UTTER

Unified Transcription and Translation for Extended Reality (UTTER)

Horizon Europe Research and Innovation Action Number: 101070631

D13 – First report on data and resources

Nature	Report	**Work Package**	WP2
Due Date	28/03/2024	**Submission Date**	22/03/2024
Main authors	Marcely Zanon Boito (NAV), Laurent Besacier (NAV)		
Co-authors	Catarina Farinha (UNB), José Souza (UNB), Alexandra Birch (UEDIN), Barry Haddow (UEDIN), Nikita Moghe (UEDIN), Laurie Burchell (UEDIN)		
Reviewers	Wilker Aziz (UVA)		
Keywords	dataset creation, survey, speech data, text data, multilinguality		
Version Control			
v0.1	**Status**	Draft	13/03/2024
v1.0	**Status**	Final	22/03/2024

This project has received funding from the European Union's Horizon Europe Research and Innovation programme under Grant Agreement No 101070631 and from the UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (Grant No 10039436).

Contents

List of Figures

1 Abstract

This report presents the different *data and resources* initiatives from the UTTER project in the context of WP2. This includes a data survey (MS2) conducted by the consortium, as well as many speech and textual resources, detailed below.

Related to the speech modality, the following contributions will be presented:

1. `mHuBERT-147` data: 90K hours of speech multilingual data in 147 languages for self-supervised training;
2. Speech-MASSIVE: a versatile multilingual speech dataset for spoken language understanding in 12 languages (Spanish, Portuguese, Polish, German, Dutch, French, Hungarian, Russian, Turkish, Vietnamese, Arabic, and Korean).

Related to the textual modality, the following contributions will also be presented:

1. ELITR-bench: data augmentation of a popular meeting dataset for the meeting assistant use case (English);
2. MAIA: dialog data collected for the customer care use case in three language pairs (English-German, English-French and English-Portuguese) and emotion annotation;
3. PMIndiaSum: a multilingual and cross-lingual headline summarization dataset covering 14 Indian languages;
4. WMT 2023: test data for the machine translation task for 8 language pairs;
5. ACES and Span-ACES: translation accuracy challenge set in 146 languages with corresponding benchmark covering 50 metrics form WMT 2022 and 2023;
6. OpenLID: an open dataset and model for language identification in 201 languages;
7. MULTI3NLU++: multilingual, multi-intent and multi-domain dataset for natual language understanding in task-oriented dialogue in Spanish, Marathi, Turkish and Amharic.

2 Contributors

Task	Who is reporting	Paper
T2.1	Boito and Besacier §4.2	Marcely Zanon Boito, Vivek Iyer, Nikolaos Lagos, Laurent Besacier, Ioan Calapodescu. *"mHuBERT-147: A Compact Multilingual HuBERT Model"*. Under review.
T2.1	Besacier §4.3	Beomseok Lee, Ioan Calapodescu, Marco Gaido, Matteo Negri, Laurent Besacier. *"Speech-MASSIVE: A multilingual Speech Dataset for SLU and Beyond"*. Under review.
T2.1	Haddow §4.4	Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, Mariya Shmatova. *"Findings of the 2023 Conference on Machine Translation (WMT23): LLMs Are Here but Not Quite There Yet"*. WMT 2023.
T2.1	Moghe §4.5	Chantal Amrhein, Nikita Moghe, Liane Guillou. *"ACES: Translation Accuracy Challenge Sets for Evaluating Machine Translation Metrics"*. WMT 2022.
T2.1	Moghe §4.5	Nikita Moghe, Arnisa Fazla, Chantal Amrhein, Tom Kocmi, Mark Steedman, Alexandra Birch, Rico Sennrich, Liane Guillou. *"Machine Translation Meta Evaluation through Translation Accuracy Challenge Sets"*. arXiv, 2024.
T2.1	Burchell §4.6	Laurie Burchell, Alexandra Birch, Nikolay Bogoychev, Kenneth Heafield. *"An Open Dataset and Model for Language Identification"*. ACL 2023.

Table 1: List of publications to be discussed

Task	Who is reporting	Paper
T2.2	Farinha §5.1	Ana C Farinha, M. Amin Farajian, Marianna Buchicchio, Patrick Fernandes, José G. C. de Souza, Helena Moniz, André F. T. Martins. *"Findings of the WMT 2022 Shared Task on Chat Translation"*. WMT 2022.
T2.2	Farinha §5.2	John Mendonça, Patrícia Pereira, Miguel Menezes, Vera Cabarrão, Ana C. Farinha, Helena Moniz, João Paulo Carvalho, Alon Lavie, Isabel Trancoso. *"Dialogue Quality and Emotion Annotations for Customer Support Conversations"*. GEM workshop, EMNLP 2024.
T2.2	Moghe §5.3	Nikita Moghe, Evgeniia Razumovskaia, Liane Guillou, Ivan Vulić, Anna Korhonen, Alexandra Birch. *"Multi3NLU++: A Multilingual, Multi-Intent, Multi-Domain Dataset for Natural Language Understanding in Task-Oriented Dialogue"*. ACL 2023.
T2.3	Haddow §6.2	Ashok Urlana, Pinzhen Chen, Zheng Zhao, Shay Cohen, Manish Shrivastava, Barry Haddow. *"PMIndiaSum: Multilingual and Cross-lingual Headline Summarization for Languages in India"*. EMNLP 2023.

Table 2: List of publications to be discussed

3 Introduction

3.1 Objectives

Proposal

"Gather, annotate and collect language data to achieve UTTER 's objectives."

Work Completed

We conducted a data survey, which allowed us to understand the data needs of the UTTER partners, and direct our data collection efforts accordingly. For speech, we gathered a massive amount of speech in 147 languages for self-supervised training. We also released a multilingual speech datasets for spoken language understanding in 12 languages (Speech-MASSIVE).

For text we released two datasets related to WP7, one for each use-case. We also released datasets for summarization (PMIndiaSum), natural language understanding (MULTI3NLU++), machine translation test data (WMT 2023), translation accuracy challenge data (ACES and SpanACES), and language identification (OpenLID).

4 Task T2.1: Identifying, collecting and evaluating monolingual and bilingual written and spoken language resources

Proposal

"We will provide an interface to access the very large multilingual text datasets created during the BigScience research workshop. Those datasets cover a variety of scenario including transcription of spoken dialogues. We will also provide a unified access to multiple downstream translation datasets. On the speech side, we will analyse existing data sets for speech recognition and their relevance for the target task. Open-sourced large-scale audio data is mostly based on read out audiobooks (such as LibriSpeech). Real-world audio data, such as conversational audio data (AMI, Fisher, Switchboard) is much smaller in size and often not freely available. Various combinations and strategies for augmenting the datasets will be investigated, including using podcasts, informed by the data governance and ethics work accomplished during the BigScience project."

Work Completed

We survey existing textual and spoken resources, making them available internally for the consortium in a convenient format. With the exit of HuggingFace from the consortium, we shift our focus from BigScience, as we do not have access to their data.

For speech, we describe the speech dataset for the training of our speech foundation model `mHuBERT-147`. We also enrich a popular natural language understanding corpus called MASSIVE with speech recordings in 12 languages. For text, we present two contributions related to machine translation: the WMT 2023 test sets, and the translation accuracy challenge data. We also present a language identification dataset covering 201 languages.

4.1 Survey of Available data (MS2)

For M3, we conducted a survey of available relevant speech and textual data necessary to accomplish UTTER objectives. We met with all WP leaders, and listed the tasks they were planning to explore, and the type of data they found necessary. The main objective of this milestone was to identify data needs and data-related bottlenecks that the consortium could potentially face during the execution of the project.

Relevant data for us meant data that matched our use cases, our target languages and/or tasks. For this project the relevant languages are: Dutch, English, French, German, Portuguese (Euopean) and Korean. Our relevant tasks are (as defined by the WP leaders): Emotion Recognition (from text), Emotion Recognition (from speech), Machine Translation, Minuting, SSL for speech/multimodal pre-training, Speech Translation, Speech Recognition, Summarization.

Format: For this survey, we opted for an interactive format. We host google Drive tables that can be filtered and sorted for easy search. We split information in three tables, that we now detail.

- **General Table:** contains basic information for all datasets: dataset name, languages, modality, task, domain, size (GB), license, download URL, reference link, UTTER use case and short description.
- **Speech Datasets:** details content of speech datasets, including (if applicable): languages (speech), languages (aligned text), duration (hours), number of utterances, speech style, number of speakers, gender information, accent information, emotion labeling, number of utterances with text, number of tokens (text), number of types (text), HuggingFace datasets page.
- **Text Datasets:** details content for speech datasets, including general statistics (number of sentences, types and tokens) for both source and target languages; emotion labeling and HuggingFace datasets page.

For each dataset included in our survey, the procedure was to first fill information in the general table. Then, depending on the modality of the dataset, speech and/or text Dataset tables were filled.

Surveyed Data: We gathered 207 datasets, corresponding to 91 datasets and approximately 10TB of data. Figure 1 presents language representation and percentage of datasets corresponding to one of our two use-cases.

For speech we collected 84 language direction pairs, corresponding to 212,758 hours of speech. This corresponded to 68,154 hours of speech with gender information, 14,783 hours of accented speech (all languages but Korean), 460 hours of emotional speech labeled (French, English and Korean), and 16 speech styles identified (top 3: “Podcasts”, “Spontaneous” and “Read”). Figure 2 presents an overview.

For text we collected 141 language direction pairs, corresponding to over 3.8 billion sentences. From these, 808,487 sentences with text summaries (English, German, French, Korean), 522,108 sentences with emotion labels (English, German, French, Korean), 152,155 sentences from dialogue datasets (English, German, Portuguese, French).

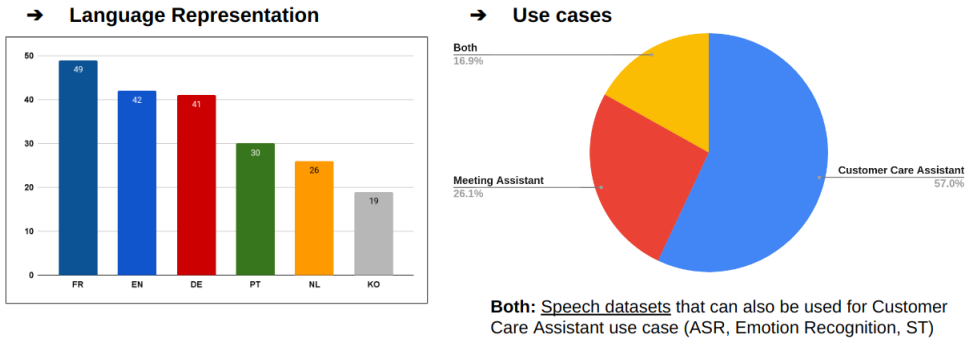

Figure 1: MS2 survey overview.

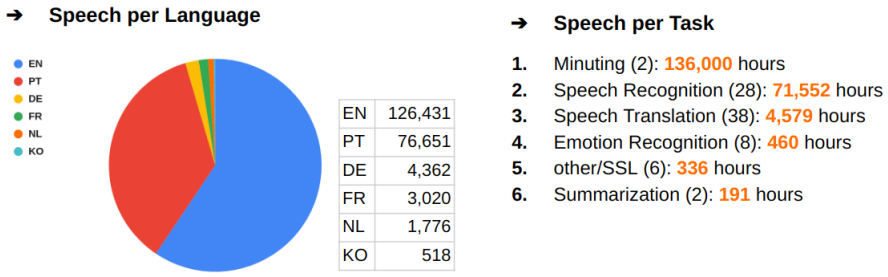

Figure 2: MS2 survey overview for speech datasets.

Findings: We verify that there is enough data for Speech Recognition and Machine Translation across all UTTER languages. For Speech Translation, there is enough data via pipeline approach (ASR+MT). Emotion Recognition (from speech) is achievable in English, French, Korean for the meeting use case. Emotion Recognition (from text) is only relevant in English, for which there is enough data. Minuting is possible in English and Portuguese. The data, made of publicly available datasets, is available internally to the consortium in a convenient format.[1]

4.2 Data collected for the `mHuBERT-147` model

We gathered 90,430 hours of speech from datasets with permissive licences in 147 languages. For this multilingual collection, our goal was to prioritize linguistic diversity over data quantity alone. Table 3 lists datasets and corresponding licences. Appendix Tables 11-16 list all languages included with amount of data per dataset. Figure 3 presents an overview of this information. In total, our training set spans 19 language families (sorted in decreasing order of data quantity): Indo-European, Niger-Congo, Uralic, Afro-Asiatic, Constructed (Esperanto), Turkic, Dravidian, Sino-Tibetan, Austronesian, Koreanic, Kra-Dai, Japonic, Language isolate (Basque), Kartvelian, Austroasiatic, Mongolic, Northwest Caucasian, Creole and Tupian.

Dataset Pre-processing and Filtering

For training `mHuBERT-147`, we follow the default HuBERT pre-processing guidelines (Hsu et al., 2021). For all datasets, the speech data is converted to 16-bit 16kHz WAV files. Volume reduction

[1] https://docs.google.com/spreadsheets/d/1swmj2SrCU3U6SP5uaVouSUCZ0YsdUJsSomcYX3cfDsY/edit?usp=sharing

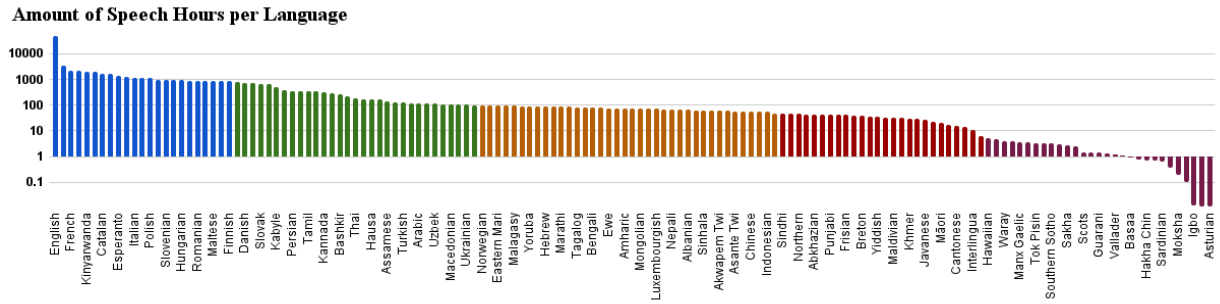

Figure 3: Speech data amount per language in a logarithmic scale. Colors correspond to different levels of speech resourcefulness: $\geq$ 800 h (blue), $\geq$ 100 h (green), $\geq$ 50 h (orange), $\geq$ 10 h (red), $\leq$ 10 h (purple). Some languages are excluded from the plot. Best seen in color.

of 5% if applied for datasets in which we observe excessive clipping during conversion. Data is filtered to the interval of $[2, 30]$s. Speech utterances longer than 30s or shorter than 2s are discarded unless stated otherwise. For TTS/ASR/ST datasets, only the training set is used, no development or test data is included in any case in order to avoid potential data contamination. We now detail dataset-specific pre-processing.

- **B-TTS:** all short utterances ($< 2s$) are concatenated to minimum size ($2s$). This process is performed for the following languages: Akwapem Twi, Asante Twi, Hausa and Yoruba.
- **IISc-MILE:** all short utterances for Kannada ($< 2s$) are concatenated to minimum size ($2s$).
- **JVS:** all short utterances for Japanese ($< 2s$) are concatenated to minimum size ($2s$).
- **VL:** manual inspection revealed that some language splits contained a considerable amount of music, noise and silence-only files. These occurrences were more frequent in less-resourced languages. In order to increase the quality of the data we feed into our self-supervised models, and to correctly estimate the amount of speech data per language they learn from, we filtered this dataset using the `inaSpeechSegmenter` tool (Doukhan et al., 2018). Using this tool's default settings, and knowing that the average utterance length for this dataset is $9s$, we classify a file as *music* if a music event is detected for longer than $2s$. Similarly, a file is classified as *noise* if there is a noise event for longer than $2s$, or if a *non energy* event (i.e. silence) is detected for longer than 5s. These settings were optimized using Cebuano, a language for which we observed a significant amount of noisy utterances. In total, we removed over 332K potentially noisy utterances (249K music, 83K noise/silence).
- **VP:** for languages that are already well represented in the multilingual collection (i.e. more than 1,000 hours of speech data in Tables 11-16), we use only the 10k subset of VP, corresponding to the years of 2019 and 2020. The languages in this setting are German, English, Spanish, French and Dutch. For the remainder, we extract talks from 2017 to 2020 (100k VP split). The languages in this setting are Bulgarian, Czech, Danish, Greek, Estonian, Finnish, Croatian, Hungarian, Italian, Lithuanian, Latvian, Maltese, Polish, Portuguese, Romanian, Slovak, Slovenian and Swedish.

Dataset	Full Name	License	Download URL
Aishell	Aishell (Bu et al., 2017)	Apache License 2.0	https://www.openslr.org/33/
Aishell-3	AISHELL-3 (Shi et al., 2015)	Apache License 2.0	https://www.openslr.org/93/
B-TTS	BibleTTS (Meyer et al., 2022)	CC BY-SA 4.0	http://www.openslr.org/129/
Clovacall	ClovaCall (Ha et al., 2020)	MIT	https://github.com/clovaai/ClovaCall
CV	Common Voice version 11.0 (Ardila et al., 2020a)	CC BY-SA 3.0	https://commonvoice.mozilla.org/en/datasets
G-TTS	High quality TTS data for Javanese (Sodimana et al., 2018)	CC BY-SA 4.0	http://www.openslr.org/41/
	High quality TTS data for Khmer (Sodimana et al., 2018)	CC BY-SA 4.0	http://www.openslr.org/42/
	High quality TTS data for Nepali (Sodimana et al., 2018)	CC BY-SA 4.0	http://www.openslr.org/43/
	High quality TTS data for Sundanese(Sodimana et al., 2018)	CC BY-SA 4.0	http://www.openslr.org/44/
	High quality TTS data for four South African languages (van Niekerk et al., 2017)	CC BY-SA 4.0	http://www.openslr.org/32/
	High quality TTS data for Bengali languages (Sodimana et al., 2018)	CC BY-SA 4.0	https://www.openslr.org/37/
IISc-MILE	IISc-MILE Tamil ASR Corpus (A et al., 2022a), (A et al., 2022b)	CC BY 2.0	https://www.openslr.org/127/
	IISc-MILE Kannada ASR Corpus (A et al., 2022a), (A et al., 2022b)	CC BY 2.0	http://www.openslr.org/126/
JVS	Japanese versatile speech (Takamichi et al., 2019)	CC BY-SA 4.0	https://sites.google.com/site/shinnosuketakamichi/research-topics/jvs_corpus
Kokoro	Kokoro	CC0	https://github.com/kaiidams/Kokoro-Speech-Dataset
kosp2e	Korean Speech to English Translation Corpus (Cho et al., 2021)	CC0	https://github.com/warnikchow/kosp2e
MS	MediaSpeech(Kolobov et al., 2021)	CC BY 4.0	https://www.openslr.org/108/
MLS	Multilingual LibriSpeech (Pratap et al., 2020a)	CC BY 4.0	http://www.openslr.org/94/
Samrómur	Samrómur Unverified 22.07 (Hedström et al., 2022)	CC BY 4.0	https://www.openslr.org/128/
THCHS-30	THCHS-30 (Dong Wang, 2015)	Apache License 2.0	https://www.openslr.org/18/
THUYG-20	THUYG-20 (Roze et al., 2015),(Rozi et al., 2015)	Apache License 2.0	https://www.openslr.org/22/
VL	VoxLingua107(Valk and Alumäe, 2021)	CC BY 4.0	https://bark.phon.ioc.ee/voxlingua107/
VP	VoxPopuli (Wang et al., 2021)	CC0	https://github.com/facebookresearch/voxpopuli/

Table 3: List of datasets used for training `mHuBERT-147` models, their corresponding abbreviation used throughout the paper (left), licenses and links used for download.

Data Overlap with Existing Pre-trained Multilingual SSL Models

	XLSR-53	XLS-R	MMS	mHuBERT-147
BABEL (Gales et al., 2014)	✓	✓	✓	
CV v2	✓	✓	✓	✓
CV v6		✓	✓	✓
CV v9			✓	✓
CV v11				✓
MLS	✓	✓	✓	✓
VP		✓	✓	**down-sampled**
VL		✓	✓	**down-sampled**
MMS-lab-U (Pratap et al., 2023)			✓	
Aishell				✓
AIshell-3				✓
B-TTS				✓
Clovacall				✓
G-TTS				✓
IIScMILE				✓
JVS				✓
Kokoro				✓
kosp2e				✓
MS				✓
Samrómur				✓
THCHS-30				✓
THUYG-20				✓

Table 4: Datasets included in the training of the multilingual models we consider the most comparable to the `mHuBERT-147` models. For `mHuBERT-147` datasets, detailed dataset information can be found in Table 3. We highlight that although here we consider previous versions of CV as being entirely comprised on later versions, this is a simplification, and some files might have been removed due to speakers requests or number of downvotes.

Table 4 presents the overlap between the datasets included in `mHuBERT-147` models and the multilingual wav2vec 2.0 models (Baevski et al., 2020): XLSR-53 (Conneau et al., 2021), XLS-R (Babu et al., 2021) and MMS (Pratap et al., 2023). We do not consider multilingual HuBERT models,

because the ones we are aware of (mHuBERT (Lee et al., 2021), Duquenne et al. (2022)), are very limited in the number of languages covered, and trained on VP only. Looking at the Table, we see that the model most similar to ours in terms of dataset overlap is XLS-R, with only BABEL not included in our training. This dataset is not freely available, and it comprises speech data in 17 low-resource African and Asian languages: Assamese, Bengali, Cantonese, Cebuano, Georgian, Haitian, Kazakh, Lao, Northern Kurdish, Pashto, Swahili, Tagalog, Tamil, Tok Pisin, Turkish, Vietnamese and Zulu. Of these, the only language we did not find an openly available dataset for was Zulu. Thus, `mHuBERT-147` covers 127 of the 128 languages of XLS-R, while supporting 20 additional languages.

This dataset is part of an Interspeech 2024 submission. Due to their anonymity rules, we do not share the data repository for now. It will be made publicly available after the 06/06/2024.

4.3 Speech-MASSIVE: a versatile multilingual speech dataset for SLU and beyond

We expand the MASSIVE dataset (FitzGerald et al., 2022) from its initial text-based format to include speech data. THe original MASSIVE features 1 million realistic, labeled virtual assistant utterances spanning 51 languages, 18 domains, 60 intents, and 55 slots. Our motivation stems from the scarcity of massively multilingual Spoken Language Understanding (SLU) datasets in the research community and the increasing demand for versatile speech datasets to evaluate foundation models (LLMs, speech encoders) for multiple tasks across diverse languages.

To meet this demand, we augment the MASSIVE dataset with spoken versions in 12 languages – Arabic, German, Spanish, French, Hungarian, Korean, Dutch, Polish, Portuguese, Russian, Turkish, and Vietnamese – leveraging the Prolific[2] crowdsourcing platform with rigourous quality controls. Our contributions include creating and sharing a multimodal, multi-task, and multilingual dataset characterized by high parallelism between tasks. We also believe that this dataset can be useful for various other tasks such as speech transcription, language identification, and speech translation.

Related Work. While several massively multilingual datasets have emerged recently, such as CommonVoice (Ardila et al., 2020b), FLEURS (Conneau et al., 2022), MaSS (Zanon Boito et al., 2020), and MLS (Pratap et al., 2020b), they primarily focus on Automatic Speech Recognition (ASR). Additionally, datasets for speech translation (ST) have been introduced, exhibiting a certain level of parallelism, which makes them applicable for both ASR and ST tasks. Examples include MuST-C (Di Gangi et al., 2019), mTEDx (Salesky et al., 2021), Europarl-ST (Iranzo-Sánchez et al., 2020), and CoVoST (Wang et al., 2020). However, it is only recently that more versatile multilingual speech datasets capable of benchmarking speech foundation models for multiple tasks have emerged, such as ML-SUPERB (Shi et al., 2023). Nevertheless, even ML-SUPERB remains confined to a few tasks, predominantly ASR, phone recognition and Language Identification (LID).

Our contribution, Speech-MASSIVE, offers unique versatility, allowing assessment across various speech technology tasks such as ASR, ST, Spoken Language Understanding (SLU), LID, and others. Additionally, it provides a reasonable degree of multilinguality, encompassing 12 languages from diverse language families.

[2] https://www.prolific.com/

4.3.1 The Speech-MASSIVE dataset

MASSIVE multilingual dataset. We extend MASSIVE (FitzGerald et al., 2022), originally a text-only dataset, to incorporate a multimodal aspect through speech. The original MASSIVE dataset was created by tasking professional translators with localizing the English-only SLURP dataset (Bastianelli et al., 2020) into 50 languages. Overall, it contains 1 million labeled utterances across 18 domains, encompassing 60 intentions and 55 slots.

Choice of 12 languages for recording. We selected 12 languages based on various criteria, aiming to record a minimum of 5,000 samples per language while maximizing the diversity of speakers. Initially, we considered the number of users on the crowdsourcing platform, Prolific, sorting the 51 languages covered in (FitzGerald et al., 2022). Languages with no registered users were excluded, followed by those with less than 200 users. Additionally, the language it-IT was removed due to the availability of the full dataset elsewhere (Koudounas et al., 2023). Finally, from the remaining 18 languages, we prioritized a balance between budget considerations and linguistic diversity, resulting in the selection of Spanish, European Portuguese, Polish, German, Dutch, French, Hungarian, Russian, Turkish, Vietnamese, Arabic, and Korean.

Summary of the Prolific data collection

We created the speech counterpart of textual MASSIVE data by recruiting native speakers[3] through the Prolific crowdsourcing platform.

Speech data collection and validation process. A first group of workers was instructed to record the spoken version of MASSIVE sentences with guidelines emphasizing the importance of accurate and natural reading, as well as proper recording conditions and strict adherence to the corresponding text. To ensure high final data quality, a second group of native speakers validated the recorded utterances. During validation, participants were directed to read the original text, listen to the recording, and label it as *valid* or *invalid*. Those marked as invalid underwent a second iteration of this two-step (recording and validation) process. After the second iteration, the process concluded, irrespective of the outcome of the second validation phase, to avoid potentially endless cycles. This decision was also informed by the observation that, upon inspecting the invalid recordings, we found some were marked as such not due to a lack of adherence of the speech to the text but because of grammatical errors in the original MASSIVE dataset text. Correcting these errors was beyond the scope of our work.

To further enhance the reliability of the collected dataset, we implemented two additional precautions. During the recording phase, we instructed participants to review their own recordings before proceeding to the next sample, allowing them to re-record if the audio was not properly acquired. Additionally, in the validation step, four speech utterances were chosen from Common Voice and inserted among the samples for validation. Out of these four quality control samples, two intentionally featured audio-transcript mismatches to be marked as invalid. The other two cases had perfect audio-transcript alignment to be marked as valid. Care was taken to select quality control samples with clear and intelligible audio. Validation results from a Prolific user were retained only if they accurately assessed all four quality control samples. Any mistakes led to the disregarding

[3] Compensated £9 per hour.

of their validations, requiring the entire set of samples from that user to be re-validated by other participants.

Overall statistics We collected speech recordings for the development and test splits of MASSIVE. Obtaining the entire training dataset, which includes 11,514 utterances per language across all 12 languages, exceeded our budget constraints. In a concession, our emphasis was placed on acquiring comprehensive training data for French and German, while we obtained limited few-shot training data consisting of 115 utterances from the training set for the remaining 10 languages (*train-115* split).

lang	split	# sample	# valid	# hrs	total spk (M/F/U)	WER	CER
ar-SA	dev	2033	2027	2.12	36 (22/14/0)	31.78	13.59
	test	2974	2962	3.23	37 (15/17/5)	33.96	14.84
de-DE	train-full	11514	11201	12.61	117 (50/63/4)	-	-
	dev	2033	2032	2.33	68 (35/32/1)	10.13	2.04
	test	2974	2969	3.41	82 (36/36/10)	10.59	2.23
es-ES	dev	2033	2024	2.53	109 (51/53/5)	6.11	1.47
	test	2974	2948	3.61	85 (37/33/15)	7.27	2.05
fr-FR	train-full	11514	11481	12.42	103 (50/52/1)	-	-
	dev	2033	2031	2.20	55 (26/26/3)	9.34	2.92
	test	2974	2972	2.65	75 (31/35/9)	10.18	3.18
hu-HU	dev	2033	2019	2.27	69 (33/33/3)	25.44	9.77
	test	2974	2932	3.30	55 (25/24/6)	20.35	4.65
ko-KR	dev	2033	2032	2.12	21 (8/13/0)	24.63	6.55
	test	2974	2970	2.66	31 (10/18/3)	25.56	7.29
nl-NL	dev	2033	2032	2.14	37 (17/19/1)	9.76	2.36
	test	2974	2959	3.30	100 (48/49/3)	9.32	2.3
pl-PL	dev	2033	2024	2.24	105 (50/52/3)	9.33	3.35
	test	2974	2933	3.21	151 (73/71/7)	11.69	5.52
pt-PT	dev	2033	2031	2.20	107 (51/53/3)	9.97	3.5
	test	2974	2967	3.25	102 (48/50/4)	10.33	3.45
ru-RU	dev	2033	2032	2.25	40 (7/31/2)	7.36	2.45
	test	2974	2969	3.44	51 (25/23/3)	7.52	2.64
tr-TR	dev	2033	2030	2.17	71 (36/34/1)	15.8	3.4
	test	2974	2950	3.00	42 (17/18/7)	17.17	3.73
vi-VN	dev	2033	1978	2.10	28 (13/14/1)	14.64	8.7
	test	2974	2954	3.23	30 (11/14/5)	12.64	7.65

Table 5: Speech-MASSIVE's overall statistics. '# hrs' displays total recording duration for all samples (including invalid), while '# spk (Male/Female/Unknown)' indicates the number of total speakers for all the samples (including invalid). The last 2 columns ('WER', and 'CER') provide whisper-large-v3 evaluation scores.

We provide statistics for the collected dataset,[4] including, for each language, the available data splits, the number of recordings, hours of speech, and speakers (total, male, female and unknown). The "# valid" column indicates the count of human-validated utterances for each data split after the two iterations. As a few speech recordings remained invalidated after our two recording-validation cycles, we retained for each utterance the candidate with the lowest Word Error Rate (WER) as transcribed using Whisper (Radford et al., 2022). This ensures speech availability for all MASSIVE utterances, even if some may not perfectly align with the reference transcript. This information is incorporated into the corpus metadata.

ASR assessment. To assess Speech-MASSIVE in multilingual ASR, we used Whisper, since it is one of the recent state-of-the-art multilingual speech recognition models. We selected Whisper-large-v3,[5] utilizing it without additional fine-tuning for our ASR evaluation. Table 5 shows WER and character error rate (CER) across languages and data splits. We compared ASR error rates to those obtained on the FLEURS dataset (Conneau et al., 2022).[6] FLEURS generally yields lower WERs/CERs compared to Speech-MASSIVE. The same observation was made for Italian in (Koudounas et al., 2023), which followed a recording methodology similar to ours. This suggests that the higher WERs are likely due to the inherent difficulty of MASSIVE utterances compared to those in FLEURS. Furthermore, there are still discrepancies between our Whisper model's hypotheses and the references in the MASSIVE dataset (e.g., numbers reported in letters in MASSIVE references), which we did not address as optimizing ASR WER was not our main goal. Finally, we calculated the correlation coefficient between WERs (CER for Korean) on Speech-MASSIVE and FLEURS, resulting in a value of 0.96. This shows that Whisper consistently performs across both datasets, despite Speech-MASSIVE being more challenging than FLEURS for ASR.

This dataset is part of an Interspeech 2024 submission. Due to their anonymity rules, we do not share the data repository for now. It will be made publicly available after the 06/06/2024.

4.4 WMT 2023 General MT Task Data

The WMT General MT task (formerly the WMT news translation task) is a long-running shared task, where participants are asked to develop machine translation systems on provided data, and to translate shared test sets with the systems. New training and test data is produced each year for the task. The translations are evaluated for quality using human and automatic evaluation, and compared to state-of-the-art commercial systems as well as human translation.

The organisation of the shared task is a large collaborative effort, and the involvement of UTTER was in selecting part of the test sets, as well as preparing some of the training data to be used by participants. For 2023 the test sets covered English↔Chinese, English↔German, English↔Hebrew, English↔Japanese, English↔Russian, English↔Ukrainian, Czech↔Ukrainian and English→Czech. We were responsible for selecting the news portions of the test sets for all languages apart from Hebrew. In the training data, we updated the news crawl for 2022 – it contains ≈ 2B lines in 59 languages.

For full details of the tasks, refer to the overview paper (Kocmi et al., 2023).[7]

[4] For brevity, we omit details on the *train-115* split for 10 languages.
[5] https://hf.co/openai/whisper-large-v3
[6] Accessible for our 12 languages except Arabic at https://github.com/openai/whisper/discussions/1762
[7] The datasets can be downloaded https://data-tmp.statmt.org/news-crawl/ and https://www2.statmt.org/wmt23/

4.5 Translation Accuracy Challenge Sets

Recent machine translation (MT) metrics calibrate their effectiveness by correlating with human judgement. However, these results are often obtained by averaging predictions across large test sets without any insights into the strengths and weaknesses of these metrics across different error types. Challenge sets are used to probe specific dimensions of metric behaviour but there are very few such datasets and they either focus on a limited number of phenomena or a limited number of language pairs. Particularly, exploring metric behaviour in the presence of accuracy errors in machine translation is crucial as undetected errors can have severe consequences in certain domains (*e.g.,* legal, medical).

We introduce ACES, a contrastive challenge set spanning 146 language pairs, aimed at discovering whether metrics can identify 68 translation accuracy errors. These phenomena range from basic alterations at the word/character level to more intricate errors based on discourse and real-world knowledge. We conducted a large-scale study by benchmarking ACES on 50 metrics submitted to the WMT 2022 and 2023 metrics shared tasks. We benchmark metric performance, assess their incremental performance over successive campaigns, and measure their sensitivity to a range of linguistic phenomena. We also investigate claims that Large Language Models (LLMs) are effective as MT evaluators, addressing the limitations of previous studies by providing a more holistic evaluation that covers a range of linguistic phenomena and language pairs and includes both low- and medium-resource languages.

Our results demonstrate that different metric families struggle with different phenomena and that LLM-based methods fail to demonstrate reliable performance. Our analyses indicate that most metrics ignore the source sentence, tend to prefer surface-level overlap and end up incorporating properties of base models which are not always beneficial. To further encourage detailed evaluation beyond singular scores, we expand ACES to include error span annotations, denoted as SPAN-ACES and we use this dataset to evaluate span-based error metrics showing these metrics also need considerable improvement.

Finally, we provide a set of recommendations for building better MT metrics, including focusing on error labels instead of scores, ensembling, designing strategies to explicitly focus on the source sentence, focusing on semantic content rather than relying on the lexical overlap, and choosing the right base model for obtaining representations. See Amrhein et al. (2022) and Moghe et al. (2024) for further details.

ACES and SPAN-ACES are publicly available on Huggingface.[8]

4.6 An Open Dataset and Model for Language Identification

Language identification (LID) is a foundational step in many NLP pipelines. It is used not only to select data in the relevant language but also to exclude 'noise'. For this reason, effective LID systems are key for building useful and representative NLP applications.

Despite their importance, recent work has found that existing LID algorithms perform poorly in practice compared to test performance (Caswell et al., 2020). The problem is particularly acute for low-resource languages: Kreutzer et al. (2022) found a positive Spearman rank correlation between quality of data and size of language for all of the LID-filtered multilingual datasets they

wmttest2023.src.zip .

[8] https://huggingface.co/datasets/nikitam/ACES.

studied. In addition, for a significant fraction of the language corpora they studied, less than half of the sentences were in the correct language. They point out that such low-quality data not only leads to poor performance in downstream tasks, but that it also contributes to 'representation washing', where the community is given a false view of the actual progress of low-resource NLP.

For applications such as corpus filtering, LID systems need to be fast, reliable, and cover as many languages as possible. There are several open LID models offering quick classification and high language coverage, such as CLD3 or the work of Costa-jussà et al. (2022). However, to the best of our knowledge, none of the commonly-used scalable LID systems make their training data public.

To address this gap, we provided OpenLID, a curated and open dataset covering 201 language varieties. The majority of our source datasets were derived from news sites, Wikipedia, or religious text, though some come from other domains (e.g. transcribed conversations, literature, or social media). We audited a sample from each source and each language making up this dataset manually to ensure quality. The final dataset contains 121 million lines of data in 201 language classes. Before sampling, the mean number of lines per language is 602,812. The smallest class contains 532 lines of data (South Azerbaijani) and the largest contains 7.5 million lines of data (English).

We also trained the OpenLID LID model using this dataset and showed that it outperformed existing open LID models, achieving a macroaverage F1 score of 0.93 over 201 language classes. We made this model publicly available and it is now used by the Wikimedia Foundation (among others).

Further details on the dataset and model are available in Burchell et al. (2023). The OpenLID dataset is publicly available through GitHub[9] and on Huggingface.[10] The OpenLID LID model trained on this dataset is publicly available either through the same GitHub repository or on HuggingFace.[11]

5 Task T2.2: Data for dialogue

Proposal

"We will develop new data sets which are useful for dialogue translation for both text and speech modalities. This will involve collecting and curating already existing conversational parallel data (e.g. from repositories such as OPUS), translating monolingual dialogue data, as well as collecting new data as part of the development of the use cases described in WP6. A multilingual chat dataset will be created, anonymised, and curated with real-world customer care bilingual dialogue by Unbabel. For speech translation, medium-size parallel speech-text datasets already exist such as MUST-C, Europarl-ST,and CoVoST2. However, those corpora deal with translation of TED talks, parliament speeches and read speech respectively. As our focus is on dialog, those corpora are rather off-domain and we will have to create datasets for the dialog translation use case. Possible starting point to obtain more in-domain speech data is to ask speakers to record speech from the source of existing (and bigger) conversational parallel texts. We may also leverage monolingual dialog speech for self-supervised pre-training."

[9] https://github.com/laurieburchell/open-lid-dataset
[10] https://huggingface.co/datasets/laurievb/open-lid-dataset
[11] https://huggingface.co/laurievb/OpenLID

Work Completed

In the context of the customer service assistant, we release the MAIA bilingual dialog dataset. We also conducted comprehensive emotion and dialogue quality annotations, that we make available to the community. Finally, we release a multilingual, multi-intent, multi-domain dataset for natural language understanding in task-oriented dialogue.

5.1 Customer Service Assistant: MAIA dataset

For the customer care use case, we compiled the Unbabel's MAIA Dataset (Martins et al., 2020) that was released on the WMT 2022 Shared Task on Chat Translation (Farinha et al., 2022). This dataset comprises original and complete bilingual conversations extracted from four distinct real flows within Unbabel. Skilled translators from the Unbabel Community translated the original exchanges between customers and agents into their respective target languages. Specifically, customer utterances were consistently translated into English, while agent responses were rendered from English into other languages. An excerpt an en↔pt-br conversation between a *customer* and an *agent* can be seen in Table 6.

In total, the dataset encompasses over 40,000 segments extracted from more than 900 conversations across three language pairs, encompassing six language translation directions: English-German, English-French, and Portuguese-English (Brazil). Statistics about the Unbabel's MAIA Dataset, used for the WMT 2022 Chat Translation Shared task, are shown in Table 8.

In compliance with the General Data Protection Regulation (GDPR) and to facilitate public access to the conversations, we anonymized them using a two-step process. Initially, we employed Unbabel's proprietary anonymization tool for automatic anonymization, followed by manual verification of the data. This process yielded 12 distinct anonymization categories, each represented by a specific token, as detailed in Table 7.

5.2 Customer Service Assistant: Dialog Quality and Emotion Annotations

Building on top of the MAIA Dataset, we conducted comprehensive emotion and dialogue quality annotations (Mendonça et al., 2023). Specifically, we annotated the dataset at various levels of granularity. At the sentence level, we perform 8-class emotion (Neutral, Anger, Confusion, Anxiety, Frustration, Empathy, Disappointment, Happiness) and local conversational quality annotations (Correctness, Templated, Engagement). At the turn-level, conversational quality annotations (IQ (Interaction Quality), Understanding, Sensibleness, Politeness), are provided. Finally, at the dialogue level, annotations for task success are included (Dropped Convesations and Task Sucess). A total of 612 dialogues amounting to over 24k sentences from en↔pt-br and en↔en language pairs were annotated.

Statistical information concerning the dialogue quality annotations of the MAIA dataset can be seen in Table 9 and emotion annotations distribution in Figure 4. Annotations are publicly available.

This dataset (Mendonça et al., 2023) was delivered on the context of MS9. It can be accessed at GitHub.[12]

[12] http://github.com/johndmendonca/MAIA-DQE

customer	source_segment: Ola, tudo bem? target_segment: Hello! How are you?
customer	source_segment: Alguns meses atras, precisei restaurar o aplicativo da #PRS_ORG# para PC. target_segment: A few months ago, I needed to restore the #PRS_ORG# PC App.
customer	source_segment: Quando fiz isso, perdi todos os meus livros comprados. target_segment: When I did that, I lost all my purchased books.
customer	source_segment: Gostaria de saber como recupera-los. target_segment: I would like to know how to recover them.
customer	source_segment: Obrigada. target_segment: Thank you.
customer	source_segment: Celular para contato: #PHONENUMBER#. target_segment: Mobile for contact: #PHONENUMBER#
agent	source_segment: Thank you for the information. target_segment: Agradeço pela informação.
agent	source_segment: I will be more than happy to assist you. target_segment: Terei todo o prazer em ajudar você.
agent	source_segment: I see all your books are in the account linked to the #EMAIL# target_segment: Vejo que todos os seus livros estão na conta vinculada ao #EMAIL#

Table 6: Excerpt of a en↔pt-br conversation between a *customer* and an *agent*.

Token	Description
#NAME#	Person's names
#PRS_ORG#	Products, Services, and Organizations
#ADDRESS#	Address
#EMAIL#	E-mail address
#IP#	IP Address
#PASSWORD#	Password
#PHONENUMBER#	Phone number
#CREDITCARD#	Credit card number
#URL#	URL Address
#IBAN#	IBAN Address
#NUMBER#	Any number (all digits)
#ALPHANUMERIC_ID#	Any alphanumeric ID

Table 7: Anonymization tokens and their description.

5.3 MULTI3NLU++: A Multilingual, Multi-Intent, Multi-Domain Dataset for Natural Language Understanding in Task-Oriented Dialogue

Task-oriented dialogue (TOD) systems (Gupta et al., 2006; Young et al., 2013), in which conversational agents assist human users to achieve their specific goals, have been used to automate telephone-based and online customer service tasks in a range of domains, including travel (Raux et al., 2003, 2005), finance and banking (Altinok, 2018), and hotel booking (Li et al., 2019). ToD systems are often implemented as a pipeline of dedicated modules (Raux et al., 2005; Young et al.,

	en↔de			en↔fr			en↔pt-br		
	source-only dev set	parallel dev set	parallel test set	source-only dev set	parallel dev set	parallel test set	source-only dev set	parallel dev set	parallel test set
Number of conversations	355	70	71	84	59	51	57	47	60
Number of total seg.	13,400	2,109	2,488	5,239	2,753	3,065	3,672	2,359	2,384
Number of agent seg.	6,389	1,006	1,113	3,305	1,750	1,937	2,007	1,353	1,381
Number of customer seg.	7,011	1,103	1,375	1,934	1,003	1,128	1,665	1,006	1,003

Table 8: Number of conversations and segments provided in the WMT 2022 Chat Translation Shared Task.

Annotation	Count
Correctness {0,1,2}	205 — 938 — 23,730
Templated {0,1}	18,174 — 6,602
Engagement {0,1}	315 — 23,712
Understanding {0,1}	136 — 9,470
Sensibleness {0,1}	127 — 9,478
Politeness {0,1}	345 — 9,390
IQ [1,5]	89 — 479 — 1,665 — 4,358 — 3,012
Dropped Conv. {0,1}	499 — 112
Task Success [1,5]	35 — 63 — 141 — 27 — 347

Table 9: Statistical information pertaining to the annotations of the MAIA dataset.

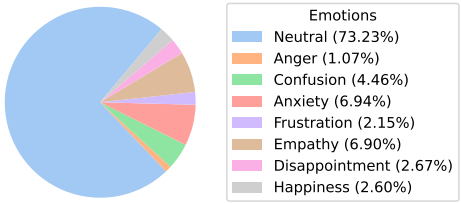

Figure 4: Emotion distribution of the MAIA dataset.

2013). Two prominent tasks in these pipelines 1) intent detection and 2) slot labelling.

Existing datasets for these tasks are **1)** predominantly limited to detecting a single intent, **2)** focused on a single domain, and **3)** include a small set of slot types (Larson and Leach, 2022; Casanueva et al., 2022). Furthermore, the success of task-oriented dialogue is **4)** often evaluated on a small set of higher-resource languages (i.e., typically English) which does not test how generalisable systems are to the diverse range of the world's languages (Razumovskaia et al., 2022).

To address all of the limitations discussed above, we proposed MULTI^{3}NLU^{++}, a **multi**lingual, **multi**-intent, **multi**-domain for training and evaluating TOD systems. MULTI^{3}NLU^{++} extends the recent monolingual English-only dataset NLU++, which is a multi-intent, multi-domain dataset for the Banking and Hotels domains. MULTI^{3}NLU^{++} adds the element of multilinguality and thus enables simultaneous cross-domain and cross-lingual training and experimentation for TOD NLU as its unique property.

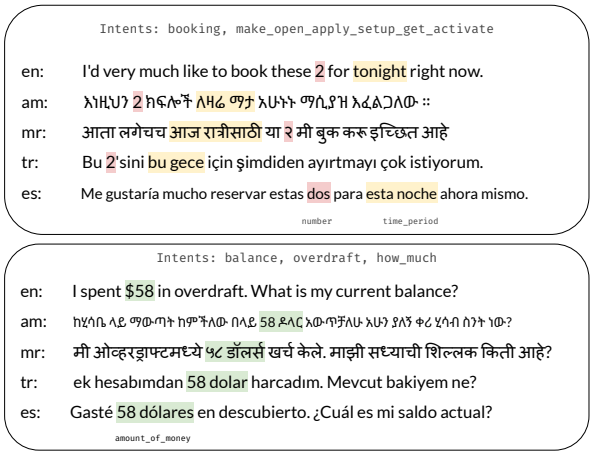

Figure 5: Examples from the MULTI^{3}NLU^{++} dataset for the Hotel and Banking domain demonstrating the complex NLU tasks of multi-label intent detection and slot labelling across multiple languages. Intent labels consist of generic and domain-specific intents. The slot values per example are highlighted in different colours with the slot labels beneath them. en: English, am: Amharic, mr: Marathi, tr: Turkish, es: Spanish.

MULTI^{3}NLU^{++} includes expert manual translations of the 3,080 utterances in NLU++ to four languages of diverse typology and data availability: Spanish, Marathi, Turkish, and Amharic. The selection of languages covers a range of language families and scripts and includes high, medium, and low-resource languages. Capturing language diversity is particularly important if we wish to design multilingual TOD systems that are robust to the variety of expressions used across languages to represent the same value or concept. We list an example from both the domains in Figure 5. Note the variety of expressing the same intent and same slot across different languages.

Using MULTI^{3}NLU^{++} we demonstrated the challenges involved in extending existing state-of-the-art Machine Translation systems and multilingual language models for NLU in TOD systems. We experiment with various intent detection and slot-filling models and highlight that (i) there is a significant drop in performance across all languages as compared to NLU++ with performance drops increasing as we progress from high- to low-resource languages; (ii) zero-shot performance improves when the source language has lower resources; (iii) cross-lingual transfer in intent detection is dependent on matching the script of the source language and amount of data during pretraining setup.

For more details, please refer to Moghe et al. (2023). MULTI^{3}NLU^{++} is publicly available on Huggingface.[13]

6 Task T2.3: Data for minuting and summarisation

Proposal

"In this task, we will gather and extend meeting summarisation datasets with attribution and relevant span annotations. The data targeted here is related to the meeting use case. We will first gather existing meeting speech datasets such as: ICSI, NIST meeting pilot, and AMI. These corpora will

[13] https://huggingface.co/datasets/uoe-nlp/multi3-nlu

provide a strong starting point to evaluate transcription and minuting on English, despite limitations, such as the fact that they were collected in smart meeting environments with microphone arrays whereas we target virtual meetings. We will also use data from a recent shared task on automatic minuting, aimed at multi-party meetings. In addition, we will consider recent Europarl-ST corpus that contains paired audio-text samples constructed using debates in the European Parliament. To our knowledge, none of the previously described corpora is annotated in term of paralinguistic features such as emotion. We thus plan to annotate a subset of them along this dimension at least for evaluation of contextualisation and emotion tracking (see WP4)."

Work Completed

Since the proposal, we witnessed the release of novel LLMs able to perform minuting with an unprecedented performance in long context mode over meeting transcripts. Due to this, the meeting assistant use case changed focus from speech minuting to interactive use-tailored minuting from meeting transcripts. We release an extension of the ELITR dataset for evaluating different chat LLMs on this task (ELITR-Bench). Focusing on text summarization, we release a cross-lingual headline summarization in 14 languages (PMIndiaSum).

6.1 ELITR-Bench Dataset

Context: Meeting assistant prototype

We built the first prototype of our UTTER meeting assistant. It is more precisely described in a youtube video presentation.[14] Our meeting assistant does not only provide an ASR transcript and a summary of the meetings. It is a *"smart assistant which attended the meeting on your behalf"*. Users can chat with it and seek information about a former meeting they attended long ago, or about a meeting they did not attend to. This allows the user to efficiently seek information without having to read the full transcriptions.

Our meeting assistant UI is based in *streamlit*[15] and the assistant is powered by *OpenAI* LLMs for the moment. In short our assistant has the following setup:

- A single instructed LLM with a long context of 16k tokens (*gpt3.5-16k*) is used to allow processing transcripts of 1h long meetings;
- We sample LLM's responses at unit temperature (*temp* parameter is set to 1.0) as a default setting;
- A particular *system prompting*, presented in Figure 6, is used for our assistant (prompts slightly differ depending on the style of the speech transcripts used).

Our augmented ELITR dataset (ELITR-Bench)

Our prototype was evaluated using several datasets including the already existing ELITR corpus which contains anonymised transcripts for meetings that took place within an EU research pro-

[14] https://tinyurl.com/UTTER-Meeting-Assistant
[15] https://streamlit.io

-Prompt MrMeeting (UTTER meetings)-

```
The following is the transcript of a meeting with
multiple participants, where each line has a
timestamp (e.g. 11:58:37 AM means 11h58mn37s am),
the speaker's name and their utterance.

<meeting-transcript>

As a professional conversational assistant, you can
respond to any questions about the meeting, and you
can make inferences from the transcripts.

<user-question>
```

-Prompt MrMeeting (ELITR meetings)-

```
The following is the transcript of a meeting with
multiple participants, where each line has an
anonymized speaker's name (for instance PERSON4 )
and their utterance.

<meeting-transcript>

As a professional conversational assistant, you can
respond to any questions about the meeting, and you
can make inferences from the transcripts.

<user-question>
```

Figure 6: Prompts used for our assistant (prompts slightly differ depending on the style of the speech transcripts used)

ject.[16] For each meeting of this corpus, we prepared questions which can be answered from the transcript, as well as their ground truth answer. Our questions are of different types: **Who** questions, **What** questions, **When** questions, and **How many** questions.

We also annotated if answer to question is in the **Beginning** (1st third), **Middle** (2nd third) or **End** (3rd third) or on **Several** blocks of the meeting transcript - in order to see if we can confirm results of the *Lost in the middle* paper from Liu et al. (2023).

Our starting point is the ELITR Minuting Corpus that consists of transcripts of meetings in Czech and English, their manually created summaries ("minutes") and manual alignments between the two. We only used the English meetings which are in the computer science domain. Each transcript has one or multiple corresponding minutes files. We worked with the official *dev* (10 meetings) and *test2* (8 meetings) sets of ELITR-English. As ELITR is open-source and anonymized we augment it with interaction logs between a user and our meeting assistant prototype, and share it on a GitHub repository.

We believe that such a dataset is interesting for open-ended evaluation of LLMs especially for tasks that require long-form answers (such as general purpose or specialised assistants). Table 10 summarizes the statistics of the different Q&A made available in our augmented version of ELITR.

[16] https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-4692

	#meetings	#questions		#positions	
dev	10	what	59	B	45
		who	51	M	29
		when	21	E	32
		how many	10	S	35
		all	**141**		141
test	8	what	57	B	43
		who	45	M	34
		when	20	E	22
		how many	8	S	31
		all	**130**		130

Table 10: Statistics of the augmented ELITR dataset - all questions and answers are annotated per question type (What, Who, When, How many) and per answer position within the meeting transcript (Begin, Middle, End or Several)

This dataset was delivered on the context of MS9. It can be accessed at GitHub.[17]

6.2 PMIndiaSum: Multilingual and Cross-lingual Headline Summarization for Languages in India

In this work, we constructed a multilingual and massively parallel summarization corpus focused on the languages spoken in India. We exploited our previous work (Haddow and Kirefu, 2020) where we crawled the Prime Minister of India's website to create a multi-parallel corpus covering 14 languages. PMIndiaSum converts this resource together with newly crawled data into a massively cross-lingual summarization corpus by aligning article bodies with headlines in up to 14 languages. The resulting corpus has 76,680 monolingual document-summary pairs across 14 languages and 620,336 cross-lingual document-summary pairs for 182 summarization language directions.

We also performed summarization experiments where we compared heuristic extractive baselines, with summarize-and-translate, and fine-tuning of two state-of-the-art multilingual pre-trained models. We showed that summarize-and-translate performed slightly better across the whole corpus, but that there is still significant room for improvement. The full details of the extraction process, the corpus, and our summarization experiments are described in our paper (Urlana et al., 2023).

Our corpus (Urlana et al., 2023) is distributed via Hugging Face.[18] Our processing scripts are publicized on GitHub.[19]

[17] https://github.com/utter-project/UTTER-MS9-meetingdata
[18] https://huggingface.co/datasets/PMIndiaData/PMIndiaSum
[19] https://github.com/ashokurlana/PMIndiaSum

7 Conclusion

In this report we detailed the many different data projects related to WP2, illustrating our progress in all three tasks. For T2.1, we surveyed massive amounts of available data, collected two speech datasets (SSL and ASR/ST/SLU), and three text datasets (MT and LID). For T2.2, we presented two rich datasets for dialogue-related tasks, one of them including emotion labels. Finally, for T2.3, we presented evaluation data for the meeting assistant use-case, and a summarization dataset.

In summary, during the first half of the UTTER project, we collected many relevant datasets that allowed us to advance in our research. For both speech and text, we enriched the collection of available resources to the consortium, increasing coverage for the UTTER languages, and covering both use-cases (meeting assistant and customer care assistant). These resources were already explored by some works that will be presented in parallel deliverables, and they will continue to be used throughout the project.

References

Madhavaraj A, Bharathi Pilar, and Ramakrishnan A G. Subword dictionary learning and segmentation techniques for automatic speech recognition in tamil and kannada, 2022a. URL https://arxiv.org/abs/2207.13331.

Madhavaraj A, Bharathi Pilar, and Ramakrishnan A G. Knowledge-driven subword grammar modeling for automatic speech recognition in tamil and kannada, 2022b. URL https://arxiv.org/abs/2207.13333.

Duygu Altinok. An ontology-based dialogue management system for banking and finance dialogue systems. In Mahmoud El-Haj, Paul Rayson, and Andrew Moore, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may 2018. European Language Resources Association (ELRA). ISBN 979-10-95546-23-8.

Chantal Amrhein, Nikita Moghe, and Liane Guillou. ACES: Translation accuracy challenge sets for evaluating machine translation metrics. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 479–513, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.44.

R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215, 2020a.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France, May 2020b. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.520.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*, 2021.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, and Verena Rieser. SLURP: A spoken language understanding resource package. *CoRR*, abs/2011.13205, 2020. URL https://arxiv.org/abs/2011.13205.

Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *Oriental COCOSDA 2017*, page Submitted, 2017.

Laurie Burchell, Alexandra Birch, Nikolay Bogoychev, and Kenneth Heafield. An open dataset and model for language identification. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 865–879, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.75. URL https://aclanthology.org/2023.acl-short.75.

Inigo Casanueva, Ivan Vulić, Georgios Spithourakis, and Paweł Budzianowski. NLU++: A multi-label, slot-rich, generalisable dataset for natural language understanding in task-oriented dialogue. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1998–2013, Seattle, United States, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.154. URL https://aclanthology.org/2022.findings-naacl.154.

Isaac Caswell, Theresa Breiner, Daan van Esch, and Ankur Bapna. Language ID in the wild: Unexpected challenges on the path to a thousand-language web text corpus. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6588–6608, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.579. URL https://aclanthology.org/2020.coling-main.579.

Won Ik Cho, Seok Min Kim, Hyunchang Cho, and Nam Soo Kim. kosp2e: Korean Speech to English Translation Corpus. In *Proc. Interspeech 2021*, pages 3705–3709, 2021. doi: 10.21437/Interspeech.2021-1040.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. Unsupervised Cross-Lingual Representation Learning for Speech Recognition. In *Proc. Interspeech 2021*, pages 2426–2430, 2021. doi: 10.21437/Interspeech.2021-329.

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

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. MuST-C: a Multilingual Speech Translation Corpus. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2012–2017, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1202. URL https://aclanthology.org/N19-1202.

Zhiyong Zhang Dong Wang, Xuewei Zhang. Thchs-30 : A free chinese speech corpus, 2015. URL http://arxiv.org/abs/1512.01882.

David Doukhan, Jean Carrive, Félicien Vallet, Anthony Larcher, and Sylvain Meignier. An open-source speaker gender detection framework for monitoring gender equality. In *Acoustics Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*. IEEE, 2018.

Paul-Ambroise Duquenne, Hongyu Gong, Ning Dong, Jingfei Du, Ann Lee, Vedanuj Goswani, Changhan Wang, Juan Pino, Benoît Sagot, and Holger Schwenk. Speechmatrix: A large-scale mined corpus of multilingual speech-to-speech translations. *arXiv preprint arXiv:2211.04508*, 2022.

Ana C Farinha, M. Amin Farajian, Marianna Buchicchio, Patrick Fernandes, José G. C. de Souza, Helena Moniz, and André F. T. Martins. Findings of the WMT 2022 shared task on chat translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 724–743, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.70.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, et al. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages. *ArXiv preprint*, abs/2204.08582, 2022. URL https://arxiv.org/abs/2204.08582.

Mark JF Gales, Kate M Knill, Anton Ragni, and Shakti P Rath. Speech recognition and keyword spotting for low-resource languages: Babel project research at cued. In *Fourth International workshop on spoken language technologies for under-resourced languages (SLTU-2014)*, pages 16–23. International Speech Communication Association (ISCA), 2014.

N. Gupta, G. Tur, D. Hakkani-Tur, S. Bangalore, G. Riccardi, and M. Gilbert. The AT&T spoken language understanding system. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):213–222, 2006. doi: 10.1109/TSA.2005.854085.

Jung-Woo Ha, Kihyun Nam, Jingu Kang, Sang-Woo Lee, Sohee Yang, Hyunhoon Jung, Hyeji Kim, Eunmi Kim, Soojin Kim, Hyun Ah Kim, Kyoungtae Doh, Chan Kyu Lee, Nako Sung, and Sunghun Kim. ClovaCall: Korean Goal-Oriented Dialog Speech Corpus for Automatic Speech Recognition of Contact Centers. In *Proc. Interspeech 2020*, pages 409–413, 2020. doi: 10.21437/Interspeech.2020-1136.

Barry Haddow and Faheem Kirefu. PMIndia—A collection of parallel corpora of languages of India. *arXiv preprint*, 2020. URL https://arxiv.org/abs/2001.09907v1.

Staffan Hedström, Ragnheiður Òrhallsdòttir, David Erik Mollberg, Smári Freyr Guðmundsson, Òlafur Helgi Jónsson, Sunneva Òrsteinsdóttir, Eydís Huld Magnúsdóttir Judy Y. Fong, and Jon Gudnason. Samrómur Unverified 22.07. Reykjavik University: Language and Voice Lab, 2022.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Adrià Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan. Europarl-st: A multilingual corpus for speech translation of parliamentary debates, 2020.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.1. URL https://aclanthology.org/2023.wmt-1.1.

Rostislav Kolobov, Olga Okhapkina, Andrey Platunov Olga Omelchishina, Roman Bedyakin, Vyacheslav Moshkin, Dmitry Menshikov, and Nikolay Mikhaylovskiy. Mediaspeech: Multilanguage asr benchmark and dataset, 2021.

Alkis Koudounas, Moreno La Quatra, Lorenzo Vaiani, Luca Colomba, Giuseppe Attanasio, Eliana Pastor, Luca Cagliero, and Elena Baralis. Italic: An italian intent classification dataset. *arXiv preprint arXiv:2306.08502*, 2023.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72, 2022. doi: 10.1162/tacl_a_00447. URL https://aclanthology.org/2022.tacl-1.4.

Stefan Larson and Kevin Leach. A survey of intent classification and slot-filling datasets for task-oriented dialog. *ArXiv preprint*, abs/2207.13211, 2022. doi: 10.48550/arXiv.2207.13211. URL https://doi.org/10.48550/arXiv.2207.13211.

Ann Lee, Hongyu Gong, Paul-Ambroise Duquenne, Holger Schwenk, Peng-Jen Chen, Changhan Wang, Sravya Popuri, Yossi Adi, Juan Pino, Jiatao Gu, et al. Textless speech-to-speech translation on real data. *arXiv preprint arXiv:2112.08352*, 2021.

Bai Li, Nanyi Jiang, Joey Sham, Henry Shi, and Hussein Fazal. Real-world conversational ai for hotel bookings. In *2019 Second International Conference on Artificial Intelligence for Industries (AI4I)*, pages 58–62, 2019. doi: 10.1109/AI4I46381.2019.00022.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts, 2023.

André F. T. Martins, Joao Graca, Paulo Dimas, Helena Moniz, and Graham Neubig. Project MAIA: Multilingual AI agent assistant. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 495–496, Lisboa, Portugal, November 2020. European Association for Machine Translation. URL https://aclanthology.org/2020.eamt-1.68.

John Mendonça, Patrícia Pereira, Miguel Menezes, Vera Cabarrão, Ana C. Farinha, Helena Moniz, João Paulo Carvalho, Alon Lavie, and Isabel Trancoso. Dialogue quality and emotion annotations for customer support conversations. 2023.

Josh Meyer, David Adelani, Edresson Casanova, Alp Öktem, Daniel Whitenack, Julian Weber, Salomon Kabongo Kabenamualu, Elizabeth Salesky, Iroro Orife, Colin Leong, Perez Ogayo, Chris Chinenye Emezue, Jonathan Mukiibi, Salomey Osei, Apelete Agbolo, Victor Akinode, Bernard Opoku, Olanrewaju Samuel, Jesujoba Alabi, and Shamsuddeen Hassan Muhammad. Bibletts: a large, high-fidelity, multilingual, and uniquely african speech corpus. In *Interspeech*. ISCA, 2022. URL https://arxiv.org/pdf/2207.03546.pdf.

Nikita Moghe, Evgeniia Razumovskaia, Liane Guillou, Ivan Vulić, Anna Korhonen, and Alexandra Birch. Multi3NLU++: A multilingual, multi-intent, multi-domain dataset for natural language understanding in task-oriented dialogue. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3732–3755, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.230. URL https://aclanthology.org/2023.findings-acl.230.

Nikita Moghe, Arnisa Fazla, Chantal Amrhein, Tom Kocmi, Mark Steedman, Alexandra Birch, Rico Sennrich, and Liane Guillou. Machine translation meta evaluation through translation accuracy challenge sets. *Computing Research Repository*, arXiv:2401.16313, 2024. URL http://arxiv.org/abs/2401.16313.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. Mls: A large-scale multilingual dataset for speech research. *ArXiv*, abs/2012.03411, 2020a.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. Mls: A large-scale multilingual dataset for speech research. In *Interspeech 2020*. ISCA, October 2020b. doi: 10.21437/interspeech.2020-2826. URL http://dx.doi.org/10.21437/Interspeech.2020-2826.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. Scaling speech technology to 1,000+ languages. *arXiv preprint arXiv:2305.13516*, 2023.

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

Antoine Raux, Brian Langner, Alan W. Black, and Maxine Eskénazi. Let's go: improving spoken dialog systems for the elderly and non-natives. *8th European Conference on Speech Communication and Technology (Eurospeech 2003)*, 2003. URL http://www.isca-speech.org/archive/eurospeech_2003/e03_0753.html.

Antoine Raux, Brian Langner, Dan Bohus, Alan W Black, and Maxine Eskenazi. Let's go public! taking a spoken dialog system to the real world. In *9th European Conference on Speech Communication and Technology, Lisbon, Portugal*, pages 432–437, 2005. URL http://www.isca-speech.org/archive/interspeech_2005/i05_0885.html.

Evgeniia Razumovskaia, Goran Glavas, Olga Majewska, Edoardo M Ponti, Anna Korhonen, and Ivan Vulic. Crossing the conversational chasm: A primer on natural language processing for multilingual task-oriented dialogue systems. *Journal of Artificial Intelligence Research*, 74: 1351–1402, 2022. URL https://doi.org/10.1613/jair.1.13083.

Askar Roze, Shi Yin, Dong Wang Zhiyong Zhang, and Askar Hamdulla. Thugy20: A free uyghur speech database. In *NCMMSC'15*, 2015.

Askar Rozi, Dong Wang, and Zhiyong Zhang. An open/free database and benchmark for uyghur speaker recognition. In *O-COCOSDA'15*, 2015.

Elizabeth Salesky, Matthew Wiesner, Jacob Bremerman, Roldano Cattoni, Matteo Negri, Marco Turchi, Douglas W. Oard, and Matt Post. The multilingual tedx corpus for speech recognition and translation. *CoRR*, abs/2102.01757, 2021. URL https://arxiv.org/abs/2102.01757.

Jiatong Shi, Dan Berrebbi, William Chen, Ho-Lam Chung, En-Pei Hu, Wei Ping Huang, Xuankai Chang, Shang-Wen Li, Abdelrahman Mohamed, Hung yi Lee, and Shinji Watanabe. Ml-superb: Multilingual speech universal performance benchmark, 2023.

Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. Aishell-3: A multi-speaker mandarin tts corpus and the baselines. 2015. URL https://arxiv.org/abs/2010.11567.

Keshan Sodimana, Knot Pipatsrisawat, Linne Ha, Martin Jansche, Oddur Kjartansson, Pasindu De Silva, and Supheakmungkol Sarin. A Step-by-Step Process for Building TTS Voices Using Open Source Data and Framework for Bangla, Javanese, Khmer, Nepali, Sinhala, and Sundanese. In *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU)*, pages 66–70, Gurugram, India, August 2018. URL http://dx.doi.org/10.21437/SLTU.2018-14.

Shinnosuke Takamichi, Kentaro Mitsui, Yuki Saito, Tomoki Koriyama, Naoko Tanji, and Hiroshi Saruwatari. Jvs corpus: free japanese multi-speaker voice corpus. *arXiv preprint arXiv:1908.06248*, 2019.

Ashok Urlana, Pinzhen Chen, Zheng Zhao, Shay Cohen, Manish Shrivastava, and Barry Haddow. PMIndiaSum: Multilingual and cross-lingual headline summarization for languages in India. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11606–11628, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.777. URL https://aclanthology.org/2023.findings-emnlp.777.

Jörgen Valk and Tanel Alumäe. VoxLingua107: a dataset for spoken language recognition. In *Proc. IEEE SLT Workshop*, 2021.

Daniel van Niekerk, Charl van Heerden, Marelie Davel, Neil Kleynhans, Oddur Kjartansson, Martin Jansche, and Linne Ha. Rapid development of TTS corpora for four South African languages. In *Proc. Interspeech 2017*, pages 2178–2182, Stockholm, Sweden, August 2017. URL http://dx.doi.org/10.21437/Interspeech.2017-1139.

Changhan Wang, Anne Wu, and Juan Miguel Pino. Covost 2: A massively multilingual speech-to-text translation corpus. *CoRR*, abs/2007.10310, 2020. URL https://arxiv.org/abs/2007.10310.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. 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 993–1003, Online, August 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.acl-long.80.

Steve J. Young, Milica Gasic, Blaise Thomson, and Jason D. Williams. POMDP-Based Statistical Spoken Dialog Systems: A Review. *Proceedings of the IEEE*, 101(5):1160–1179, 2013. doi: 10.1109/JPROC.2012.2225812. URL https://doi.org/10.1109/JPROC.2012.2225812.

Marcely Zanon Boito, William Havard, Mahault Garnerin, Éric Le Ferrand, and Laurent Besacier. MaSS: A large and clean multilingual corpus of sentence-aligned spoken utterances extracted from the Bible. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6486–6493, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.799.

8 Appendix

IETF code	Language	Dataset	# Hours
abk	Abkhazian	CV VL	34.0 10.0
afr	Afrikaans	G-TTS VL	3.3 100.0
amh	Amharic	VL	74.0
ara	Arabic	CV VL	61.0 53.6
asm	Assamese	CV VL	1.0 143.0
ast	Asturian	CV	0.01
aze	Azerbaijani	CV VL	0.05 55.6
bak	Bashkir	CV VL	213.4 54.7
bas	Basaa	CV	0.9
bel	Belarusian	CV VL	1101.6 126.4
ben	Bengali	CV G-TTS VL	33.2 2.0 45.9
bod	Tibetan	VL	92.4
bos	Bosnian	VL	97.7
bre	Breton	CV VL	4.4 33.1
bul	Bulgarian	CV VL VP	4.6 47.7 773.3
cat	Catalan	CV VL	1638.7 80.3
ceb	Cebuano	VL	3.9
ces	Czech	CV VL VP	38.2 62.4 866.1
chv	Chuvash	CV	17.3
ckb	Central Kurdish	CV	90.5
cnh	Hakha Chin	CV	0.7
cym	Welsh	CV VL	100.6 65.6
dan	Danish	CV VL VP	3.5 25.0 728.4
deu	German	CV MLS VL VP	1091.0 1966.5 36.3 274.8
div	Maldivian	CV	32.4

Table 11: List of included languages, with corresponding amount of speech data per dataset.

IETF code	Language	Dataset	# Hours
ell	Greek	CV	13.0
		VL	60.5
		VP	573.2
eng	English	CV	2213.9
		MLS	44659.7
		VL	43.5
		VP	293.6
epo	Esperanto	CV	1355.6
		VL	8.9
est	Estonian	CV	29.9
		VL	34.7
		VP	682.7
eus	Basque	CV	79.0
		VL	25.5
ewe	Ewe	B-TTS	76.6
fao	Faroese	VL	56.4
fas	Persian	CV	312.4
		VL	52.0
fil	Tagalog	VL	82.7
fin	Finnish	CV	5.0
		VL	29.5
		VP	816.1
fra	French	CV	836.0
		MLS	1076.6
		VL	63.2
		VP	290.8
fry	Frisian	CV	41.2
gle	Irish	CV	3.2
glg	Galician	CV	4.4
		VL	66.0
glv	Manx Gaelic	VL	3.5
grn	Guarani	CV	0.4
		VL	1.0
guj	Gujarati	VL	39.5
hat	Haitian Creole	VL	86.3
hau	Hausa	B-TTS	85.6
		CV	2.3
		VL	81.1
haw	Hawaiian	VL	5.4
heb	Hebrew	VL	89.7
hin	Hindi	CV	5.0
		VL	73.2
hrv	Croatian	VL	109.4
		VP	836.6
hsb	Upper Sorbian	CV	1.5

Table 12: List of included languages, with corresponding amount of speech data per dataset.

IETF code	Language	Dataset	# Hours
hun	Hungarian	CV	9.5
		VL	68.8
		VP	851.0
hye	Armenian	CV	1.0
		VL	66.5
ibo	Igbo	CV	0.01
ina	Interlingua	CV	8.3
		VL	2.1
ind	Indonesian	CV	20.6
		VL	34.4
isl	Icelandic	Samròmur	2088.3
		VL	81.2
ita	Italian	CV	271.9
		MLS	247.4
		VL	46.6
		VP	613.8
jav	Javanese	G-TTS	3.5
		VL	24.4
jpn	Japanese	CV	37.0
		JVS	26.4
		kokoro	60.0
		VL	50.1
kab	Kabyle	CV	516.6
kan	Kannada	IISc-MILE	273.8
		VL	39.9
kat	Georgian	CV	6.6
		VL	93.4
kaz	Kazakh	CV	0.6
		VL	72.6
khm	Khmer	G-TTS	4.0
		VL	27.0
kin	Kinyarwanda	CV	1982.7
kir	Kyrgyz	CV	32.6
kmr	Northern Kurdish	CV	45.8
kor	Korean	clovacall	38.1
		kosp2e	190.9
		VL	71.5
lao	Lao	VL	36.1
lat	Latin	VL	30.9
lav	Latvian	CV	3.0
		VL	37.0
		VP	868.4
lin	Lingala	B-TTS	54.0
		VL	77.8
lit	Lithuanian	CV	7.1
		VL	78.4
		VP	796.2

Table 13: List of included languages, with corresponding amount of speech data per dataset.

IETF code	Language	Dataset	# Hours
ltz	Luxembourgish	VL	71.8
lug	Ganda	CV	363.0
mal	Malayalam	CV VL	0.5 42.5
mar	Marathi	CV VL	11.9 74.7
mdf	Moksha	CV	0.2
mhr	Eastern Mari	CV	97.5
mkd	Macedonian	CV VL	0.2 103.9
mlg	Malagasy	VL	94.5
mlt	Maltese	CV VL VP	3.9 62.9 818.1
mon	Mongolian	CV VL	6.6 65.9
mri	Māori	VL	20.5
mrj	Western Mari	CV	6.5
msa	Malay	VL	71.8
mya	Burmese	VL	31.5
myv	Erzya	CV	1.1
nan-tw	Taiwanese Hokkien	CV	0.7
nep	Nepali	CV G-TTS VL	0.01 2.8 65.6
nld	Dutch	CV MLS VL VP	72.1 1554.2 37.2 277.0
nno	Norwegian Nynorsk	CV VL	0.4 43.3
nor	Norwegian	VL	98.2
oci	Occitan	VL	13.7
ori	Odia	CV	0.8
pan	Punjabi	CV VL	1.0 42.1
pol	Polish	CV MLS VL VP	129.4 103.6 76.2 841.9
por	Portuguese	CV MLS VL VP	102.3 161.0 58.6 851.6
pus	Pashto	VL	44.7
rm-sursilv	Sursilvan	CV	2.4

Table 14: List of included languages, with corresponding amount of speech data per dataset.

IETF code	Language	Dataset	# Hours
rm-vallader	Vallader	CV	1.2
ron	Romanian	CV VL VP	8.5 59.0 834.8
rus	Russian	CV VL	149.1 67.5
sah	Sakha	CV	2.7
san	Sanskrit	VL	4.6
sat	Santali	CV	0.4
sco	Scots	VL	1.5
sin	Sinhala	VL	60.6
skr	Saraiki	CV	1.3
slk	Slovak	CV VL VP	12.9 36.6 644.5
slv	Slovenian	CV VL VP	7.6 112.3 832.1
sna	Shona	VL	21.5
snd	Sindhi	VL	48.2
som	Somali	VL	94.9
sot	Southern Sotho	G-TTS	3.2
spa	Spanish	CV MLS VL VP	380.7 917.7 33.9 301.5
sqi	Albanian	VL	67.2
srd	Sardinian	CV	0.7
srp	Serbian	CV VL	0.8 47.5
sun	Sundanese	G-TTS VL	2.1 43.6
swa	Swahili	CV VL	304.2 57.6
swe	Swedish	CV VL VP	29.5 31.3 827.4
tam	Tamil	CV IISc-MILE VL	186.1 132.2 44.1
tat	Tatar	CV VL	20.3 93.4
tel	Telugu	VL	64.3
tgk	Tajik	VL	60.5
tha	Thai	CV VL	134.7 50.8

Table 15: List of included languages, with corresponding amount of speech data per dataset.

IETF code	Language	Dataset	# Hours
tig	Tigre	CV	0.01
tpi	Tok Pisin	CV	3.3
tsn	Tswana	G-TTS	3.5
tuk	Turkmen	VL	81.8
tur	Turkish	CV MS VL	59.3 10.0 54.5
tw-akuapem	Akwapem Twi	B-TTS	59.8
tw-asante	Asante Twi	B-TTS	56.8
uig	Uyghur	CV	94.9
uig	Uyghur	THUYG-20	20.7
ukr	Ukrainian	CV VL	52.5 49.7
urd	Urdu	CV VL	38.8 35.2
uzb	Uzbek	CV VL	69.7 42.2
vie	Vietnamese	CV VL	3.8 55.5
vot	Votic	CV	0.1
war	Waray	VL	3.9
xho	Xhosa	G-TTS	3.1
yid	Yiddish	VL	35.9
yor	Yoruba	B-TTS VL	24.8 66.0
yue	Cantonese	CV	15.6
zh-CN	Chinese (PRC)	Aishell Aishell-3 CV THCHS-30 VL	151.1 60.6 104.6 25.5 40.9
zh-HK	Chinese (Hong Kong)	CV	89.5
zh-TW	Chinese (Taiwan)	CV	56.3
	Total		**90,429.5**

Table 16: List of included languages, with corresponding amount of speech data per dataset.

ENDPAGE

UTTER

HORIZON-CL4-2021-HUMAN-01 101070631

D13 First report on data and resources