 2: FLEURS vs New Recordings from 2M-BELEBELE for sentences in passages.

Speech recordings. We commission human recordings for the part of the BELEBELE dataset that is not covered by existing FLEURS recordings, as well as for elements of BELEBELE that do not exist in FLEURS (i.e. questions and answers). Recording participants must be native speakers of the languages they record. They must have an impeccable grasp of the conventions used in their respective languages for the narration of texts. The three tasks that participants are asked to perform are: (1) Read aloud and record the text passages provided (from FLORES-200); (2) Read aloud and record the provided written questions; (3) Read aloud and record the provided written answers. For the task, we provide the participants with (a) the text of the sentences to be recorded in TSV format (the number of passages may differ from language to language), (b) the written questions (900 per language, and (c) the written answer options (3,600 per language). Additional details on the recording guidelines provided to annotators are reported in the appendix B. We verify the quality of the recordings by randomly selecting 270 recordings (30% of sample size) and ensuring that the recordings do not contain background or ambient noise and that the voices of the participants are clearly audible.

Sign recordings. To obtain ASL sign recordings, we provide translators of ASL and native signers with the English text version of the sentences to be recorded. The interpreters are then asked to translate these sentences into ASL, create glosses for all sentences, and record their interpretations into ASL one sentence at a time. The glosses are subjected to an additional quality check by expert annotators to harmonize the glossing format. To harmonize the recording conditions and eliminate visual bias, the videos are recorded against plain monochrome backgrounds (e.g., white or green), and signers are requested to wear monochrome upper body clothing (e.g., black). All videos are captured in 1920x1080p resolution with all of the signing space covered in FOV. The recordings are done in 60 frames per second to address most of the motion blur that happens during signing.

2M-BELEBELE Statistics. The final dataset is composed of 75 languages (74 in speech, 1 in sign). Each of the languages' respective subsets includes 2,000 utterances organized in 488 distinct passages, 900 questions, and 4 multiple choice answers per question. For our recorded data (the red portion of Figure 2 plus questions and answers), we have one audio file or two per sentence, depending on the number of available participants (one participant only in 23 languages, and two participants in 51 languages). When two speakers are available, we request that one should represent a higher-pitch range, and the other a lower-pitch range for each passage. More details are available in Appendix A.

In addition, the data set includes video recordings in ASL for 2,000 FLORES sentences (not including the test partition) and is similarly organized in 488 distinct passages, as well as 900 questions and 4 multiple-choice answers for each question (see summary table 3). The ASL dataset was recorded by two interpreters, but, contrary to what was possible in other languages, each interpreter could only cover one-half of the dataset each.

4 Experiments

We evaluate 2M-BELEBELE, and compare performance across modalities. Our comparison is limited in number of systems and combination of modalities. 2M-BELEBELE offers the opportunity to check multimodal comprehension by combining speech/text/sign passages; questions and answers. In our case, we only provide results for entire text passages, questions and answers and speech passages, text questions and answers. A more comprehensive set of experiments is out of the scope of this paper, which aims at unblocking such experimentation by open-sourcing the dataset itself.

Systems. We use the speech section of the 2M-BELEBELE dataset to evaluate the speech comprehension task with a cascaded system consisting of first speech recognition (ASR) using the WHISPER-LARGE-V3 model (Radford et al., 2022) (hereinafter, WHISPER) and SEAMLESSM4T (corresponding to SEAMLESSM4T-LARGE V2) model (SEAMLESSCommunicationTeam, 2025) feeding into LLAMA-3[1]. We also provide results with a unified system SPIRITLM (Nguyen et al., 2024), which is a multimodal language model that freely mixes text and speech. Since the size of this model is 7B and is based on LLAMA-2, we also add a comparison to the LLAMA-2 model. We compare these results with LLAMA-3 and LLAMA-3-CHAT using the BELEBELE text passage as input. For these systems, we report the results in 5-shot in-context learning and zero-shot on 59 at the intersection of WHISPER and 2M-BELEBELE and 39 languages at the intersection of WHISPER, SEAMLESSM4T and 2M-BELEBELE (see Appendix A).

Zero-shot Evaluation. We use the same evaluation strategy as in (Bandarkar et al., 2023). SPIRITLM is not available in chat mode.

5-shot In-Context Learning. The few-shot examples are taken randomly from the English training set and they are prompted as *text* format to the model. Different from (Bandarkar et al., 2023), we do not pick the answer with the highest probability but directly assess the predicted letter of the answer. For 5-shot and zero-shot settings, our instruction prompt is as follows *"Given the following passage, query, and answer choices, output the letter corresponding to the correct answer. Do not write any explanation. Only output the letter within A, B, C,*

[1] https://ai.meta.com/blog/meta-llama-3/

Dataset	Model	Size	Vocab	#Lang	AVG	% $\geq$ 50	% $\geq$ 70	Eng	non-Eng
5-Shot In-Context Learning (examples in English)									
BELEBELE	LLAMA-3	70B	128K	59	85.4	96.6	94.9	94.8	85.2
2M-BELEBELE	WHISPER + LLAMA-3	70B	128K	59	77.4	88.1	72.9	94.4	77.1
BELEBELE	LLAMA-3	70B	128K	39	84.9	97.4	94.9	94.8	84.7
2M-BELEBELE	WHISPER + LLAMA-3	70B	128K	39	77.1	89.7	71.8	94.4	76.6
2M-BELEBELE	SEAMLESSM4T + LLAMA-3	70B	128K	39	81.7	94.9	92.7	93.5	81.4
2M-BELEBELE	WHISPER + LLAMA-2	7B	32K	1	-	-	-	49.9	-
2M-BELEBELE	SPIRITLM	7B	37K	1	-	-	-	25.9	-
Zero-Shot									
BELEBELE	LLAMA-3-CHAT	70B	128K	59	87.5	98.3	96.6	95.8	87.3
2M-BELEBELE	WHISPER + LLAMA-3-CHAT	70B	128K	59	79.4	93.2	78.0	95.7	79.2
BELEBELE	LLAMA-3-CHAT	70B	128K	39	87.0	97.4	94.9	95.8	86.7
2M-BELEBELE	WHISPER + LLAMA-3-CHAT	70B	128K	39	79.1	92.3	76.9	95.7	78.7
2M-BELEBELE	SEAMLESSM4T + LLAMA-3-CHAT	70B	128K	39	84.8	94.9	94.9	95.5	84.5

Table 1: Summary of accuracy results on 2M-BELEBELE compared to BELEBELE across models and evaluation settings. % $\geq$ 50/70 refers to the proportion of languages for which a given model performs above 50/70% for question and answer in text and passage in speech.

or D that corresponds to the correct answer." and we report the averaged accuracy over 3 runs[2].

Results. Table 1 reports the summary of the results at the intersection of languages between system availability (either 59 or 39 as reported in detail in Table 2). The English drop from direct text to speech task does not vary much between 5-shot and zero-shot strategies, being slightly higher in the zero-shot setting (coherently with previous LLAMA-3 results that show better performance in zero-shot in other tasks[3]). When comparing speech and text comprehension, we observe that speech decreases performance in about 10% when comparing for 59 languages (using WHISPER for ASR). However, this decrease shortens (to about 2-3% average) when comparing for 39 languages (using SEAMLESSM4T for ASR). 2M-BELEBELE accuracy results per language compared to BELEBELE are shown in Figure 3 in Appendix D. Differences in speech and text vary slightly depending on the languages. Low-resource languages have a greater variation between text and speech BELEBELE. The ten languages with the largest gap are: Burmese, Maltese, Assamese, Mongolian, Southern Pashto, Sindhi, Telugu, Javanese, Tajik, Georgian.

Additionally, Table 1 reports English results for SPIRITLM, a direct multimodal model. One of the reasons SPIRITLM may be performing worse is that 5-shot examples are in text, while the passage on the asked question is in speech. Best results in average for speech comprehension are achieved with the SEAMLESSM4T + LLAMA-3 cascade.

ASL We know from previous large-scale translation attempts (Albanie et al., 2021; Müller et al., 2022) that models struggle to generalize over both individuals/appearance and large domain of discourse. Compared to speech and text models, sign language models suffer from having to learn generalized representations from high-dimensional inputs, i.e. videos, without overfitting to limited training dataset. Previous attempts have been made to create a more generalizable abstraction layer in the form of subunits (Camgoz et al., 2020), similar to phonemes for speech, which achieved promising results on a translation task with a small discourse domain. However, this work is yet to be applied to large discourse domain translation tasks. The best results in the FLORES domain have been achieved with close models that are not available (Zhang et al., 2024). Trying (Rust et al., 2024) as an open model did not perform above chance in the final reading comprehension dataset. However, we believe that the release of this new dataset with the additional gloss annotation will help in training models that generalize over individuals better and improve large-scale sign language translation.

5 Conclusions

The 2M-BELEBELE dataset[4] allows to evaluate natural language comprehension in a large number of languages, including ASL. 2M-BELEBELE is purely human-made and covers the BELEBELE passages, questions, and answers for 74 languages in the speech modality and ASL. As a by-product, 2M-FLORES extends FLEURS by 20% [5]

[2]Random seeds: 0, 1, 2.

[3]https://ai.meta.com/blog/meta-llama-3-1/ and https://ai.meta.com/blog/meta-llama-3/

[4]2M-BELEBELE dataset is freely available in BLIND

[5]2M-FLORES is freely available in BLIND

Limitations and ethical considerations

Our speech annotations do not have the entire set completed with two annotators. Due to the high volume of the dataset, not every recording has been thoroughly verified. Some of the languages in 2M-BELEBELE are low-resource languages, which pose a challenge in sourcing professionals to record. Therefore, some of the audios were recorded in home settings and may contain minor background noise, static noise, echoes, and, occasionally, the speech could be slightly muffled or soft. All annotators are native speakers of the target language, but they may have regional accents in their speech, and their personal speech styles may be present in the audio as well. However, the mentioned imperfections should not affect intelligibility; all the recordings can be clearly understood by human standards. Note that we are planning to release more languages as reported in Appendix C.

We can group the ASL limitations under two categories, namely visual and linguistic. For visual limitations, ASL sequences are recorded in what can be considered laboratory environments with few signer variance. This makes it harder for models trained on them to generalize to unseen environments and signers. For linguistic limitations, ASL sequences are collected one sentence at a time. Although this enables pairwise training and evaluation, such as classical text-based NMT, the generated sequences may not be fully realistic in terms of real-world signing. An example would be the use of placement. In sentence-per-sentence sequence generation, a signer would refer to an entity with their sign each sentence, whereas in long-form conversation, a signer would place the entity in their signing space after first reference and refer them in via use of placement in the following sentences.

Our benchmarking is limited compared to the potential capabilities of the dataset. For example, since we have spoken questions, passages and responses, instead of just using a fix modality (spoken passages, text questions and responses), we could explore the performance when using all combinations among modalities (e.g., question in speech, answer in speech, passage in speech; or question in speech, answer in text, passage in speech; or question in speech, answer in speech and passage in text.)

In terms of compute budget, we estimate it as 47K Nvidia A100 hours by taking into account the product of following factors: number of languages (59 / 39), number of random seeds (3), number of GPUs required by model (8), number of experiment setups (5) and estimated number of hours per experiment (10).

Speakers and signers were paid a fair rate. Our recorded data reports self-identified gender by participant. Each of the speakers and signers signed a consent form agreeing on the dataset and its usage that they were participating in.

References

Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, et al. 2021. Bbc-oxford british sign language dataset. *arXiv preprint arXiv:2111.03635*.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2023. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. *Preprint*, arXiv:2308.16884.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Multi-channel transformers for multi-articulatory sign language translation. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 301–319. Springer.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2022. Fleurs: Few-shot learning evaluation of universal representations of speech. *Preprint*, arXiv:2205.12446.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, Online. Association for Computational Linguistics.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren. 2021. Common sense beyond English: Evaluating and improving multilingual language models for commonsense reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1274–1287, Online. Association for Computational Linguistics.

Mathias Müller, Malihe Alikhani, Eleftherios Avramidis, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Sarah Ebling, Cristina España-Bonet, Anne Göhring, Roman Grundkiewicz, et al. 2023. Findings of the second wmt shared task on sign language translation (wmt-slt23). In *Proceedings of the Eighth Conference on Machine Translation (WMT23)*, pages 68–94.

Mathias Müller, Sarah Ebling, Eleftherios Avramidis, Alessia Battisti, Michèle Berger, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Cristina España-Bonet, Roman Grundkiewicz, et al. 2022. Findings of the first wmt shared task on sign language translation (wmt-slt22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 744–772.

Eliya Nachmani, Alon Levkovitch, Roy Hirsch, Julian Salazar, Chulayuth Asawaroengchai, Soroosh Mariooryad, Ehud Rivlin, RJ Skerry-Ryan, and Michelle Tadmor Ramanovich. 2023. Spoken question answering and speech continuation using spectrogram-powered llm. *Preprint*, arXiv:2305.15255.

Tu Anh Nguyen, Benjamin Muller, Bokai Yu, Marta R. Costa-jussa, Maha Elbayad, Sravya Popuri, Paul-Ambroise Duquenne, Robin Algayres, Ruslan Mavlyutov, Itai Gat, Gabriel Synnaeve, Juan Pino, Benoit Sagot, and Emmanuel Dupoux. 2024. Spirit-lm: Interleaved spoken and written language model. *Preprint*, arXiv:2402.05755.

NLLBTeam. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630:841–846.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2024. Scaling speech technology to 1,000+ languages.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Phillip Rust, Bowen Shi, Skyler Wang, Necati Cihan Camgoz, and Jean Maillard. 2024. Towards privacy-aware sign language translation at scale. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8624–8641, Bangkok, Thailand. Association for Computational Linguistics.

SEAMLESSCommunicationTeam. 2025. Joint speech and text machine translation for up to 100 languages. *Nature*, 637:587–593.

Bowen Shi, Diane Brentari, Greg Shakhnarovich, and Karen Livescu. 2022. Open-domain sign language translation learned from online video. *arXiv preprint arXiv:2205.12870*.

Ozge Mercanoglu Sincan, Necati Cihan Camgoz, and Richard Bowden. 2023. Is context all you need? scaling neural sign language translation to large domains of discourse. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1955–1965.

Garrett Tanzer. 2024. Fleurs-asl: Including american sign language in massively multilingual multitask evaluation. *Preprint*, arXiv:2408.13585.

Dave Uthus, Garrett Tanzer, and Manfred Georg. 2024. Youtube-asl: A large-scale, open-domain american sign language-english parallel corpus. *Advances in Neural Information Processing Systems*, 36.

Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. Including signed languages in natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7347–7360.

Biao Zhang, Garrett Tanzer, and Orhan Firat. 2024. Scaling sign language translation. *Preprint*, arXiv:2407.11855.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. *Preprint*, arXiv:2402.07827.

A Languages

Table 2 reports details on languages covered by FLEURS, TTS and ASR.

B Annotation Guidelines

Recording process. Find a quiet place free from distractions and noises, and choose a headphone that is comfortable to wear and a good quality microphone that will not distort or break your voice. Read aloud and record the scripts in a pleasant tone and at a constant and even pace, as if you were reading a formal document. Try not to speak too quickly or slowly and aim for a natural pace that is easy to follow. The audio files below provide examples of paces that are expected, too fast, or too slow, for the sentence. The hearing also marks the date for the suspect's right to a rapid trial.

To achieve the best sound quality when recording, position the microphone close to your mouth so that the voice will sound clear and present, but not too close that it sounds muddy or you can hear a puff of air. Clearly enunciate the words and avoid mumbling. Be sure to provide a 2-second pause between sentences to add clarity and keep the overall pace down. When dealing with long, complicated sentences that contain multiple clauses or phrases, there are several approaches to ensure clarity and a natural flow as follows. Break it down: Separate the sentence into smaller parts or clauses. Practice reading aloud several times before starting the recording. This can help you get a feel for the rhythm and pacing of the sentence. Pace yourself: Try to maintain a steady, even pace. If the sentence is particularly long, it is possible to take a brief pause at a natural breakpoint to catch your breath. You should read the provided passages aloud without repairs (a repair is the repetition of a word that was incorrectly pronounced to correct its pronunciation).

To achieve this, familiarize yourself beforehand with the correct pronunciation of difficult words, proper nouns, and transliterated words, as well as signs and symbols, dates and times, numbers, abbreviations, and punctuation marks. Some elements may have more than one correct pronunciation. In this case, use the one that comes the more naturally to you, as long as it is an accepted pronunciation (i.e., it is acknowledged in your language's dictionaries). Practice reading the passages aloud several times to become more comfortable with the material. Please pay particular attention to the following items:

Numbers. Number formats can vary from language to language; it is important to follow the pronunciation rules in your language. Here are some general guidelines and examples: Decimal numbers: Read the whole part of the number as a whole number and then individually read every number after the decimal point. For example, in English, the decimal number 3.14 should be read as "three point one four." Different languages may have different rules, and you should follow the rules that are appropriate for your language. Cardinal numbers represent quantities or amounts. Ordinal numbers represent positions or ranks in sequential order and should be read with the appropriate suffix. For example, in English, the ordinal number 1st is read "first" (not "onest") and 5th is read "fifth" (not "fiveth"). Different languages may have different rules, and you should follow the rule that is appropriate for your language.

Roman numerals are a collection of seven symbols that each represent a value: I = 1, V = 5, X = 10, L = 50, C = 100, D = 500, and M = 1,000. The can be pronounced in slightly different ways depending on the context, but they are never pronounced as individual letters. For example, in English, VIII in Henry VIII is pronounced "Henry the eighth", while Superbowl LVIII is pronounced "Superbowl fifty-eight", but they are never pronounced "Henry v i i i" or "Superbowl l v i i i". Different languages may have different rules, and you should follow the rules that are appropriate for your language. Punctuation marks: As a general rule, punctuation marks should not be pronounced, except quotation marks.

For example, in English, punctuation marks such as periods, commas, colons, semicolons, question marks, and exclamation points are typically not pronounced. For example, the sentence. As a result of this, a big scandal arose. will be pronounced "As a result of this a big scandal arose" - not "As

Language	**Code**	**Script**	**Family**	FLEURS	SeamlessM4T	Whisper	2M-BELEBELE
Mesopotamian Arabic	acm_Arab	Arab	Afro-Asiatic				
Afrikaans	afr_Latn	Latn	Indo-European	✓		✓	✓(1)
Tosk Albanian	als_Latn	Latn	Indo-European				
Amharic	amh_Ethi	Ethi	Afro-Asiatic	✓			✓(2)
North Levantine Arabic	apc_Arab	Arab	Afro-Asiatic				
Modern Standard Arabic	arb_Arab	Arab	Afro-Asiatic				
Modern Standard Arabic	arb_Latn	Latn	Afro-Asiatic				
Najdi Arabic	ars_Arab	Arab	Afro-Asiatic				
Moroccan Arabic	ary_Arab	Arab	Afro-Asiatic				
Egyptian Arabic	arz_Arab	Arab	Afro-Asiatic	✓		✓	✓(2)
Assamese	asm_Beng	Beng	Indo-European	✓	✓	✓	✓(2)
North Azerbaijani	azj_Latn	Latn	Turkic	✓			✓(1)
Bambara	bam_Latn	Latn	Niger-Congo				
Bengali	ben_Beng	Beng	Indo-European	✓	✓	✓	✓(2)
Bengali	ben_Latn	Latn	Indo-European				
Standard Tibetan	bod_Tibt	Tibt	Sino-Tibetan				
Bulgarian	bul_Cyrl	Cyrl	Indo-European	✓	✓	✓	✓(2)
Catalan	cat_Latn	Latn	Indo-European	✓	✓	✓	✓(2)
Cebuano	ceb_Latn	Latn	Austronesian	✓			✓(1)
Czech	ces_Latn	Latn	Indo-European	✓		✓	✓(2)
Central Kurdish	ckb_Arab	Arab	Indo-European	✓			
Danish	dan_Latn	Latn	Indo-European	✓		✓	✓(2)
German	deu_Latn	Latn	Indo-European	✓	✓	✓	✓(2)
Greek	ell_Grek	Grek	Indo-European	✓	✓	✓	✓(2)
English	eng_Latn	Latn	Indo-European	✓	✓	✓	✓(2)
Estonian	est_Latn	Latn	Uralic	✓		✓	✓(1)
Basque	eus_Latn	Latn	Basque				
Finnish	fin_Latn	Latn	Uralic	✓	✓	✓	✓(2)
French	fra_Latn	Latn	Indo-European	✓	✓	✓	✓(2)
Fulfulde (Nigerian)	fuv_Latn	Latn	Atlantic-Congo				
Oromo (West Central)	gaz_Latn	Latn	Afro-Asiatic	(✓)			
Guarani	grn_Latn	Latn	Tupian				
Gujarati	guj_Gujr	Gujr	Indo-European	✓	✓	✓	✓(1)
Haitian Creole	hat_Latn	Latn	Indo-European				
Hausa	hau_Latn	Latn	Afro-Asiatic	✓	(✓)		✓(2)
Hebrew	heb_Hebr	Hebr	Afro-Asiatic	✓	✓	✓	✓(2)
Hindi	hin_Deva	Deva	Indo-European	✓	✓	✓	✓(2)
Hindi	hin_Latn	Latn	Indo-European				
Croatian	hrv_Latn	Latn	Indo-European	✓			✓(2)
Hungarian	hun_Latn	Latn	Uralic	✓	✓	✓	✓(2)
Armenian	hye_Armn	Armn	Indo-European	✓		✓	✓(1)
Igbo	ibo_Latn	Latn	Atlantic-Congo	✓			✓(1)
Ilocano	ilo_Latn	Latn	Austronesian				
Indonesian	ind_Latn	Latn	Austronesian	✓	✓	✓	✓(2)
Icelandic	isl_Latn	Latn	Indo-European	✓	✓	✓	✓(1)
Italian	ita_Latn	Latn	Indo-European	✓		✓	✓(2)
Javanese	jav_Latn	Latn	Austronesian	✓	✓	✓	✓(1)
Japanese	jpn_Jpan	Jpan	Japonic	✓		✓	✓(2)
Jingpho	kac_Latn	Latn	Sino-Tibetan				
Kannada	kan_Knda	Knda	Dravidian	✓			
Georgian	kat_Geor	Geor	Kartvelian	✓		✓	✓(2)
Kazakh	kaz_Cyrl	Cyrl	Turkic	✓	✓	✓	✓(1)
Kabuverdianu	kea_Latn	Latn	Indo-European	✓			✓(1)

Language	**Code**	**Script**	**Family**	FLEURS	SeamlessM4T	Whisper	2M-BELEBELE
Mongolian	khk_Cyrl	Cyrl	Mongolic	(✓)		✓	✓(2)
Khmer	khm_Khmr	Khmr	Austroasiatic	✓			✓(1)
Kinyarwanda	kin_Latn	Latn	Atlantic-Congo				
Kyrgyz	kir_Cyrl	Cyrl	Turkic	✓			
Korean	kor_Hang	Hang	Koreanic	✓	✓	✓	✓(1)
Lao	lao_Laoo	Laoo	Kra-Dai	✓			
Lingala	lin_Latn	Latn	Niger-Congo	✓			
Lithuanian	lit_Latn	Latn	Indo-European	✓		✓	✓(2)
Ganda	lug_Latn	Latn	Atlantic-Congo	✓			✓(1)
Luo	luo_Latn	Latn	Atlantic-Congo	✓			✓(2)
Standard Latvian	lvs_Latn	Latn	Indo-European	(✓)		✓	✓(2)
Malayam	mal_Mlym	Mlym	Dravidian	✓	✓	✓	✓(2)
Marathi	mar_Deva	Deva	Indo-European	✓			
Macedonian	mkd_Cyrl	Cyrl	Indo-European	✓		✓	✓(2)
Maltese	mlt_Latn	Latn	Afro-Asiatic	✓			
Maori	mri_Latn	Latn	Austronesian	✓			
Burmese	mya_Mymr	Mymr	Sino-Tibetan	✓	✓	✓	✓(2)
Dutch	nld_Latn	Latn	Indo-European	✓	✓	✓	✓(2)
Norwegian Bokmål	nob_Latn	Latn	Indo-European	✓			✓(2)
Nepali	npi_Deva	Deva	Indo-European	✓		✓	✓(2)
Nepali	npi_Latn	Latn	Indo-European				
Northern Sotho	nso_Latn	Latn	Atlantic-Congo	✓			
Nyanja	nya_Latn	Latn	Afro-Asiatic	✓			
Odia	ory_Orya	Orya	Indo-European	✓			✓(1)
Eastern Panjabi	pan_Guru	Guru	Indo-European	✓	✓	✓	✓(2)
Southern Pashto	pbt_Arab	Arab	Indo-European	(✓)		✓	✓(1)
Western Persian	pes_Arab	Arab	Indo-European	(✓)		✓	✓(1)
Plateau Malagasy	plt_Latn	Latn	Austronesian				
Polish	pol_Latn	Latn	Indo-European	✓	✓	✓	✓(2)
Portuguese	por_Latn	Latn	Indo-European	✓	✓	✓	✓(2)
Romanian	ron_Latn	Latn	Indo-European	✓	✓	✓	✓(2)
Russian	rus_Cyrl	Cyrl	Indo-European	✓	✓	✓	✓(2)
Shan	shn_Mymr	Mymr	Tai-Kadai				
Sinhala	sin_Latn	Latn	Indo-European				
Sinhala	sin_Sinh	Sinh	Indo-European				
Slovak	slk_Latn	Latn	Indo-European	✓		✓	✓(1)
Slovenian	slv_Latn	Latn	Indo-European	✓		✓	✓(2)
Shona	sna_Latn	Latn	Atlantic-Congo	✓	✓	✓	✓(2)
Sindhi	snd_Arab	Arab	Indo-European	✓		✓	✓(2)
Somali	som_Latn	Latn	Afro-Asiatic	✓			
Southern Sotho	sot_Latn	Latn	Atlantic-Congo				
Spanish	spa_Latn	Latn	Indo-European	✓	✓	✓	✓(2)
Serbian	srp_Cyrl	Cyrl	Indo-European	✓		✓	✓(2)
Swati	ssw_Latn	Latn	Atlantic-Congo				
Sundanese	sun_Latn	Latn	Austronesian				
Swedish	swe_Latn	Latn	Indo-European	✓	✓	✓	✓(2)
Swahili	swh_Latn	Latn	Atlantic-Congo	✓	✓	✓	✓(1)
Tamil	tam_Taml	Taml	Dravidian	✓	✓	✓	✓(2)
Telugu	tel_Telu	Telu	Dravidian	✓	✓	✓	✓(2)
Tajik	tgk_Cyrl	Cyrl	Indo-European	✓	✓	✓	✓(1)
Tagalog	tgl_Latn	Latn	Austronesian	(✓)	✓	✓	✓(2)
Thai	tha_Thai	Thai	Tai-Kadai	✓	✓	✓	✓(2)
Tigrinya	tir_Ethi	Ethi	Afro-Asiatic				
Tswana	tsn_Latn	Latn	Atlantic-Congo				

Language	Code	Script	Family	FLEURS	SeamlessM4T	Whisper	2M-BELEBELE
Tsonga	tso_Latn	Latn	Afro-Asiatic				
Tsonga	tso_Latn	Latn	Afro-Asiatic				
Turkish	tur_Latn	Latn	Turkic	✓	✓	✓	✓(1)
Ukranian	ukr_Cyrl	Cyrl	Indo-European	✓			
Urdu	urd_Arab	Arab	Indo-European	✓	✓	✓	✓(2)
Urdu	urd_Latn	Latn	Indo-European				
Northen Uzbek	uzn_Latn	Latn	Turkic	✓			
Vietnamese	vie_Latn	Latn	Austroasiatic	✓	✓	✓	✓(2)
Waray	war_Latn	Latn	Austronesian				
Wolof	wol_Latn	Latn	Atlantic-Congo	✓			✓(1)
Xhosa	xho_Latn	Latn	Atlantic-Congo	✓			✓(1)
Yoruba	yor_Latn	Latn	Atlantic-Congo	✓	✓	✓	✓(2)
Chinese	zho_Hans	Hans	Sino-Tibetan	✓			✓(2)
Chinese	zho_Hant	Hant	Sino-Tibetan	(✓)			
Standard Malay	zsm_Latn	Latn	Austronesian	(✓)			
Zulu	zul_Latn	Latn	Atlantic-Congo	✓			
American Sign Language	ase	-	Sign Language				✓(2)

Table 2: Languages details. Column FLEURS reports the languages covered by Speech BELEBELE v1. Column ASR shows the languages reported in the experiment section, note that Hausa is covered by WHISPER-LARGE-V3 but not for SEAMLESSM4T. The number in brackets shows the number of annotations per language.

Passages		Questions/Answers	
Distinct Passages	488	Distinct Q	900
Questions per passage	1-2	Multiple-choice A	4
Avg words (std)	79.1 (26.2)	Avg words Q (std)	12.9 (4.0)
Avg sentences (std)	4.1 (1.4)	Avg words A (std)	4.2 (2.9)

Table 3: Statistics for 2M-BELEBELE, which covers 74 spoken languages plus ASL. Average words are computed for English.

a result of this comma a big scandal arose period”. However, in formal-register English (in the news, for example), a difference is made between content created by the news team and content that should be attributed to someone else by explicitly pronouncing quotation marks. For example, the news transcript The fighter said: ”I am here to try to win this.” will be pronounced: ”The fighter said, quote, I am here to try to win this. End of quote.” In this case, different languages may have different rules, and you should follow the rules that are appropriate for your language. Signs and symbols. Signs and symbols need to be pronounced as they would be heard in a speech-only setting. Attention should be paid: (a) to potential number or gender agreement (for example, in English, ”40%” should be read as ”forty percent” — not ”forty percents”) (b) to potential differences between the place of the sign or symbol in writing and in speech (for example, in English, the ”$” sign should be read as ”dollar” and should be read after the number it precedes; i.e. ”$22” should be read as ”twenty-two dollars” — not ”dollars twenty-two”) (c) to the way the sign or symbol gets expanded in speech (for example, in English, ”Platform 9 ¾” should be read ”platform nine and three quarters” — not ”platform nine three quarters”). Similarly, 50 km/h would be pronounced ”fifty kilometers per hour” — not ”fifty kilometers hour”). Different languages may have different rules, and you should follow the rules that are appropriate for your language.

Proper nouns and foreign expressions. Even the same language may have at least 2 different ways to pronounce foreign expressions of proper nouns: (a) one way is to try to approach the way they would sound in the foreign language from which they come (for example, in English, Louis in Louis XIV is pronounced ”lewee” as it would be in French); (b) the other way is to pronounce them according to the rules of the adopting language (for example, in English, Louis in the City of St Louis is pronounced as in the English proper noun ”Lewis”)

Abbreviations. Abbreviations should be expanded as much as possible. However, it is suggested to refrain from expanding them if their expansion results in unnatural speech. For example, in English, abbreviations such as Dr. or etc. are pronounced ”doctor” and ”et cetera”, respectively (not ”d r” nor ”e t c”). However, abbreviations such as AM or PhD are pronounced as a sequence of letters without being expanded (”a m” and ”p h d”, respectively - not ”ante meridiem” nor ”philosophy doctorate"). Different languages may have different conventions, and you should follow the conventions that are appropriate for your language.

C Extra languages pending for collection

We plan to collect in total 91 languages with both high-pitched and low-pitched. This is the list of all the languages in planning.

- Central Kurdish
- Nigerian Fulfulde
- West Central Oromo
- Kannada
- Kyrgyz
- Lao
- Lingala
- Marathi
- Maltese
- Maori
- Northern Sotho
- Chewa
- Somali
- Ukrainian
- Northern Uzbek
- Malay
- Zulu

D Detailed results per Language

E Ablation study: Synthetic extension in speech evaluation datasets

In this part of our work, we aim to analyze the feasibility of synthetically extending text benchmarks to speech using TTS systems, thereby creating multimodal datasets. Our goal is to understand if it would have been feasible to obtain the speech version of BELEBELE by using state of the art TTS systems, instead of human recordings.

For this study we use FLEURS dataset, that contains ASR data in the same domain as BELEBELE. We chose to perform this study in the ASR task because it is simpler compared to other speech tasks, due to its monotonic alignment process and minimal need for reasoning. This ensures that the overall model performance and the complexity of the task are less likely to influence the results.

For our experiments, we generate a synthetic copy of the FLEURS dataset using the MMS TTS (Pratap et al., 2024) system on the FLEURS transcripts. Then, we benchmark state-of-the-art models (WHISPER, SEAMLESSM4T and MMS ASR) on both the original and synthetic datasets and analyze whether the conclusions remain consistent across both datasets. [6]

It is important to note that a decrease in system performance is expected when using synthetic data. However, if this decrease occurs proportionally across all models, the synthetic data could still be useful to benchmark models. Conversely, if the model performance ranking changes, we can conclude that synthetic data is not reliable when benchmarking models.

To measure the variability in model rankings between the original and the synthetic data, we track the inversions that occur in the order of the models in the two settings. We define an inversion as a swap between two models that appear in adjacent positions on the list. We count how many swaps are needed in the ranking obtained using synthetic data to match the ranking from the original dataset.

	SEAMLESSM4T		WHISPER		MMS		
	Hum	Syn	Hum	Syn	Hum	Syn	Inv
Bengali	14.1	21.1	114.7	105.8	14.6	25.0	
Catalan	8.2	13.2	6.7	16.4	10.3	21.8	✓
Dutch	9.9	20.0	8.5	19.7	12.4	28.3	
English	6.0	11.7	4.5	9.8	12.3	19.2	
Finnish	20.1	20.8	12.5	18.9	13.1	18.4	✓
French	9.5	10.8	6.7	11.3	12.4	16.6	✓
German	8.5	13.9	5.2	12.3	10.5	20.8	
Hindi	11.9	13.4	33.5	28.7	11.1	18.3	✓
Indonesian	12.1	12.8	8.7	14.2	13.2	21.9	✓
Korean	25.7	40.3	15.4	29.9	47.8	61.2	
Polish	13.0	14.7	8.1	13.3	11.6	18.1	✓
Portuguese	9.0	8.0	4.1	6.9	8.7	10.4	✓
Romanian	12.6	11.7	13.5	25.4	12.0	15.4	✓
Russian	10.2	18.6	5.6	17.4	18.8	34.3	
Spanish	6.3	9.1	3.4	10.0	6.4	10.8	✓
Swahili	19.5	19.0	64.2	58.4	14.2	19.0	✓
Swedish	15.4	20.1	11.3	19.1	21.0	27.8	
Telugu	27.4	28.0	132.2	133.9	24.2	27.8	
Thai	127.8	135.5	104.0	121.3	99.8	99.9	
Turkish	18.6	23.0	8.4	16.5	19.2	30.3	
Ukrainian	15.0	23.5	9.8	21.8	18.1	34.7	
Vietnamese	16.0	20.1	10.2	14.2	25.8	25.3	

Table 4: WER(↓) results on the ASR task. Last column marks if the language has at least 1 inversion in ASR performance ranking comparing human vs TTS inputs.

[6]Note that we perform the study on the FLEURS languages that are covered by all MMS, WHISPER and SEAMLESSM4T.

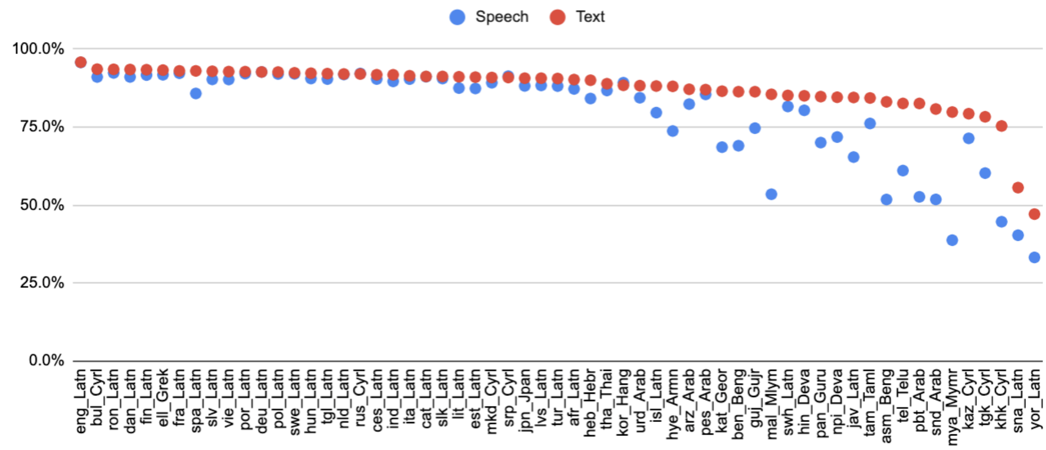

Figure 3: Speech and Text BELEBELE accuracy results in 59 languages. We compare text performance with LLAMA-3-CHAT (zero-shot) and speech performance with WHISPER +LLAMA-3-CHAT (asr+zero-shot).

In Table 4 we see that in the ASR setting, conclusions regarding model performance can vary depending on whether human or synthetic data is used. Although these conclusions are specific to the evaluated tasks and datasets, we demonstrate that even with the outstanding performance of current TTS methods, this does not guarantee the reliability of the data they generate when it comes to evaluation purposes. This is true not only for low-resource languages, but also for high-resource languages such as French or Spanish. These findings show that speech benchmarks might not be reliable if synthetically generated even in widely researched areas, further supporting the creation of evaluation datasets by humans.