Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6486–6493
Marseille, 11–16 May 2020

MaSS: A Large and Clean Multilingual Corpus of Sentence-aligned Spoken Utterances Extracted from the Bible

Marcely Zanon Boito[*1], **William N. Havard**[*1,2], **Mahault Garnerin**[1,2], **Éric Le Ferrand**[1], **Laurent Besacier**[1]
[1]LIG, Univ. Grenoble Alpes, CNRS, Grenoble INP, F-38000 Grenoble, France
[2]LIDILEM, Univ. Grenoble Alpes, F-38000 Grenoble, France
{first.last-name}@univ-grenoble-alpes.fr
[*]Both authors have contributed equally to this paper.

Abstract

The *CMU Wilderness Multilingual Speech Dataset* (Black, 2019) is a newly published multilingual speech dataset based on recorded readings of the New Testament. It provides data to build Automatic Speech Recognition (ASR) and Text-to-Speech (TTS) models for potentially 700 languages. However, the fact that the source content (the Bible) is the same for all the languages is not exploited to date. Therefore, this article proposes to add multilingual links between speech segments in different languages, and shares a large and clean dataset of 8,130 parallel spoken utterances across 8 languages (56 language pairs). We name this corpus MaSS (**M**ultilingu**a**l corpus of **S**entence-aligned **S**poken utterances). The covered languages (Basque, English, Finnish, French, Hungarian, Romanian, Russian and Spanish) allow researches on speech-to-speech alignment as well as on translation for typologically different language pairs. The quality of the final corpus is attested by human evaluation performed on a corpus subset (100 utterances, 8 language pairs). Lastly, we showcase the usefulness of the final product on a bilingual speech retrieval task.

Keywords: parallel speech corpus, multilingual alignment, speech-to-speech alignment, speech-to-speech translation, speech retrieval

1. Introduction

Recently, a remarkable work introduced the *CMU Wilderness Multilingual Speech Dataset* (Black, 2019).[1] Based on readings of the New Testament from *The Faith Comes By Hearing* website, it provides data to build Automatic-Speech-Recognition (ASR) and Text-to-Speech (TTS) models for potentially 700 languages. Such a resource allows the community to experiment and to develop speech technologies on an unprecedented number of languages. However, the fact that the initial language material from these monolingual corpora (the Bible) is the same for all languages, thus constituting a multilingual and comparable[2] spoken corpus, is not exploited to date.

Therefore, this article proposes an automatic pipeline for adding multilingual links between small speech segments in different languages. We apply our method to 8 languages (Basque, English, Finnish, French, Hungarian, Romanian, Russian and Spanish), resulting in 56 language pairs for which we obtain speech-to-speech, speech-to-text and text-to-text alignments. In order to ensure the quality of the pipeline, a human evaluation was performed on a corpus subset (8 language pairs, 100 sentences) by bilingual native speakers. The current version of our dataset (named MaSS for **M**ultilingu**a**l corpus of **S**entence-aligned **S**poken utterances) is freely available to the community, together with instructions and scripts allowing the pipeline extension to new languages.[3]

We believe the obtained corpus can be useful in several applications, such as speech-to-speech retrieval (Lee et al., 2015), multilingual speech representation learning (Harwath et al., 2018a) and direct speech-to-speech translation (so far, mostly direct speech-to-text translation has been investigated (Bérard et al., 2016; Weiss et al., 2017; Bansal et al., 2017; Bérard et al., 2018)). Moreover, typological and dialectal fields could use such a corpus to solve some of the following novel tasks using parallel speech: word alignment, bilingual lexicon extraction, and semantic retrieval.

This paper is organized as follows: after briefly presenting related works in Section 2, we review the dataset source and extraction pipeline in Section 3. Section 4 describes the human verification performed and comments on some of the linguistic features present in the covered languages. Section 5 presents a possible application of the dataset: speech-to-speech retrieval. Section 6 concludes this work.

2. Related Work

2.1. End-to-end Speech Translation

Previous Automatic Speech-to-Text Translation (AST) systems operate in two steps: source language Automatic Speech Recognition (ASR) and source-to-target text Machine Translation (MT). However, recent works have attempted to build end-to-end AST without using source language transcription during learning or decoding (Bérard et al., 2016; Weiss et al., 2017), or by using it at training time only (Bérard et al., 2018). Very recently several extensions of these pioneering works were introduced: low-resource AST (Bansal et al., 2019), unsupervised AST (Chung et al., 2018), end-to-end speech-to-speech translation (*Translatotron*) (Jia et al., 2019b). Improvements for end-to-end AST were also proposed by using weakly supervised data (Jia et al., 2019a), or by adding a second attention mechanism (Sperber et al., 2019).

[1]Available at `http://www.festvox.org/cmu_wilderness/index.html`

[2]Our definition of a *comparable* corpus in this work is the following: a non-sentence-aligned corpus, parallel at a broader granularity (e.g. chapter, document).

[3]Available at `https://github.com/getalp/mass-dataset`

2.2. Multilingual Approaches

Multilingual approaches for speech and language processing are growing ever more popular. They are made possible by the availability of massively parallel language resources covering an increasing number of languages of the world. These resources feed truly multilingual approaches, such as machine translation (Aharoni et al., 2019), syntax parsing (Nivre et al., 2016), automatic speech recognition (Schultz and Schlippe, 2014; Adams et al., 2019), lexical disambiguation (Navigli and Ponzetto, 2010; Sérasset, 2015), and computational dialectology (Christodoulopoulos and Steedman, 2015).

2.3. Corpora for End-to-end Speech Translation

To date, few datasets are available for multilingual automatic speech translation (only a few parallel corpora publicly available[4]). For instance, *Fisher* and *Callhome* Spanish-English corpora (Post et al., 2013) provide 38 hours of speech transcriptions of telephonic conversations aligned with their translations. However, these corpora are only medium size and contain low-bandwidth recordings. Microsoft Speech Language Translation (MSLT) corpus (Federmann and Lewis, 2016) also provides speech aligned to translated text, but this corpus is rather small (less than 8 hours per language). A 236 hours extension of *Librispeech* with French translations was proposed by Kocabiyikoğlu et al. (2018). They exploited automatic alignment procedures, first at the text level (between transcriptions and translations), and then between the text and the corresponding audio segments.

Inspired by this work, Di Gangi et al. (2019) created MuST-C, a multilingual speech translation corpus for training end-to-end AST systems from English into 8 languages.[5] Similar in size, the English-Portuguese dataset *How2* (Sanabria et al., 2018) was created by translating English short tutorials into Portuguese using a crowd-sourcing platform. More recently, Iranzo-Sánchez et al. (2020) introduced a multilingual speech corpus including several source languages. The remark that can be made on all these corpora is that they are limited to Indo-European languages and thus typologically similar.

3. A Large and Clean Subset of Sentence Aligned Spoken Utterances (MaSS)

In this section we present the source material for our multilingual corpus (Section 3.1.), we briefly explain the CMU speech-to-text pipeline (Section 3.2.), and we detail our speech-to-speech pipeline (Section 3.3.).

3.1. The Source Material: Bible.is

The *Faith Comes By Hearing* website[6] (or simply *bible.is*) is an online platform that provides audio-books of the Bible with transcriptions in 1,294 languages. These recordings are a collection of field, virtual and partner recordings. In all cases, only native speakers participate in the recordings, and the number of different voices can go from one up to twenty five. Moreover, the recordings can be performed in *drama* and *non-drama* fashion, the former being an acted version of the text, corresponding to less tailored realizations. Finally, based on exchanges with the target users (the native community), background music can be added to the recordings.[7] In summary, while the written content is always the same across different languages, the corresponding speech can be quite different in terms of realization (drama and non-drama), number of speakers, acoustic quality (field, virtual or partner recordings), and can sometimes contain background noise (music).

3.2. The CMU Wilderness Multilingual Speech Corpus

The CMU Wilderness corpus (Black, 2019) is a speech dataset containing over 700 different languages for which it provides audio excerpts aligned with their transcription. Each language accounts for around 20 hours of data extracted from readings of the New Testament, and available at the *bible.is* website. Segmentation was made at the sentence level, and alignment between speech and corresponding text can be obtained with the pipeline provided along with the dataset. This pipeline, notably, can process a large amount of languages without using any extra resources such as acoustic models or pronunciation dictionaries.

However, for most of the languages on the website, several recording versions are available, each of them having significant differences in speech content, as explained in Section 3.1. As this pipeline extracted the soundtracks from the defaults links, audio excerpts often contain music, and it is unknown if drama or non-drama versions were selected. Thus, although the quality of the alignment is good for many languages, it could be inaccurate (or noisy) for an unknown subset.

Lastly, the final segmentation from chapters was obtained through the use of punctuation marks. While efficient for a speech-to-text monolingual scenario, this strategy does not allow accurate multilingual alignment, since different languages and translations may result in different sentence segmentation and ordering.

3.3. Our Pipeline: from Speech-to-text to Speech-to-speech Alignment

As far as multilingual alignment is concerned, Bible chapters are inherently aligned at the *chapter* level. But Bible chapters are very long excerpts, with an average duration of 5 minutes. Alignments at this broad granularity are not relevant for research in speech-to-speech translation or speech-to-speech alignment. Thus, we propose a new extraction methodology that allows us to obtain fully aligned speech segments at a much smaller granularity (segments between 8 to 10 seconds). Our pipeline is summarized in Figure 1 and described below.

3.3.1. Alignment pipeline

1. Extracting clean spoken chapters. Starting from the pipeline described in the last section, which provides scripts

[4]Table 1 in (Di Gangi et al., 2019) provides a good survey.

[5]Available at `https://ict.fbk.eu/must-c`

[6]Available at `https://www.bible.is`

[7]More information available at `https://www.faithcomesbyhearing.com/mission/recordings`

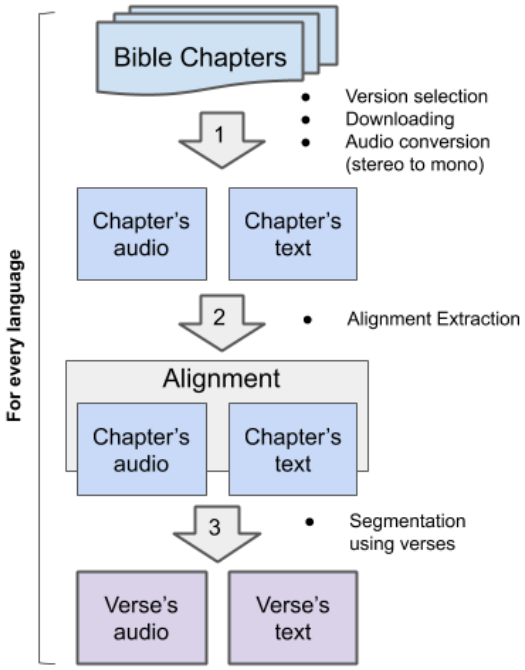

Figure 1: The pipeline for a given language in the *bible.is* website.

for downloading audio data and transcriptions from the *bible.is* website, we downloaded all the 260 chapters from the New Testament in several languages. We selected (after having manually sampled the website) *non-drama* versions (as opposed to *drama*) that contain standard speech and pronunciation, and mostly, no background music. The audios are also converted from stereo to mono for the purpose of the following steps.

2. Aligning speech and text for each chapter. For each chapter, we extracted speech-to-text alignments through the *Maus* forced aligner[8] (Kisler et al., 2017) online platform. During this step, we kept languages with good audio quality and for which an acoustic model was available in the off-the-shelf forced aligner tool. Our final set was reduced to the following eight languages: Basque, English, Finnish, French, Hungarian, Romanian, Russian and Spanish.

3. Segmenting chapters into verses. Any written chapter of the Bible is inherently segmented into *verses*. A *verse* is the minimal segmentation unit used in the Bible and corresponds to a sentence, or more rarely to a phrase or a clause. In order to segment our audio files in such smaller units, we aligned our *TextGrid files* (from step 2) with a written version of the Bible containing *verse* information. This alignment is rather trivial, since, after removing punctuation, both texts have the same content. After this step, all audio chapters are segmented into verses and receive *IDs* based on their English chapter name, and their verse number (e.g. "Matthew_chapter1_verse3").

[8]Available at `https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface/WebMAUSBasic`

3.3.2. Result and Comparison

Considering that all chapters consist of the same set of verses, the verse numbers give us a multilingual alignment between all verses for all the language pairs.[9] Thus, the output of our pipeline is a set of 8,160 audios segments, aligned at verse-level, in eight different languages, with an average of 20 hours of speech for each language. Finally, corpus statistics are presented in Table 1.

For justifying the need of extending the approach presented in Section 3.2, Table 2 presents a comparison between our corpus (bottom) output and theirs (top). This comparison takes the speech file numbering on their pipeline as multilingual alignment clue, since no other information is available. We can observe that by segmenting based on punctuation, the multilingual alignment quickly becomes incorrect: the segmentation on the third file, based on a punctuation mark not present in the English text, shifts the alignment for the rest of the chapter.

3.3.3. Reproducibility

The presented pipeline performs automatic verse-level alignment using Bible chapters. All the scripts used in this work are available, together with the resulting dataset.[3] For extending it to a new language, here are some recommendations:

- **Bible version:** as discussed in Section 3.1, a language can have several versions available on the website. For ensuring the best quality possible, manual inspection in one chapter can be quickly performed to identify a non-drama version, but it is not mandatory.
- **Alignment Tool:** for generating verse-level alignment, a chapter-level alignment between speech and text is needed. While we use the *Maus forced aligner* for this task, any aligner able to provide a *TextGrid file* as output can be used at this stage.

4. Resource Evaluation and Analysis

4.1. Human Evaluation: Speech Alignment Quality

Having obtained multilingual alignments between spoken utterances, we attest their quality by performing a human evaluation on a corpus subset, covering the eight language pairs for which we were able to find bilingual judges.

We implemented an online evaluation platform with 100 randomly selected verses in these 8 different language pairs. Judges were asked to evaluate the spoken alignments using a scale from 1 to 5 (1 meaning the two audio excerpts do not have any information in common, and 5 meaning they are perfectly aligned). Aiming at the most uniform evaluation possible, we provided guidelines and examples to our evaluators. Transcriptions were also displayed as a cognitive support in evaluation.

The eight language pairs are the following: French-English (FR-EN), French-Spanish (FR-ES), French-Romanian (FR-RO), English-Spanish (EN-ES), English-

[9]This is mostly true, but for a small subset of chapters, due to different Bible versions and different translation approaches, the number of aligned speech verses will differ slightly.

Languages	# types	# tokens	Types per verse	Tokens per verse	Avg. token length	Audio length (h)	Avg. verse length (s)
(EN) English	6,471	176,461	18.03	21.52	3.82	18.50	8.27
(ES) Spanish	11,903	168,255	17.90	20.52	4.17	21.49	9.58
(EU) Basque	14,514	128,946	14.88	15.78	5.55	22.76	9.75
(FI) Finnish	18,824	134,827	15.04	16.44	5.66	23.16	10.21
(FR) French	10,080	183,786	19.25	22.36	4.02	19.41	8.62
(HU) Hungarian	20,457	135,254	15.01	16.46	5.07	21.12	9.29
(RO) Romanian	9,581	169,328	18.19	20.61	4.14	23.11	10.16
(RU) Russian	16,758	129,973	14.50	15.82	4.44	22.90	9.70

Table 1: Statistics of the MaSS corpus.

Alignment from Black (2019)		
Files	**French**	**English**
00001	Matthieu	Matthew
00002	Jésus descend de la montagne et des foules nombreuses le suivent.	When he came down from the mountainside, large crowds followed him.
00003	Un lépreux s'approche, il se met à genoux devant Jésus et lui dit :	A man with leprosy came and knelt before him and said, *"Lord, if you are willing, you can make me clean."*
00004	Seigneur, si tu le veux, tu peux me guérir !	*Jesus reached out his hand and touched the man. "I am willing," he said. "Be clean!" Immediately he was cured of his leprosy.*
Our alignment		
Verses	**French**	**English**
00	Matthieu 8	Matthew 8
01	Lorsque Jésus fut descendu de la montagne une grande foule le suivit	When he came down from the mountain great crowds followed him
02	Et voici un lépreux s'étant approché se prosterna devant lui et dit : Seigneur si tu le veux tu peux me rendre pur	And behold a leper came to him and knelt before him saying Lord if you will you can make me clean
03	Jésus étendit la main le toucha et dit : Je le veux sois pur Aussitôt il fut purifié de sa lèpre	And Jesus stretched out his hand and touched him saying I will be clean And immediately his leprosy was cleansed

Table 2: A comparison between CMU's multilingual alignment and ours. Text in italic shows alignment mismatches between English and French. We used a slightly different (*non-drama*) version of the Bible, hence the small differences in the displayed texts.

Finnish (EN-FI), English-Hungarian (EN-HU), English-Romanian (EN-RO) and English-Russian (EN-RU). This selection is a trade-off between the difficulty of finding judges and the desire to provide a good typological variety in our evaluation data. Basque was also chosen due to the fact it is language isolate, that is, a language that has no known connection to any other language. However, we were unable to find judges to perform the evaluation on any language pair including it.

Table 3 summarizes the results of the human evaluation. Evaluation scores are good, with a mean value of 4.41. Moreover, for every language pair evaluated (except for FR-ES), the median score is the maximum score, hence confirming the quality of the alignment. However, when trying to quantify rater's agreement, we obtained mixed results. Percentage of agreement with tolerance 1 (meaning raters differing by one-scale degree are interpreted as agreeing) varies from 59.6% (EN-RO) to 95.96% (EN-HU).

	$\bar{x}$	σ	**med**	**min**	**max**	**# Eval.**
EN - ES	4.56	0.62	5	3	5	2
EN - FI	4.37	0.92	5	1	5	1
EN - HU	4.44	0.88	5	1	5	2
EN - RO	4.24	0.97	5	1	5	6
EN - RU	4.56	0.83	5	1	5	3
FR - EN	4.38	0.79	5	1	5	5
FR - ES	4.22	0.89	4	2	5	2
FR - RO	4.51	0.90	5	1	5	1
All	**4.36**	**0.88**	**5**	**1**	**5**	**22**

Table 3: Result of the manual inspection of the speech alignment quality performed on 8 language pairs (100 sentences). Scale is from 1 to 5 (higher is better). Last column refers to the number of evaluators for a given language pair.

4.2. Corpus Linguistic Analysis

Regarding content, the corpus features languages belonging to different families. These are listed as follows:

- Indo-European:
  - Romance: **French**, **Romanian**, **Spanish**
  - Germanic: **English**
  - Slavic: **Russian**
- Uralic:
  - Ugric: **Hungarian**
  - Finnic: **Finnish**
- Language Isolate: **Basque**

It should be noted that these languages are very different from a typological point of view. First of all, Basque, Finnish, Hungarian, Romanian and Russian mainly use case marking to indicate the function of a word[10] in a sen-

[10]Case markers are small grammatical morphemes added to a

tence, while English, French and Spanish rely on word position and prepositions for the same purposes. Basque, Finnish and Hungarian are agglutinative languages, while English, French, Romanian, Russian and Spanish are fusional languages. Thus, for the former group, grammatical markers will bear only one meaning, while in the latter, grammatical markers will bear several meanings at the same time.[11]

Basque is even more special as this language features ergative-absolutive marking while the other languages use nominative-accusative marking. In languages using ergative-absolutive marking, the subject of an intransitive verb and the patient of a transitive verb are treated alike and receive the same case marker, while the agent of a transitive verb is treated differently than the subject of an intransitive verb. Romanian also presents an interesting morphological characteristic regarding determiners: the definite article is suffixed to the word whereas indefinite articles are usually prefixed, for instance: "un-băiat" (INDEF-boy: "a boy") and "băiat-ul" (boy-DEF: "the boy"). Finnish and Russian on the other hand do not have any article, neither definite nor indefinite.

Another interesting linguistic phenomenon to observe is the existence of grammatical genders. Russian features three genders (feminine, masculine and neutral) whereas French features only two (feminine and masculine), and Basque and Finnish present no grammatical genders at all. From a syntactic point of view, English, French and Spanish have a relatively fixed word order (and mainly follow the Subject-Verb-Object (SVO) pattern), while word order is more flexible in Basque, Finnish, Hungarian, Romanian and Russian, mainly due to the fact that these languages use case markers.

Due to all the diverse linguistic features described in this section, we believe this dataset could be used for a wide variety of tasks, such as natural language grammar induction from raw speech, automatic typological features retrieval, speech-to-speech translation, and speech-to-speech retrieval. The latter is illustrated on Section 5. Moreover, this dataset could also serve as a benchmark for evaluating computational language documentation techniques that work on speech inputs.

5. Use Case: Multilingual Speech Retrieval Task Baseline

In this section we showcase the usefulness of our corpus on a multilingual setting. We perform speech-to-speech retrieval by adapting a model for visually grounded speech (Harwath et al., 2018b), and we discuss the results for our *baseline model*.

word to indicate its grammatical function (eg. subject, object, etc.) within a clause/sentence.

[11] Compare Hungarian "ház-ak-nak" (house-PL-DAT) and Russian "дом-ам" (house-PL.DAT). Words in agglutinative languages are comparatively longer than their equivalent in fusional languages.

5.1. Task and Model Definition

For performing multilingual speech retrieval, we adapted the model[12] proposed by Harwath et al. (2018b). This model was primarily designed to retrieve images from speech utterances, and it is made of two networks: a speech and a image encoder. By projecting both representations to the same shared space, the model is thus able to learn the relationship between speech segments and the image contents. For our speech-to-speech task, we replaced the image encoder by a (second) speech encoder.[13]

Both speech encoders consist of a convolution bank (Wang et al., 2017) followed by two layers of bidirectional LSTM (Hochreiter and Schmidhuber, 1997), and of an attention mechanism (Bahdanau et al., 2015) which computes a weighted average of the LSTM's activations. The convolution bank consists of a set of $K = 16$ 1D-convolution filters, where the k^{th} convolution has a kernel of width k. Each convolution filter consists of 40 units with ReLU activation and stride of 1. The batch-normed output of each convolution is then stacked and the resulting matrix is linearly projected to fit the LSTM's input dimension of size 256.

Our model's inputs are Mel filterbank spectrograms (40 mel coefficients with a Hamming window size of 25ms and stride of 10ms) extracted from raw speech. The network is trained to minimize the contrastive loss function in Equation 1, which minimizes the cosine distance d between a verse in a given language A, and its corresponding verse in a given language B. It does so by maximizing the distance between mismatching verses pairs (with a given margin α). Thus, verses corresponding to direct translations should lie close in the embedding space. Finally, contrary to Harwath et al. (2018a), in which only one negative example for caption is sampled, we adopted the method from Chrupała et al. (2017), considering every other verse in the batch as a negative example.

$$L(v_A, v_B, \alpha) = \sum_{v_A, v_B} \left(\sum_{v'_A} \max[0, \alpha + d(v_A, v_B) - d(v'_A, v_B)] + \sum_{v'_B} \max[0, \alpha + d(v_A, v_B) - d(v_A, v'_B)] \right) \quad (1)$$

5.2. Results

We trained an instance of this model for seven language pairs, always keeping English as source language. The 8,160 common verses were randomly split between train (80%), validation (10%) and test (10%) sets. Batches were of size 16, and models were all trained for 100 epochs. Table 4 presents our results for the retrieval task.

Results show that, while such a speech-to-speech task is challenging, it is possible to obtain bilingual speech embeddings that perform reasonably well on a multilingual retrieval task. The recall and rank results are far above the

[12] Available at `https://github.com/dharwath/DAVEnet-pytorch`

[13] Modified code available at `https://github.com/getalp/BibleNet`

Query	R@1	R@5	R@10	$\widetilde{r}$
EN-EU	0.173	0.395	0.523	9
EN-ES	0.130	0.341	0.469	12
EN-HU	0.116	0.319	0.455	13
EN-RU	0.102	0.308	0.414	16
EN-RO	0.085	0.289	0.396	17
EN-FR	0.092	0.259	0.364	22
EN-FI	0.076	0.202	0.293	26

Table 4: Recall at top 1, 5, and 10 retrieval. Median rank $\widetilde{r}$ on a verse-to-verse retrieval task is also provided. Results are reported on the test set (816 verses). Chance recalls are 0.001 (R@1), 0.006 (R@5) and 0.012 (R@10). Chance median $\widetilde{r}$ is 408.5.

chance values. We also scored a simple baseline that uses utterance length to retrieve spoken verses (in other words, it uses only distance between spoken utterances' lengths to solve the retrieval task). With this baseline, medium ranks are better than chance level ($\widetilde{r} = 408$) but vary from $\widetilde{r} = 136$ (EN-FR) to $\widetilde{r} = 219$ (EN-FI), which is very poor compared to our baseline model. Interestingly, our best results, obtained for EN-EU ($\widetilde{r} = 9$) and EN-ES ($\widetilde{r} = 12$), illustrate that speech-to-speech retrieval task is feasible even for pairs of typologically different languages.

Following this experiment, we investigated the correlation between the median rank and two variables: the quality of the alignment (human evaluation) and the syntactic distance between the language pairs (using the *lang2vec* library (Littell et al., 2017)). Results are provided at Table 5. While there is no correlation between the rank and the syntactic distance, there is a strong negative correlation with respect to the human evaluation (significant for $p < 0.1$). One possible explanation for this result may be that higher quality alignments (measured by the human evaluation $\widetilde{x}$) lead to a slightly easier corpus for the speech-speech retrieval task (difficulty being measured by the rank $\widetilde{r}$). If confirmed, this result would suggest that speech-to-speech retrieval scores are a good proxy for rating alignment corpus quality, as performed for text by Schwenk et al. (2019) through the use of NMT.

Languages	$\widetilde{r}$	Quality ($\bar{\text{x}}$)	Syntactic dist.
EN - EU	9	*NA*	0.61
EN - ES	12	4.56	0.40
EN - HU	13	4.44	0.57
EN - RU	16	4.56	0.49
EN - RO	17	4.51	0.53
EN - FR	22	4.38	0.46
EN - FI	26	4.37	0.53
Correlation		-0.76	-0.21

Table 5: Correlation between median rank and 1) alignment quality (from manual evaluation) 2) syntactic distance between languages (measured with *lang2vec*).

6. Conclusion

In this paper, we presented the creation of an automatically generated clean and controlled parallel corpus based on the *CMU Wilderness Multilingual Speech Dataset*. Our resource, called MaSS, contains 20 hours of speech in eight languages (Basque, English, Finnish, French, Hungarian, Romanian, Russian and Spanish) and presents both speech-to-text and speech-to-speech alignments. The quality of the corpus was verified on a subset of 100 sentences in 8 language pairs by native speakers. The pipeline used for creating this dataset, as well as the computed forced alignments for each of the chosen languages, are openly accessible.[3] Only eight languages are currently covered, but we believe the same methodology could easily be applied for extending it to new languages.

7. Bibliographical References

Adams, O., Wiesner, M., Watanabe, S., and Yarowsky, D. (2019). Massively multilingual adversarial speech recognition. In *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 96–108, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Aharoni, R., Johnson, M., and Firat, O. (2019). Massively multilingual neural machine translation. In *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 3874–3884, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. In *ICLR 2015*, pages 3104–3112, San Diego, California, USA.

Bansal, S., Kamper, H., Lopez, A., and Goldwater, S. (2017). Towards speech-to-text translation without speech recognition. In Mirella Lapata, et al., editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*, pages 474–479. Association for Computational Linguistics.

Bansal, S., Kamper, H., Livescu, K., Lopez, A., and Goldwater, S. (2019). Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. In Jill Burstein, et al., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 58–68. Association for Computational Linguistics.

Bérard, A., Pietquin, O., Servan, C., and Besacier, L. (2016). Listen and translate: A proof of concept for end-to-end speech-to-text translation. In *NIPS Workshop on End-to-end Learning for Speech and Audio Processing*, Barcelona, Spain, December.

Bérard, A., Besacier, L., Kocabiyikoğlu, A. C., and Pietquin, O. (2018). End-to-end automatic speech translation of audiobooks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing,*

ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018, pages 6224–6228, April.

Black, A. W. (2019). Cmu wilderness multilingual speech dataset. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5971–5975, May.

Christodoulopoulos, C. and Steedman, M. (2015). A massively parallel corpus: the bible in 100 languages. *Language Resources and Evaluation*, 49(2):375–395.

Chrupała, G., Gelderloos, L., and Alishahi, A. (2017). Representations of language in a model of visually grounded speech signal. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 613–622. Association for Computational Linguistics.

Chung, Y.-A., Weng, W.-H., Tong, S., and Glass, J. R. (2018). Towards unsupervised speech-to-text translation. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7170–7174.

Di Gangi, M. A., Cattoni, R., Bentivogli, L., Negri, M., and Turchi, M. (2019). Must-c: a multilingual speech translation corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, Minneapolis, MN, USA, June.

Federmann, C. and Lewis, W. (2016). Microsoft speech language translation (mslt) corpus: The iwslt 2016 release for english, french and german. In *Proceedings of IWSLT 2016*, December.

Harwath, D., Chuang, G., and Glass, J. R. (2018a). Vision as an interlingua: Learning multilingual semantic embeddings of untranscribed speech. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*, pages 4969–4973.

Harwath, D., Recasens, A., Surís, D., Chuang, G., Torralba, A., and Glass, J. R. (2018b). Jointly discovering visual objects and spoken words from raw sensory input. In Vittorio Ferrari, et al., editors, *ECCV (6)*, volume 11210 of *Lecture Notes in Computer Science*, pages 659–677. Springer.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November.

Iranzo-Sánchez, J., Silvestre-Cerdà, J. A., Jorge, J., Roselló, N., Giménez, A., Sanchis, A., Civera, J., and Juan, A. (2020). Europarl-ST: A Multilingual Corpus For Speech Translation Of Parliamentary Debates. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Jia, Y., Johnson, M., Macherey, W., Weiss, R., Cao, Y., Chiu, C.-C., Ari, N., Laurenzo, S., and Wu, Y. (2019a). Leveraging weakly supervised data to improve end-to-end speech-to-text translation. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7180–7184, 05.

Jia, Y., Weiss, R. J., Biadsy, F., Macherey, W., Johnson, M., Chen, Z., and Wu, Y. (2019b). Direct Speech-to-Speech Translation with a Sequence-to-Sequence Model. In *Proc. Interspeech 2019*, pages 1123–1127.

Kisler, T., Reichel, U., and Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language*, 45:326 – 347.

Kocabiyikoğlu, A. C., Besacier, L., and Kraif, O. (2018). Augmenting librispeech with French translations: A multimodal corpus for direct speech translation evaluation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Lee, L.-s., Glass, J., Lee, H.-y., and Chan, C.-a. (2015). Spoken content retrieval beyond cascading speech recognition with text retrieval. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(9):1389–1420.

Littell, P., Mortensen, D. R., Lin, K., Kairis, K., Turner, C., and Levin, L. (2017). Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, volume 2, pages 8–14.

Navigli, R. and Ponzetto, S. P. (2010). BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden, July. Association for Computational Linguistics.

Nivre, J., de Marneffe, M., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R. T., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016.*

Post, M., Kumar, G., Lopez, A., Karakos, D., Callison-Burch, C., and Khudanpur, S. (2013). Improved speech-to-text translation with the fisher and callhome spanish-english speech translation corpus. In *International Workshop on Spoken Language Translation (IWSLT 2013)*.

Sanabria, R., Caglayan, O., Palaskar, S., Elliott, D., Barrault, L., Specia, L., and Metze, F. (2018). How2: a large-scale dataset for multimodal language understanding. In *Proceedings of the Workshop on Visually Grounded Interaction and Language (ViGIL)*. NeurIPS.

Schultz, T. and Schlippe, T. (2014). Globalphone: Pronunciation dictionaries in 20 languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014.*, pages 337–341.

Schwenk, H., Chaudhary, V., Sun, S., Gong, H., and Guzmán, F. (2019). Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from Wikipedia. *Preprint*, abs/1907.05791.

Sérasset, G. (2015). Dbnary: Wiktionary as a lemon-based multilingual lexical resource in rdf. *Semantic Web*,

6(4):355–361.

Sperber, M., Neubig, G., Niehues, J., and Waibel, A. (2019). Attention-passing models for robust and data-efficient end-to-end speech translation. *Transactions of the Association for Computational Linguistics*, 7:313–325.

Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q. V., Agiomyrgiannakis, Y., Clark, R., and Saurous, R. A. (2017). Tacotron: Towards end-to-end speech synthesis. In Francisco Lacerda, editor, *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, pages 4006–4010. ISCA.

Weiss, R. J., Chorowski, J., Jaitly, N., Wu, Y., and Chen, Z. (2017). Sequence-to-sequence models can directly translate foreign speech. In Francisco Lacerda, editor, *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, pages 2625–2629. ISCA.