Khan Academy Corpus: A multilingual corpus of Khan Academy lectures

Dominika Ďurišková, Daniela Jurášová,
Matúš Žilinec, Eduard Šubert, Ondřej Bojar
Charles University, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics,
Prague, Czech Republic
dominika.duriskova.17@gmail.com, daniela.jurasova@centrum.sk,
edasubert@gmail.com, zilinec.m@gmail.com, bojar@ufal.mff.cuni.cz

Abstract

We present the Khan Academy Corpus totalling 10122 hours in 87394 recordings across 29 languages, where 43% of recordings (4252 hours) are equipped with human-written subtitles. The subtitle texts cover a total of 137 languages. The dataset was collected from open access Khan Academy lectures, benefiting from their manual transcripts and manual translations of the transcripts. The dataset can serve in creation or evaluation of multilingual speech recognition or translation systems, featuring a diverse set of subject domains.

Keywords: Khan Academy, multilingualism, corpus, alignment, speech recognition, speech translation

1. Introduction

A fundamental prerequisite for creating a well-performing system for any NLP task is high-quality in-domain data. This is especially true with today's neural network-based speech recognition, machine translation and speech translation models. Due to the large number of natural languages and the lack of parallel data between many pairs thereof, it is worthwhile to search for new resources.

While multilingual pretrained models such as Whisper (Radford et al., 2023) are openly available, the associated training data (amounting to tens of thousands of hours of non-English speech) required to train such models is not released, presenting a challenge for open research.

Our main contribution is the creation of a multilingual corpus, a collection of audio files and their transcripts in a large number of languages, including several low-resource ones. For this purpose, we use publicly available data from the Khan Academy.[1]

Khan Academy is a non-profit educational organization that offers a library of learning resources across a wide variety of subjects. These resources include video tutorials and their transcripts in the form of video subtitles, which are translated into several languages.

The language diversity of Khan Academy makes it an ideal resource for obtaining texts for low-resource languages and subsequent translation of texts. The materials are distributed under the ShareAlike Creative Commons license.[2]

Our goal is to make use of the large multilingual library of Khan Academy's materials and convert it to a collection of audio and text data in a form better suitable for training diverse models for tasks such as speech recognition and spoken language translation.

2. Related work

2.1. Multilingual voice corpora

Several multilingual corpora across various audio domains have become available recently. Here, we list and compare a selection of recent work.

SpeechMatrix (Duquenne et al., 2022) is an extensive multilingual corpus featuring speech-to-speech translations from the authentic recordings of European Parliament sessions. It uses VoxPopuli corpus (Wang et al., 2021) as its source of unlabeled speech. The corpus contains speech alignments in 136 language pairs, amounting to a total of 418,000 hours of speech. Speech alignments (i.e. direct audio-to-audio alignment links) cover a total of 17 languages. The domain is parliamentary speech which covers a wide range of topics but in a rather specific style, typically read or spontaneous monologue. Several models were trained to evaluate the quality of the data. Due to the extensive multilingual nature of SpeechMatrix, the authors conducted experiments in multilingual speech-to-speech translation. Both the data and models are available for free to the public.

[1] https://www.khanacademy.org/

[2] https://creativecommons.org/licenses/by-sa/4.0/deed.en

LREC-COLING 2024, pages 9743–9752

Dataset	Languages	Transcribed audio
SpeechMatrix	17	418,000 hrs
VoxPopuli	23	1,800 hrs
Common Voice	144	28,800 hrs*
Multilingual LibriSpeech	8	50,000 hrs
FLEURS	102	1,200 hrs
this work	29	4,200 hrs

Table 1: Comparison of related multilingual audio datasets.
*Common Voice statistics from October 2023, 19,200 hrs validated.

The Common Voice corpus (Ardila et al., 2019) is an open-source, multi-language collection of speech and text data. Speech data is obtained from volunteer contributors from around the world. Contributors record their voice by reading individual open domain sentences. At the time of writing of this article, the dataset consists of more than 28,000 hours of speech out of which more than 19,000 hours in 114 languages are validated, i.e. another speaker of the given language has listened to the audio snippet and confirmed that it matches the text. Each dataset entry is composed of a unique audio file linked with its respective text file. Common Voice can be used mainly for Automatic Speech Recognition but also for other purposes, such as language identification. Its construction (volunteers reading prepared texts) plays a critical role in the final distribution of content: the vocabulary and variety of texts is limited but the number of speakers is much higher compared to other speech corpora.
The Multilingual LibriSpeech (Pratap et al., 2020) is a read speech dataset based on LibriVox's[3] audiobooks, suitable for training speech recognition systems. It covers 8 languages with a total of 50,500 hours of speech. The corpus contains recordings in the form of very long read speech (read fragments of books). LibriSpeech also provide trained language models for each language. The corpus is freely available for download.
Basic statistics of the mentioned corpora are shown in Table 1.

2.2. Low-resource language processing

Due to the lack of parallel data between low-resource languages, typically needed to train high-quality translation systems, it is advantageous to study mechanisms for end-to-end textless translation. The study of Liu et al. (2023) provides a systematic overview of current work on speech-to-speech translation, and reports a rise in research on low-resource languages in recent years. The authors also observe an algorithm shift from traditional neural networks to Transformer-based models.

[3] https://librivox.org/

One of the tools that enables research on speech technology in low-resource languages is FLEURS (Conneau et al., 2022), an n-way parallel speech dataset in 102 languages. It can be used for various speech tasks, such as automatic speech recognition, speech language identification, translation, and others. Unlike other datasets containing n-way parallel speech and text, FLEURS provides natural human speech and high-quality transcripts for each language. Additionally, it uses a bottom-up approach of collecting spoken utterances for aligned segments, thus obtaining higher-quality alignments compared to datasets using automatic methods.

3. Data collection

We now describe our data collection pipeline consisting of a) indexing of data sources, b) data download, and c) filtering and pre-processing.
As a first step, we manually collect all available YouTube channels containing Khan Academy content in different languages. Out of these, we filter out channels that we deem unreliable.[4] We further exclude channels that do not contain videos with human-written subtitles – this is ensured by picking a sample of 200 videos from each channel and checking for subtitles. If none of the sampled videos have manual subtitles, we consider the channel unreliable. For each kept channel, we extract all YouTube video IDs, which we mark for download. This process yields 43 channels containing approximately 120,000 videos.
To obtain the data, we use the `youtube-dl` framework.[5] We opt to limit the download to only the audio and subtitle files, discarding the video content.[6]
In order to guarantee the highest possible quality

[4] We consider a channel reliable if it has a large number of subscribers (in the order of thousands) and a large number of educational videos (in the order of hundreds).

[5] https://github.com/ytdl-org/youtube-dl

[6] Arguably, the video content, which often contains hand-written notes such as sample calculations, would be very useful for multi-modal research. We preserve YouTube video IDs, so linking the relevant videos back to our data is easy.

Most frequent	Count	Least frequent	Count
English (en)	12721	Yoruba (yo)	1
Bulgarian (bg)	7596	Belarusian (be)	1
Portuguese (pt)	6347	Igbo (ig)	1
Azerbaijani (az)	6117	Aragonese (an)	1
Korean (ko)	5437	Kurdish (ku)	1
Turkish (tr)	5018	Serbian (sr-Cyrl)	1
Brazilian Portuguese (pt-BR)	4863	Sanskrit (sa)	1
Czech (cs)	4719	Rusyn (rue)	1
Russian (ru)	4120	Guarani (gn)	1
Thai (th)	3447	Fijian (fj)	1

Table 2: The most and least frequent languages of subtitles in the corpus, with their subtitle counts.

Language Combination	Number of videos
Portuguese (Brazilian and European) (pt-BR, pt-PT)	2170
Azerbaijani and English (az, en)	564
Azerbaijani, Bulgarian, and English (az, bg, en)	305
Azerbaijani, Bulgarian, English, and Korean (az, bg, en, ko)	192

Table 3: The most frequent combinations of languages (subtitles) for one video.

of the resulting dataset, we download only those subtitles which were marked as human-generated by YouTube.

Next, we employ simple pre-processing in order to obtain a data format suitable for machine learning tasks. We do not modify the audio files because they are sufficiently clean. They predominantly consist of speech and contain minimal intervals of silence or extraneous noise. We also retain the original subtitles in `.vtt` format, which consist of transcribed speaker utterances with timestamp information. However, we find the timestamps to only approximately match the audio content. In certain videos, especially those related to mathematics, it is not trivial to directly map the audio to sentences. We additionally provide a sentence-segmented variant of all subtitle files. This is achieved by concatenating all subtitle segments in the file and subsequently splitting the concatenated segments into sentences using the punctuation-agnostic `wtp-split` segmenter (Minixhofer et al., 2023).

Finally, we create train-validation-test subsets for each audio language with more than 10 available recordings (see Table 6 in the Appendix).

4. Dataset Statistics

The final dataset consists of 87 394 recordings, out of which:

- 37588 (43%) recordings (4252 hours) have human-written subtitles in at least one language
- 49868 (57%) recordings (5870 hours) do not have any human-written subtitles, as indicated in the video meta-data

There are in total 112 040 subtitle files, which averages to 2.98 subtitle files per video (considering only the videos that have subtitles). In the following, we use terms like “subtitle hours” etc. to refer to the number of hours across the set of recordings equipped with manual subtitles.

The corpus is linguistically diverse, representing a total of 137 subtitle languages (languages in which subtitles are written) and 29 audio languages (languages covered in the speech data). The most frequent languages are English, Bulgarian, and Portuguese. The least frequent are languages like Yoruba, Belarusian, and Igbo. The 10 most frequent languages and some of the least frequent languages can be found in Table 2.

We also examine what combinations of languages found in subtitles are available for each video. Contrary to the mathematically established average of 2.98 subtitles per video, there is usually only 1 subtitle file per video, and if there are more subtitles, the combinations of languages are rather unique. A list of the 4 most frequent combinations of languages can be found in Table 3, revealing a surprisingly well-covered Azerbaijani-English pair.

Note that the pairs of Portuguese video subtitle files (pt-BR and pt-PT) are mostly identical, and if they differ, it is only in a few sentences and the rest of the text is the same. The high number of videos with two versions of Portuguese is thus observed

Language	Files with transcript		All audio files		Transcribed
	Hours	Recordings	Hours	Recordings	In %
English (en)	1346.53	11271	1498.78	12333	89.84
Portuguese (pt)	835.78	7119	935.14	7958	89.38
Russian (ru)	462.82	3554	612.44	4887	75.57
Arabic (ar)	280.02	1473	477.85	2814	58.60
Gujarati (gu)	276.99	2588	396.29	3542	69.90
Turkish (tr)	235.23	3186	1134.69	10498	20.73
Hebrew (he)	199.28	1408	217.87	1457	91.47
Vietnamese (vi)	150.64	1599	210.99	2631	71.40
Spanish (es)	74.58	761	1032.90	7830	7.22
Japanese (ja)	52.74	690	157.46	1671	33.49
Georgian (ka)	52.27	708	600.94	5208	8.70
Polish (pl)	50.17	499	493.57	4428	10.17
Bulgarian (bg)	35.71	285	222.42	2144	16.06
Greek (el)	34.99	560	74.77	890	46.81
Czech (cs)	29.73	320	140.87	1528	21.10
Tamil (ta)	27.17	458	213.74	2165	12.71
Hungarian (hu)	27.16	384	31.10	442	87.33
Ukrainian (uk)	22.40	176	22.92	180	97.74
Chinese (zh)	12.85	88	266.31	1896	4.82
Latvian (lv)	12.23	178	12.53	184	97.65
German (de)	8.01	61	132.36	1186	6.05
Armenian (hy)	7.20	108	334.64	3854	2.15
Norwegian (no)	5.43	63	24.09	298	22.52
French (fr)	0.75	13	581.94	4605	0.13
Serbian (sr)	0.52	9	16.69	198	3.09
Hindi (hi)	0.47	4	191.39	1576	0.25
Italian (it)	0.41	6	68.55	797	0.60
Korean (ko)	0.27	5	2.43	47	11.19
Lithuanian (lt)	0.02	1	6.97	142	0.23

Table 4: Number of recordings and total hours for each of the 29 audio languages. Note that audio files do not necessarily have manually revised subtitles in the same language as the audio. The first two columns ("Files with transcripts") consider all files where the subtitles cover the given language. For contrast, the second two columns consider all audio files by their spoken language. Some languages such as French, Hindi or Lithuanian have less than 1 % audios transcribed. Automatic language detection was used to generate per-language file counts.

only because the same subtitle files are labelled for both regions.

Table 4 summarizes the number of hours of speech available across the covered languages. Available hours of speech for each pair of subtitled languages are listed in Figure 2 in the Appendix.

5. Cross-Lingual Subtitle Alignment

Alignment in a multi-language corpus serves as a foundational step for many subsequent tasks. It allows for the comparison and analysis of linguistic components, such as words, phrases, or sentence structures, across different languages. To this end, we created automatic alignment at the level of subtitles, as discussed in this section.[7]

[7]We do not handle audio-to-text alignment. It is to some extent provided in the subtitles themselves in the `.vtt` files.

5.1. Using temporal information

Each subtitle file is in the `.vtt` file format, containing metadata and timestamps indicating when the associated text was spoken. This temporal information offers an opportunity to attempt cross-language text alignment. However, upon manual analysis, we determined that achieving precise alignment using this timing data is impossible. While absolute timestamps were in general conserved across languages, the same timestamp would often point to different parts of a sentence in two languages. Additionally, some subtitles had long or inaccurate timing windows, making fine-grained alignment difficult.

5.2. Using sentence similarity

Given the extensive amount of multilingual data we were handling, which contained less common languages, finding an appropriate alignment tool

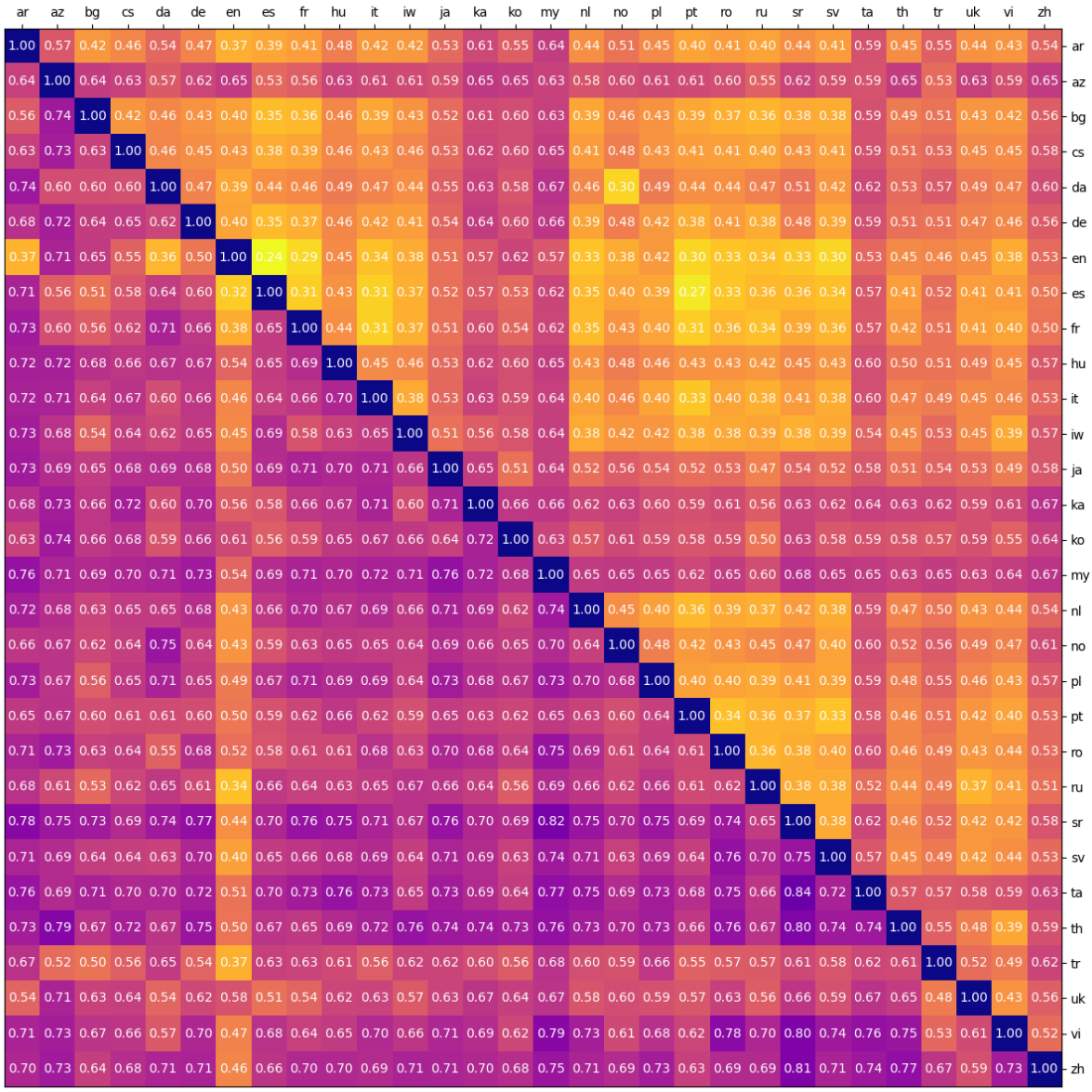

Figure 1: Subtitle alignment quality per every language pair. The upper right triangle shows the average alignment score from `vecalign`. The bottom left triangle shows the proportion of 1-1 alignments, i.e. alignments where the subtitle chunk in the source language corresponds to exactly one subtitle chunk in the target language.

was challenging.

We experimented with several aligners, however, many support only a limited number of languages or necessitate additional files like dictionaries, translations, or pre-existing alignments. Moreover, the tested aligners struggled with languages utilizing scripts different from the Latin script, such as Chinese, resulting in a large number of misaligned sentences.

In the end, we use the `vecalign` multilingual aligner (Thompson and Koehn, 2019), which empirically outperformed the statistical aligner Hunalign (Varga et al., 2007) in our experiments. The aligner is based on sentence embedding similarity. We generate sentence embeddings using LASER (Artetxe and Schwenk, 2019). For each video with more than one subtitle language, we run the aligner and obtain alignment of subtitle chunks and full sentences for all possible language pairs.

The quality (as reported by `vecalign`) of the obtained alignment prepared at the level of individual subtitles for each language pair is displayed in Figure 1. We also create alignments of the sentences as automatically identified by `wtpsplit` and present their quality in Appendix Figure 3. According to the scores, the alignment at the level of sentences seems lower.

	Base WER $\mu \pm \sigma$	Large-v2 WER $\mu \pm \sigma$
Russian (ru)	0.55 +- 0.50	0.10 +- 0.06
Portuguese (pt)	0.42 +- 0.39	0.12 +- 0.11
German(de)	0.40 +- 0.30	0.13 +- 0.04
Spanish (es)	0.35 +- 0.30	0.13 +- 0.07
Turkish (tr)	0.77 +- 0.73	0.15 +- 0.12
Czech (cs)	0.68 +- 0.34	0.18 +- 0.15
Polish (pl)	0.55 +- 0.51	0.18 +- 0.18
Hungarian (hu)	0.74 +- 0.56	0.20 +- 0.08
Ukrainian (uk)	0.63 +- 0.40	0.20 +- 0.15
Latvian (lv)	1.14 +- 0.45	0.21 +- 0.05
Norwegian (no)	0.52 +- 0.18	0.23 +- 0.07
English (en)	0.34 +- 0.72	0.24 +- 0.91
Vietnamese (vi)	0.60 +- 0.35	0.25 +- 0.20
Greek (el)	0.70 +- 0.33	0.30 +- 0.20
Tamil (ta)	0.78 +- 0.25	0.40 +- 0.20
Hebrew (he)	1.26 +- 0.81	0.49 +- 0.48
Japanese (ja)	1.45 +- 1.17	0.52 +- 0.23
French (fr)	0.61 +- 0.44	0.56 +- 0.47
Armenian (hy)	1.42 +- 0.55	0.60 +- 0.11
Chinese (zh)	3.31 +- 2.75	0.96 +- 0.06
Gujarati (gu)	1.00 +- 0.06	0.99 +- 0.02
Georgian (ka)	1.50 +- 0.65	1.01 +- 0.07
Bulgarian (bg)	1.13 +- 0.16	1.12 +- 0.20

Table 5: Word error rates for Whisper models.

6. Speech Recognition Experiments

To complement our released data with a baseline, we test a state-of-the-art multilingual speech recognition model with it.

We do not train any models ourselves, but to allow for such a comparison later, we divide the corpus into training, development and test sections. Table 6 in the Appendix summarizes the sizes of these sections across all audio languages.

We evaluate multilingual Whisper Base and Large v2 (Radford et al., 2023), two pretrained speech recognition models with 74M and 1550M parameters, respectively, pretrained on 99 languages where 17% is non-English audio+text. We do not initialize the model with a target language token, relying instead on a built-in language detector using the first 30 seconds of each recording.

We calculate Word Error Rate (WER) on model outputs with minimal postprocessing, which comprises of stripping any non-alphanumeric characters in both reference and model output. The results are presented in Table 5.

We observe that in line with our expectation, model performance improves with the number of parameters as well as with language availability in Whisper training data. Interestingly, the best performing languages are Russian, Portuguese, German and Spanish, not English. One possible explanation for this result is that the English speech need not be native accent. The model performs comparatively well in Slavic languages incl. Czech or Polish, while underperforming mainly in languages with a non-Latin script.

7. Availability

Khan Academy Corpus is available under the ShareAlike Creative Commons license (CC-BY-SA 4.0) at the following URL:

http://hdl.handle.net/11234/1-5475

8. Conclusion

We created Khan Academy Corpus, a multilingual speech and translation corpus using Khan Academy materials. The resulting corpus includes data from 87,394 videos, with 43 % having manual subtitles. We supplemented the corpus with automatically created bilingual alignments of subtitles for all available language pairs for the purposes of machine and speech translation training.

We envision this dataset can be used to improve the performance of automatic systems for subtitling or translation of speech where advanced vocabulary is needed, like technical conferences, lectures, and similar. We show that the state-of-the-art pretrained Whisper models are practically usable for speech recognition on clean recordings in this domain.

9. Acknowledgements

This work was partially supported by GAČR EXPRO grant NEUREM3 (19-26934X) and by the Grant Agency of Charles University in Prague (GAUK 244523).

10. References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

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.

Paul-Ambroise Duquenne, Hongyu Gong, Ning Dong, Jingfei Du, Ann Lee, Vedanuj Goswani, Changhan Wang, Juan Pino, Benoît Sagot,

and Holger Schwenk. 2022. Speechmatrix: A large-scale mined corpus of multilingual speech-to-speech translations. *arXiv preprint arXiv:2211.04508*.

Hsiao-Chuan Liu, Min-Yuh Day, and Chih-Chien Wang. 2023. Speech-to-speech low-resource translation. In *2023 IEEE 24th International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 91–95.

Benjamin Minixhofer, Jonas Pfeiffer, and Ivan Vulić. 2023. Where's the point? self-supervised multilingual punctuation-agnostic sentence segmentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7215–7235, Toronto, Canada. Association for Computational Linguistics.

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

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 1342–1348.

Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2007. Parallel corpora for medium density languages. *Amsterdam Studies In The Theory And History Of Linguistic Science Series 4*, 292:247.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. *arXiv preprint arXiv:2101.00390*.

Appendix

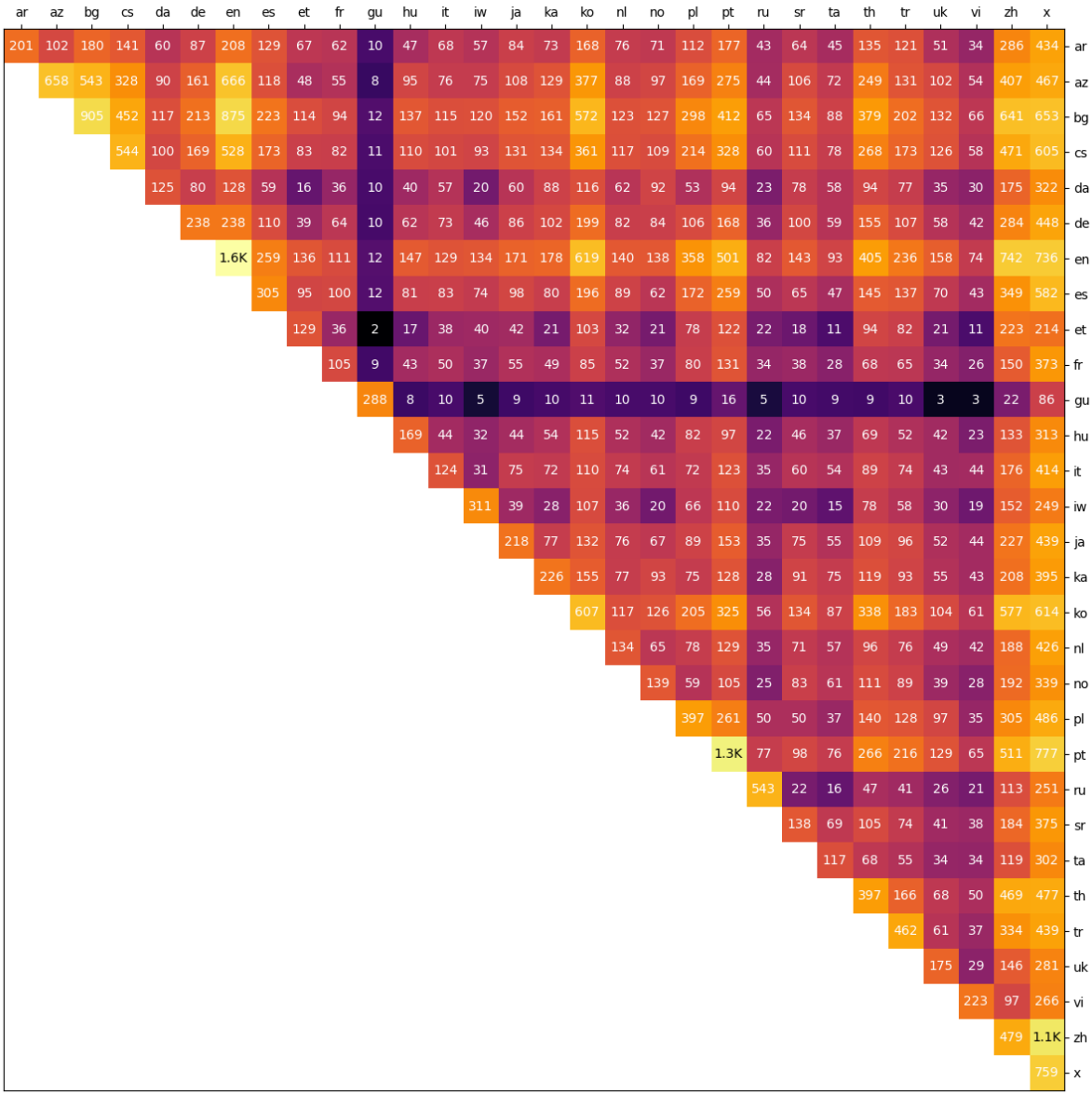

Figure 2: The heatmap shows available subtitle-hours for each pair of languages. Where multiple dialects exist (e.g. en-US, en-GB), we report the sum thereof. The total available subtitle-hours for each language (irrespective of audio language) is shown in the diagonal. Note that this is different from the row sum since the same segment of audio translated into multiple languages is counted only once here. Languages with less than 100 hours in total were aggregated (denoted by 'X') due to space limitations.

Language	Training		Validation		Testing	
	Recordings	Hours	Recordings	Hours	Recordings	Hours
Arabic (ar)	1179	223.4	147	28.8	148	28.0
Bulgarian (bg)	228	28.8	29	3.6	29	3.5
Czech (cs)	256	23.6	32	3.0	32	3.1
German (de)	48	6.4	6	0.8	7	0.8
Greek (el)	448	28.2	56	3.4	56	3.4
English (en)	9020	1081.8	1127	133.0	1128	134.0
Spanish (es)	611	60.0	76	7.3	77	7.6
French (fr)	10	0.6	1	<0.1	2	0.1
Gujarati (gu)	2070	221.9	259	26.8	259	28.3
Hebrew (he)	1126	158.6	141	21.4	141	19.3
Hungarian (hu)	307	22.0	38	2.7	39	2.5
Armenian (hy)	88	6.0	11	0.5	11	0.9
Japanese (ja)	552	42.5	69	5.1	69	5.2
Georgian (ka)	566	41.6	71	5.3	71	5.4
Latvian (lv)	142	9.2	18	1.5	18	1.5
Norwegian (no)	50	4.3	6	0.4	7	0.6
Polish (pl)	433	44.3	54	5.4	55	5.6
Portuguese (pt)	5696	667.1	712	85.6	713	83.3
Russian (ru)	2844	371.0	355	44.9	356	47.1
Tamil (ta)	366	21.9	46	2.8	46	2.5
Turkish (tr)	2554	189.0	319	22.9	320	24.1
Ukrainian (uk)	140	18.4	18	1.9	18	2.1
Vietnamese (vi)	1281	120.6	160	15.1	161	15.3
Chinese (zh)	70	10.3	9	1.1	9	1.4

Table 6: Resulting dataset statistics after splitting into training, validation and testing subsets, in ratio of 80/10/10% for each language. Languages with less than 10 available recordings were excluded.

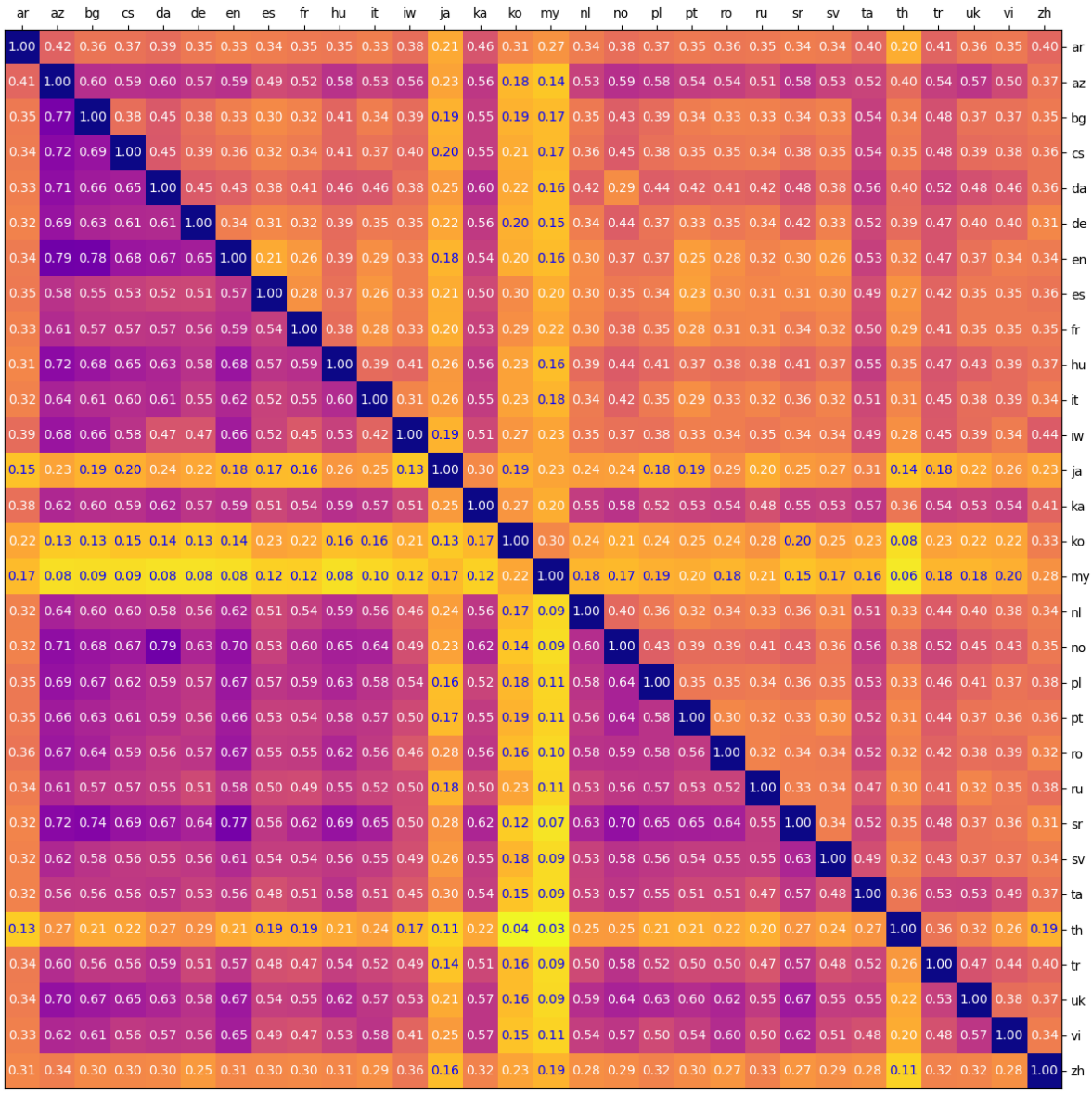

Figure 3: Alignment quality after `wtpsplit` sentence segmentation per every language pair. The upper right triangle shows average aligment score from `vecalign`. The bottom left triangle shows average identity of alignment, which we define as the proportion of subtitle chunks in source language aligned to exactly one subtitle chunk in target language.