Evaluating Multilingual Speech Translation Under Realistic Conditions with Resegmentation and Terminology

Elizabeth Salesky[J] **Kareem Darwish**[A] **Mohamed Al-Badrashiny**[A]
Mona Diab[M] **Jan Niehues**[K]
[J]Johns Hopkins University [A]aiXplain [M]Meta AI [K]Karlsruhe Institute of Technology
esalesky@jhu.edu

Abstract

We present the ACL 60/60 evaluation sets for multilingual translation of ACL 2022 technical presentations into 10 target languages. This dataset enables further research into multilingual speech translation under realistic recording conditions with unsegmented audio and domain-specific terminology, applying NLP tools to text and speech in the technical domain, and evaluating and improving model robustness to diverse speaker demographics.

1 Introduction

The NLP and speech communities are rapidly expanding, which has motivated increased interest in multilingual scientific communication and accessibility. From the automatic captioning at NAACL 2019 provided by Microsoft to the current ACL 60-60 initiative[1] for the 60th anniversary of ACL at 2022, it is clear that transcription and translation in the technical domain is needed, desired, and still a disproportionate challenge for current models compared to standard datasets in these spaces.

Translating technical presentations presents challenging conditions, from domain-specific terminology and adaptation, to recordings often captured with a laptop microphone and light background noise, diverse speaker demographics as well as unsegmented speech typically 10-60 minutes in duration. We have curated evaluation sets from presentations at ACL 2022 which have been professionally transcribed and translated with the support of ACL and the 60-60 initiative. In this paper we describe the methodology to create this dataset, considerations and methods to evaluate speech translation models with it, and open challenges we believe this dataset may support research towards. We release all data and intermediate steps to support further research in this space.

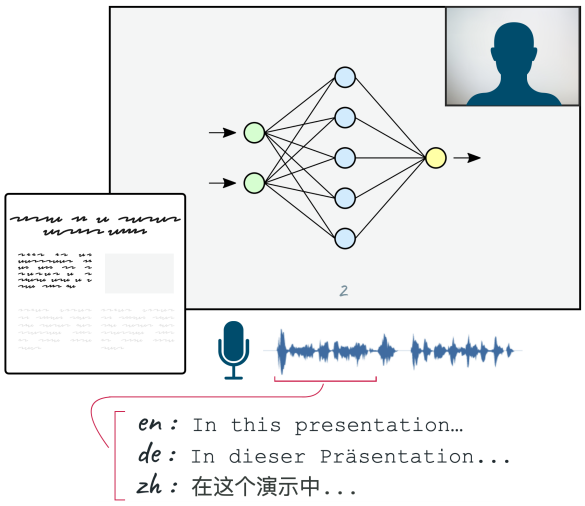


Figure 1: Multilingual translation of ACL presentations.

We present the ACL 60/60 evaluation sets to enable greater development of tools by the field for the field. Specifically, we hope that this data enables further research into speech translation and other NLP applications in the technical domain with resegmentation and terminology, given a diverse speaker set and realistic recording conditions, with the goal of increased accessibility and multilinguality. Our dataset is publicly available through the ACL Anthology.[2]

2 Evaluation under realistic conditions

To evaluate transcription and translation under realistic conditions may require different metrics than with e.g. provided segmentation. Here we present the necessary metrics in order to discuss the dataset creation process.

2.1 Resegmentation

While most offline speech translation models are trained with provided segmentation, in an application setting segmentation is unlikely to be provided.

[1]https://www.2022.aclweb.org/dispecialinitiative

[2]https://aclanthology.org/2023.iwslt-1.2

Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023), pages 62–78

Most models are typically unable to maintain output quality given audio of typical talk lengths (10+ minutes), necessitating the use of automatic segmentation methods. In order to evaluate output with variable segmentation, resegmentation to a fixed reference is necessary.

The standard tool within the field for many years has been mwerSegmenter (Matusov et al., 2005), which resegments model output to match a reference segmentation for downstream evaluation with various metrics. This is done by dynamically resegmenting the output using a given tokenization to minimize word error rate to the reference.[3] We use mwerSegmenter for all scores in this paper and suggest that resegmentation be the scoring standard for the ACL 60/60 dataset.

2.2 Evaluation metrics

We compare a variety of evaluation metrics to analyze both transcription and translation quality using the evaluation sets, as well as the results of intermediate steps in corpus creation such as post-editing.

For translation, we compare chrF (Popović, 2015) which is tokenization-agnostic and more appropriate for a wider array of target languages than BLEU; BLEU (Papineni et al., 2002) as computed by SACREBLEU (Post, 2018); and the model-based metric COMET (Rei et al., 2020), which often has higher correlation with human judgements (Mathur et al., 2020) though is limited by language coverage in pretrained models. For BLEU we use the suggested language-specific tokenizers in SACREBLEU for our non-space delimited target languages, Japanese (MeCab[4]) and Chinese (character-level).

To analyze both automatic and post-editing transcription quality, we use word error rate (WER). We note that we use case-sensitive and punctuation-sensitive WER here as these are both maintained in system output during dataset creation in order to be post-edited and translated. For downstream evaluation of ASR model quality using the final dataset, it may be desired to compute WER without case and without punctuation; if so, the scores would not be directly comparable to those presented here. We also use translation error rate (TER) (Snover et al., 2006) to assess the expected level of editing necessary to match the final reference quality.[5]

We caution against using any one translation metric in isolation, and suggest chrF and COMET as the standard evaluation metrics for this dataset.

3 Creating the ACL 60/60 evaluation sets

3.1 Languages

All data is originally spoken in English and then transcribed and translated to ten diverse languages from the 60/60 initiative for which publicly available speech translation corpora are available (see Table 5: §A.3): Arabic, Mandarin Chinese, Dutch, French, German, Japanese, Farsi, Portuguese, Russian, and Turkish. The resulting dataset contains three-way parallel (*speech, transcripts, translations*) one-to-many data for ten language pairs, and multi-way parallel text data for 100 language pairs.

3.2 Data selection

Data was selected from the ACL 2022 paper presentations for which precorded audio or video presentations were provided to the ACL Anthology. Talks were selected such that each of the two evaluation sets, development and evaluation, would have approximately one hour total duration. Oral presentations were advised to be up to 12 minutes per recording, resulting in 5 talks for each set with relatively balanced durations of ∼11.5 minutes each.

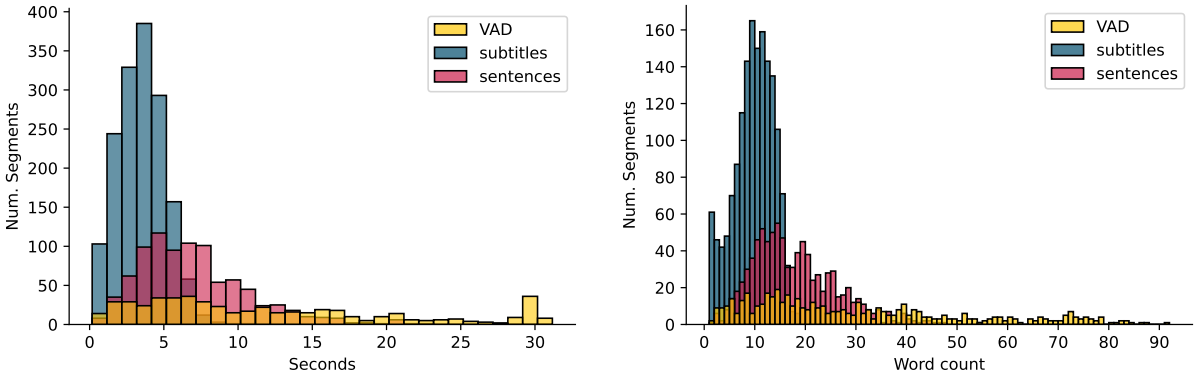
From the 324 available recordings, the final 10 were selected in order to balance speaker demographics, accents, and talk content, while lightly controlling for recording conditions. The majority of recordings were created using laptop microphones in quiet conditions, but background noise, microphone feedback, speech rate and/or volume in some cases affected understanding of the content. We selected talks with representative but minimal noise where conditions did not affect understanding of the content. We aimed for a gender balance representative of conference participation,[6] resulting in a 3:7 female:male speaker ratio. This is also a global field with a wide variety of native and non-native English accents, which remains a necessary challenge for speech models to address to mitigate performance biases (Sanabria et al., 2023; Feng et al., 2021; Koenecke et al., 2020; Tatman and Kasten, 2017). Talks were chosen and assigned to each set to maximize accent diversity, aiming for L1s from all continents with language families fre-

[3] We use word-level tokenization for all languages except Japanese and Chinese here, where we use character-level.

[4] https://taku910.github.io/mecab/

[5] We calculate TER with `--ter-normalized` and `--ter-asian-support` in SACREBLEU.

[6] Aggregate conference participation statistics provided by ACL 2022; see §A.2.



(a) Speech segment length distribution

(b) Text segment length distribution

Figure 2: Distribution of English segment lengths via speech duration (seconds) and text length (word count) for each of three segmentations: VAD, subtitles, and sentences.

quently represented in the ACL community while balancing topic diversity and gender. We note native language and country where available. Talks were chosen to cover a diverse set of tracks and topics and therefore diverse technical vocabulary representative of the needs of the field. Where presentations were chosen within the same track, they covered different focuses and methodology, e.g. math word problems versus release note generation or few-shot adaptation for structured data. Metadata for all talks with exact durations and track and speaker annotations are shown in Table 3 in §A.1.

Holding out speakers and topics per set optimizes for overall system generalization but reduces the match between dev and eval sets; this e.g. reduces the benefit of finetuning on the dev set to maximize test set performance and overfitting the model or chosen hyperparameters to the dev set will adversely affect test set performance. However, high performance on both sets is more likely to indicate generalizable systems and representative performance beyond these data points than if the dev and eval data were more closely matched.

3.3 Automatic transcription

The first pass through the data used automatic segmentation and transcription to provide initial transcripts. We used the Azure API speech-to-text service,[7] which has the best cost and quality balance of currently available models. In addition to transcription, the service performs speaker diarization, with implicit voice activity detection (VAD), segmenting the initially ∼11.5 minute audio files into segments of approximately 30 seconds or less based on pauses, speech, and non-speech phenomena. Figure 2 shows the resulting distribution of segment lengths. Evaluating these initial automatic transcripts against the final released version with resegmentation (§2.1), the automatic transcription yielded a WER of 15.4 and 22.4 for the development and evaluation sets, respectively.

[7] https://azure.microsoft.com/en-us/products/cognitive-services/speech-to-text

3.4 Human post-editing: Transcription

We contracted with aiXplain Inc. to professionally post-edit the ASR output. There was a three tier review process: an initial annotator post-edited per segment, followed by a quality assurance (QA) annotator who went through each full talk to ensure quality and consistency, and then finally 10-20% of the segments were randomly chosen for a final check. In addition to semantic content, annotators may theoretically also fix segmentation boundaries but in practice this rarely occurs. The annotators provided additional information about the speakers, namely gender (male, female) and age (child, young adult, adult, elderly). The annotators were also shown the video of the presentation to aid them ing recognizing technical terms, which may appear in the slides. Disfluencies were standardized such that false starts and repetitions were kept where there were perceivable pauses between them, and two hesitation spelling variations (*ah*, *um*) were used. The annotator guidelines and LabelStudio interface are shown in §A.4. After the professional post-editing pass, a domain expert verified and corrected the technical terms.

Post-editing analysis. ASR output is strongly monotonic with respect to the original speech, and accordingly most post-edits are for incorrectly tran-

REF:	we	find	a	BILSTM	**	CRF	model	using	flare
HYP:	we	find	a	BIAS	TM	CRF	model	using	flare
				S	D				

REF:	also	FASTTEXT	CHARACTER	EMBEDDINGS	
HYP:	also	FASTTEX	KITCHEN	BEDDINGS	
		S	S	S	

REF:	multilingual	BERT	PERFORMS	better	than	BETO	
HYP:	multilingual	BIRD	PERFORM	better	than	BETTER	
		S	S			S	

Figure 3: Sample ASR errors from dev using SCLITE. Corrections are emphasized with CASE.

scribed words, case, and punctuation. 93% of words were correctly transcribed by the initial ASR pass. Spurious punctuation and casing in the ASR output (ex 'Thank. You.') accounted for 43% of the errors captured by WER. Setting punctuation and case aside, in the professional post-editing pass, 60% of sentences had at least one correction made. The majority of post-edits were word-level substitutions for incorrectly transcribed words (62%). Dropped words were not common, with only 1.6% of words dropped by the ASR model and later inserted. Slightly more common (1.8%) were insertions due to words incorrectly transcribed as multiple tokens by the ASR system, and later corrected. Examples are shown in Figure 3.

Further corrections by a domain expert were made for 3% of words. While the majority were corrections to terminology requiring technical context (*'CONEL'* → *'CONLL'* or *'position or'* → *'positional'*), some fixes were for subtle number and tense changes in the ASR transcription possibly influenced by recording conditions or pronunciation.

Technical terms. The subset of technical terms appearing in the terminology lists created by the 60-60 initiative were automatically tagged on the source side (see Figure 4). These lists were not exhaustive but provide an initial keyword set to bootstrap identification and translation of technical terms and their evaluation, and which future work may find beneficial.

Technical terms comprised the majority of ASR errors. 86% of the tagged terminology were correctly transcribed the ASR model, 8% were corrected by the professional post-editors, and the remaining 6% were corrected by a domain expert.

3.5 Sentence segmentation

While it is common in speech corpora to segment based on voice activity detection or subtitle-like criteria, this may result in segments which are not parallel across languages (in the case of multilingual speech), which are too short to translate without additional context, or which are too long for effective system evaluation. For a multilingual dataset intended to be multi-way parallel and to be used for translation, it is critical to have consistent segmentation across all languages and for all segments to contain the necessary context to translate to the desired target languages.

> And in fact, [automatically] [detecting] [lexical] borrowings ah has proven to be useful [for] [NLP] [downstream] [tasks] such as [parsing], [text]-to-[speech] synthesis or [machine translation].

Figure 4: Example of tagged terminology from dev. Terminology lists were not exhaustive; [text-to-speech] did not appear, leading [text] and [speech] to be tagged separately.

The VAD segments facilitated transcription, but resulted in a wide distribution of segment lengths, some just one to two words long, and others containing multiple sentences, potentially skewing downstream evaluation metrics and providing a mismatch to common training conditions. One option would be to subdivide the segments using subtitle guidelines,[8] where those segments which do not conform to particular length guidelines are realigned into smaller segments which is done using forced alignment. However, subtitle segments often contain partial sentences, which, particularly when including languages with different word orders or degrees of reordering from the source language (English), may place verbs across segment boundaries for some languages and not others. Sentences, then, may be a more appropriate unit for multi-way parallel segments. We resegmented the final post-edited English transcriptions into sentences manually to avoid noise from currently available tools. Examples of all three segmentations (VAD, subtitles, and sentences) are shown in Figure 12 in § A.8. To ensure the speech and text were correctly aligned given the final sentence segments, they were re-force aligned using WHISPER-TIMESTAMPED (Louradour, 2023), an extension of OpenAI's Whisper model (Radford et al., 2022) which uses DTW (Giorgino, 2009) to time align at the word level, and were manually rechecked by the annotators.

[8] Subtitle guidelines are shown in § A.7.

	Metric	*ar*	*de*	*fa*	*fr*	*ja*	*nl*	*pt*	*ru*	*tr*	*zh*
dev	chrF	75.3	72.8	54.9	80.0	56.9	82.7	82.3	59.3	69.0	60.5
	BLEU	54.1	48.3	25.3	63.0	50.7	63.6	65.9	30.5	39.1	65.9
	COMET	86.2	83.6	76.8	84.5	89.1	88.1	87.9	82.5	85.9	87.4
eval	chrF	77.2	71.7	56.3	83.7	53.6	86.6	84.8	65.3	77.0	62.7
	BLEU	55.4	48.5	27.1	68.3	47.3	71.5	68.7	39.4	51.6	67.9
	COMET	86.2	83.6	79.5	84.5	89.1	88.1	87.9	82.5	85.9	87.4

Table 1: Evaluating the initial commercial MT from ground-truth transcripts against the final released references. BLEU scores in grey are calculated using language-specific tokenization (*ja*) or at the character-level (*zh*); see §2.2.

We compare the distribution of segment lengths for each of the three approaches (VAD, subtitles, and sentences) in terms of both duration (seconds) and number of words (English) in Figure 2. VAD results in the most uneven distribution, with segments ranging from <1 second to >30 seconds. Subtitles result in more uniform but distinctly shorter segments, with 58% containing less than 10 words and 19% shorter than two seconds, likely too short for some downstream tasks or metrics. Sentences result in less extreme segment lengths. Examples of each segmentation are shown in §A.8. The final data contains 468 sentences in the development set and 416 sentences in the evaluation set.

3.6 Machine translation

The first translation pass used publicly available bilingual MT models to translate the final sentence segments. We used the ModernMT API[9] for the 9 of 10 language pairs supported, and the Azure API[10] for English-Farsi. We evaluate the commercial machine translation output against the final released translation references (§3.7) using the metrics discussed in §2.2, shown in Table 1.

Each metric suggests a different story about translation quality and the degree to which it is language-specific. While COMET suggests relatively consistent performance across languages, chrF and BLEU do not. chrF and BLEU suggest significantly worse performance for a subset of target languages, including all but one of the non-Latin script and non-Indo European languages. BLEU yields 1.7× greater variance than chrF. By all metrics, though, MT quality was consistent between the development and evaluation sets. We see in the next section that the amount of post-editing required to create the final references, however, is not necessarily indicated by these metrics.

[9] https://www.modernmt.com/api/

[10] https://azure.microsoft.com/en-us/products/cognitive-services/translator

3.7 Human post-editing: Translation

Post-editing has become the industry standard due its increased productivity, typically reducing processing time and cognitive load compared to direct translation, particularly for domain-specific texts (O'Brien, 2007; Groves and Schmidtke, 2009; Tatsumi, 2009; Plitt and Masselot, 2010).

We contracted with Translated to professionally post-edit the MT output. There was a two tier review process: an initial annotator who was a native speaker of the target language post-edited per segment, followed by a second to review the output and consistency of the first. Annotator guidelines and the post-editing interface are shown in §A.5.

Technical terms. Terminology was not handled separately during the MT step nor automatically tagged, given that the MT systems may omit or incorrectly translate technical terms. We did not use constrained decoding given the terminology lists translations as their validity could be context-dependent and some terms had multiple possible translations. Instead, translation post-editors were instructed to correct the translations of tagged terminology on the source if they were not maintained and then tag the appropriate target translations for each source tagged source span. Capitalized acronyms and terminology not on the lists and unknown to the translators was left in English.

Post-editing analysis. While the metrics in the previous section give a sense for the automatic translation quality, they do not necessarily reflect the effort required to post-edit the translations to final reference quality. Using TER to assess the degree of post-editing necessary, we see in Figure 5 that this varies by language. Most noticeably, we see that Farsi, Russian, Japanese as target languages required the highest amount of post-editing.

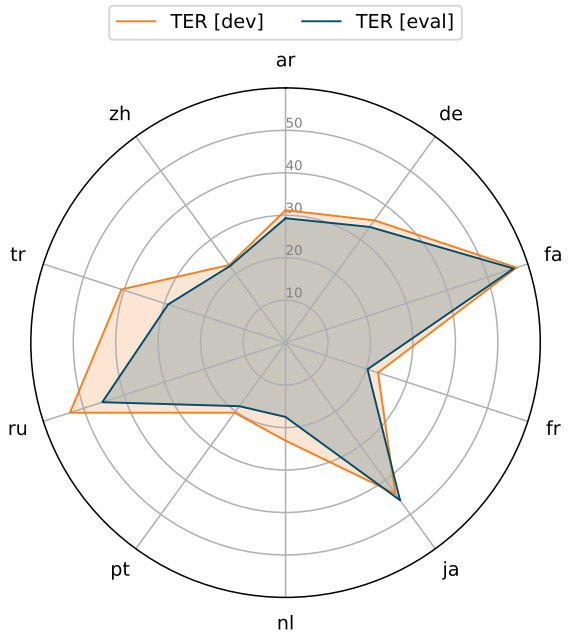


Figure 5: Estimated translation post-editing effort required per target language, as measured by TER.

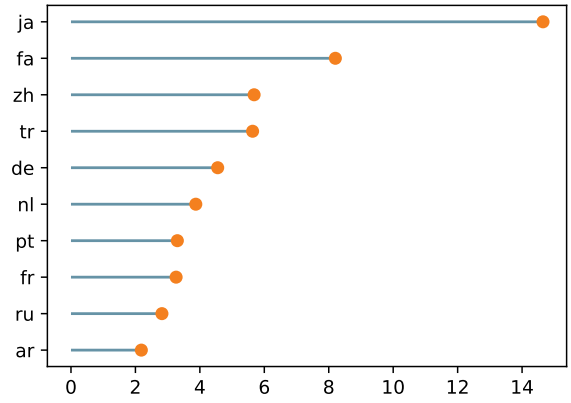


Figure 6: Degree of reordering done in MT post-editing.

For Farsi and Japanese, we see that this is predominantly due to reordering. Isolating reordering from semantic corrections by looking only at those tokens[11] which did not need to be corrected, we use Levenshtein distance to assess the degree of reordering from the MT output required. We observed a strong bias towards source language word order in the machine translation output, causing a greater degree of post-editing for languages with differing word orders. Figure 6 shows that reordering requirements are moderately correlated with overall post-editing effort for most languages ($\rho = 0.41$), while TER is only weakly suggested by COMET ($\rho = 0.29$) and is negatively correlated with chrF and BLEU (-0.63, -0.21 respectively).

For most target languages, there was no significant difference in post-editing effort between dev and test, but where there was a difference it was the dev talks that required additional editing, most noticeably for Turkish and Russian and to a lesser degree Dutch. Dividing the data into individual talks, which each vary in content within the technical domain, there was some variation in the quality of the first-pass MT (Figure 7). We found that which talks require similar levels of post-editing is moderately to strongly correlated across languages, suggesting this was due to topic rather than language, with the exception of Farsi and Japanese (Figure 8). This correlation does not appear to be influenced by language family and was not related to the proportion of tagged terminology per talk. For Russian and Turkish, a particular talk skewed overall dev TER, possibly due to a greater proportion of polysemous terms with domain-specific meaning in that area.

Terminology. Tagged terminology was more often correctly automatically transcribed than translated. Between 70-75% of the tagged spans were translated correctly by the initial MT model depending on the target language, as measured by an exact match with the final tagged post-edited span. The remaining 25-30% were manually corrected by the post-editors. In addition, 2-5% of words overall were left in English, predominantly made up of additional terminology and names.

4 Challenges to Address with ACL 60/60

4.1 Segmentation

Speech translation datasets customarily provide a segmentation for translation and evaluation, segmented either manually (e.g. CoVoST) or automatically (e.g. MuST-C). In realistic use cases, such segmentation is unavailable and long audio cannot be processed directly, resulting in mismatched conditions at inference time. There can be a noticeable performance gap between manual segmentation and automatic methods (Tsiamas et al., 2022).

We illustrate the impact of different speech segmentations on downstream transcription and translation quality by comparing manual sentence segmentation to the initial VAD segments as well as to SHAS (Tsiamas et al., 2022), using the top line commercial ASR and MT systems used during the

[11] Characters rather than words were used for this analysis for Japanese and Chinese.

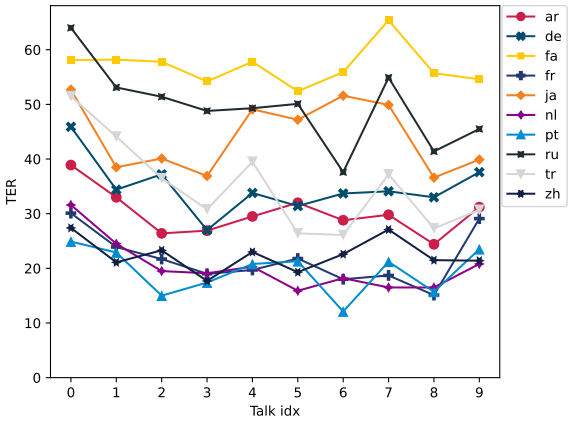


Figure 7: Range in TER by talk per language.

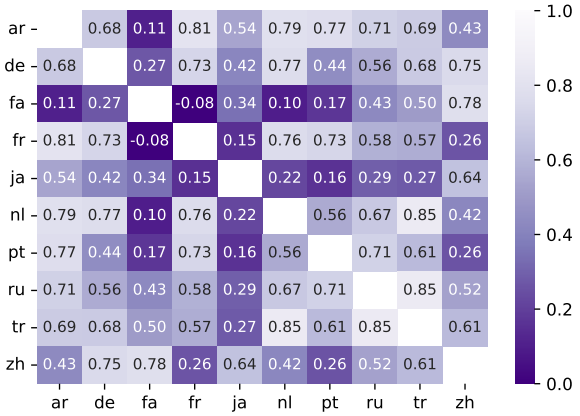


Figure 8: Correlation in TER across languages.

dataset creation pipeline. As seen in Table 2,[12] under certain circumstances automatic segmentation methods can perform as well as manual sentence segmentation, though this is not always the case and small resulting differences in ASR performance may cascade into larger performance gaps in downstream MT, meriting further research.

Variation due to segmentation also depends on model training conditions. Models are typically optimized for the segment lengths observed in training and/or may use additional internal segmentation. For example, when we compare the Whisper $_{\text{LARGE}}$ model (Radford et al., 2022) which is trained on longer segments, sentences are suboptimal compared to SHAS and VAD (0.1-0.9 WER), and when they are further segmented up to $4\times$ by its internal VAD this cascades to disproportionately worse downstream MT performance (by up to 8 chrF) than with the Azure ASR.

	ASR		MT	
Segmentation	*dev*	*test*	*dev*	*test*
Manual sentences	15.2	21.4	69.4	71.5
Commercial VAD	15.4	22.4	62.0	59.6
SHAS	16.4	21.5	61.9	60.4

Table 2: Comparison between manual sentence segmentation and high quality automatic segmentation for ASR and cascaded ST in WER and avg. chrF, respectively.

Segmentation is an important open challenge, and we suggest that this dataset be used to evaluate segmentation by making the dataset standard scoring with resegmentation.

[12]chrF for individual languages is shown in Table 6.

4.2 Demographic fairness

The field is diverse and rapidly growing with a wide variety of speaker demographics and native and non-native English accents. As we train increasingly large and multilingual models it is important to evaluate their fairness to ensure any biases we may find decrease rather than increase over time, which we believe this dataset may help with.

The variety of speaker demographics in both the field and these evaluation sets remain disproportionately challenging to current ASR models. Looking at the average WER among talks of each gender, we see a margin of 10.5. 15% of dev sentences and 26% of eval sentences were misclassified as non-English languages when using the multilingual Whisper $_{\text{BASE}}$ model, showing a bias against varied pronunciations and L1s that it is necessary to address when pursuing multilingual modelling. WER is 23% better when the model is prompted to generate English only, however, there is still a further 16% gap to the English-only BASE model. Moving to the larger multilingual model, the discrepancy in performance with and without language prompting becomes $2.4\times$ larger, though overall performance improves. At worst, the ΔWER between speakers is 62.2, and at best, 8.0, highlighting a significant discrepancy which needs to be improved.

Demographic fairness is an important issue which requires targeted research to address. We hope these evaluations sets may facilitate further research in this space, despite their small size.

4.3 Domain adaptation and terminology

Terminology. Constrained decoding of technical terms or domain-specific translations is an area

of active research (Hu et al., 2019; Post and Vilar, 2018; Hokamp and Liu, 2017). The terminology lists were not exhaustive, containing just over 250 terms, but provide an initial keyword set to bootstrap identification and translation of technical terms in context and their evaluation, which future work may find beneficial.

We highlight the reduction in terminology recall between the strong ASR and MT systems used in the dataset creation pipeline below in Figure 9. It is clear that even commercial systems struggle with domain-specific terminology particularly without adaptation. While there are discrepancies across language pairs, terminology recall is strongly correlated with overall translation performance ($\rho = 0.8$) as measured by chrF.

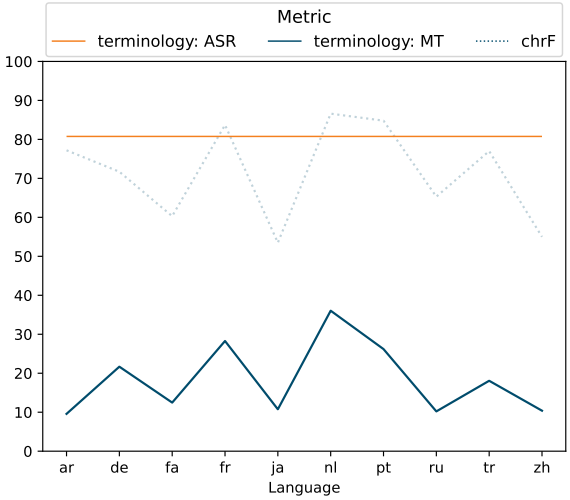


Figure 9: Terminology recall of ASR vs MT, with overall translation performance shown behind (chrF).

Lightweight domain adaptation. There are few publicly available datasets with technical content, and fewer translated. While it is possible to scrape in-domain material e.g. from the ACL Anthology, this would be in the source language (English) only rather than the target languages. While only having target-domain data in the source language is a realistic scenario, it is not the setting typically found in current research or approaches, and highlights the need for new methods for domain adaptation which can make use of this data. We additionally provide paper titles and abstracts, which are likely to contain both particularly important vocabulary and cue the talk topic. We hope this data may prove beneficial for lightweight methods to adapt to the technical domain or specific talk settings or to lexically constrain or prompt particular translations.

5 Related work

Previous work has studied data from the ACL Anthology for term mining and identification (Schumann and Martínez Alonso, 2018; Jin et al., 2013) and concept relation (Gábor et al., 2016) in the scientific domain.

Few speech translation datasets in the technical domain exist but those that do such as the QCRI Educational Corpus (Abdelali et al., 2014; Guzman et al., 2013) have primarily targeted educational lectures and videos. Additional datasets specifically for speech translation evaluation (Conneau et al., 2023) are primarily 'general domain.'

Significant previous work has studied various aspects of translation post-editing, including post-editing effort (Scarton et al., 2019), evaluating post-editing quality and reference bias (Bentivogli et al., 2018), bias from the initial MT quality and output patterns (Zouhar et al., 2021; Picinini and Ueffing, 2017), and the the efficacy of post-editing in highly technical domains (Pinnis et al., 2016) and resulting translation biases (Čulo and Nitzke, 2016).

The impact of automatic segmentation quality on various ST metrics has been evaluated in recent IWSLT shared tasks (Ansari et al., 2020; Anastasopoulos et al., 2021, 2022) and research (Tsiamas et al., 2022; Sen et al., 2022; Ansari et al., 2021) using other datasets (TED) with longer reference segmentations than ours. With longer sequences there is greater potential for variation, and past campaigns have observed larger differences between segmentations than seen here and even improvements over the provided segmentation. Significant additional work has been done in the simultaneous translation space, which we do not address here.

6 Conclusions

We introduced a new dataset to evaluate multilingual speech translation from English into ten target languages specifically in the technical NLP domain. We have discussed in detail the steps to create the corpus and the tools and considerations required. We have also provided a further view into evaluation methodology mimicking realistic conditions where segmentation is not provided. We hope that this dataset may be useful for the field to study the effectiveness of the tools we develop both for translation and additional applications in the technical domain in an increasingly multilingual space.

Limitations

While we have done our best to create high-quality evaluation data, there are limitations that should be kept in mind when using these datasets. It is known that creating translations by post-editing may bias data towards the output of the MT systems used for initial translations; however, many transcription and translation vendors now exclusively use post-editing rather than translation from scratch and so direct translation may not be an option in all cases. This could influence metrics toward similar MT systems. The presented evaluation sets are moderately sized compared to datasets in other domains with plentiful mined data, and may be best used in conjunction by reporting on both the development and evaluation sets for statistical significance. The evaluation sets also have a necessarily limited set of speakers which may not be fully representative. Systems which tune to the development set run the risk of over-fitting to specific speakers or content. We do not perform a comparison to human evaluation here, but refer interested readers to the IWSLT'23 evaluation campaign findings paper which runs this comparison for a variety of systems with the ACL 60/60 data (Agarwal et al., 2023).

Ethical Considerations

This dataset is constructed from a small set of speakers where each speaker may be the only representative of certain cross-sectional axes, and as such, even reporting aggregate metadata may break anonymity. While we do not distribute speaker annotations with the data some information is inherently recoverable due to the link to the Anthology. We nonetheless believe this data will be beneficial to the community in order to study language processing on technical data, and it is necessary to have a diverse evaluation set to provide a more realistic and representative measure for generalization. It is difficult and costly to construct datasets with human-edited transcripts and translations and this was the largest set possible to collect. Post-editors were compensated with professional wages.

Acknowledgements

We are very grateful to funding and support from ACL and the 60/60 initiative to create this dataset. We thank our annotators and the generous support of aiXplain and Translated. Elizabeth Salesky is supported by the Apple Scholars in AI/ML fellowship.

References

Ahmed Abdelali, Francisco Guzman, Hassan Sajjad, and Stephan Vogel. 2014. The AMARA corpus: Building parallel language resources for the educational domain. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1856–1862, Reykjavik, Iceland. European Language Resources Association (ELRA).

Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, William Chen, Khalid Choukri, Alexandra Chronopoulou, Anna Currey, Thierry Declerck, Qianqian Dong, Yannick Estéve, Kevin Duh, Marcello Federico, Souhir Gahbiche, Barry Haddow, Benjamin Hsu, Phu Mon Htut, Hirofumi Inaguma, Dávid Javorský, John Judge, Yasumasa Kano, Tom Ko, Rishu Kumar, Pengwei Li, Xutai Ma, Prashant Mathur, Evgeny Matusov, Paul McNamee, John P. McCrae, Kenton Murray, Maria Nadejde, Satoshi Nakamura, Matteo Negri, Ha Nguyen, Jan Niehues, Xing Niu, Atul Ojha Kr., John E. Ortega, Proyag Pal, Juan Pino, Lonneke van der Plas, Peter Polák, Elijah Rippeth, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Yun Tang, Brian Thompson, Kevin Tran, Marco Turchi, Alex Waibel, Mingxuan Wang, Shinji Watanabe, and Rodolfo Zevallos. 2023. Findings of the IWSLT 2023 Evaluation Campaign. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*. Association for Computational Linguistics.

Antonios Anastasopoulos, Loïc Barrault, Luisa Bentivogli, Marcely Zanon Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, Clara Emmanuel, Yannick Estève, Marcello Federico, Christian Federmann, Souhir Gahbiche, Hongyu Gong, Roman Grundkiewicz, Barry Haddow, Benjamin Hsu, Dávid Javorský, Věra Kloudová, Surafel Lakew, Xutai Ma, Prashant Mathur, Paul McNamee, Kenton Murray, Maria Nădejde, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, John Ortega, Juan Pino, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Yogesh Virkar, Alexander Waibel, Changhan Wang, and Shinji Watanabe. 2022. Findings of the IWSLT 2022 evaluation campaign. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 98–157, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Antonios Anastasopoulos, Ondřej Bojar, Jacob Bremerman, Roldano Cattoni, Maha Elbayad, Marcello Federico, Xutai Ma, Satoshi Nakamura, Matteo Negri, Jan Niehues, Juan Pino, Elizabeth Salesky, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Alexander Waibel, Changhan Wang, and Matthew Wiesner. 2021. FINDINGS OF THE IWSLT 2021 EVALUATION CAMPAIGN. In *Proceedings of the 18th*

International Conference on Spoken Language Translation (IWSLT 2021), pages 1–29, Bangkok, Thailand (online). Association for Computational Linguistics.

Ebrahim Ansari, Amittai Axelrod, Nguyen Bach, Ondřej Bojar, Roldano Cattoni, Fahim Dalvi, Nadir Durrani, Marcello Federico, Christian Federmann, Jiatao Gu, Fei Huang, Kevin Knight, Xutai Ma, Ajay Nagesh, Matteo Negri, Jan Niehues, Juan Pino, Elizabeth Salesky, Xing Shi, Sebastian Stüker, Marco Turchi, Alexander Waibel, and Changhan Wang. 2020. FINDINGS OF THE IWSLT 2020 EVALUATION CAMPAIGN. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 1–34, Online. Association for Computational Linguistics.

Ebrahim Ansari, Ondřej Bojar, Barry Haddow, and Mohammad Mahmoudi. 2021. SLTEV: Comprehensive evaluation of spoken language translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 71–79, Online. Association for Computational Linguistics.

Luisa Bentivogli, Mauro Cettolo, Marcello Federico, and Christian Federmann. 2018. Machine translation human evaluation: an investigation of evaluation based on post-editing and its relation with direct assessment. In *Proceedings of the 15th International Conference on Spoken Language Translation*, pages 62–69, Brussels. International Conference on Spoken Language Translation.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805.

Oliver Čulo and Jean Nitzke. 2016. Patterns of terminological variation in post-editing and of cognate use in machine translation in contrast to human translation. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pages 106–114.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 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 1 (Long and Short Papers)*, pages 2012–2017, Minneapolis, Minnesota. Association for Computational Linguistics.

Siyuan Feng, Olya Kudina, Bence Mark Halpern, and Odette Scharenborg. 2021. Quantifying bias in automatic speech recognition. *ArXiv*, abs/2103.15122.

Kata Gábor, Haïfa Zargayouna, Davide Buscaldi, Isabelle Tellier, and Thierry Charnois. 2016. Semantic annotation of the ACL Anthology corpus for the automatic analysis of scientific literature. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3694–3701, Portorož, Slovenia. European Language Resources Association (ELRA).

Toni Giorgino. 2009. Computing and visualizing dynamic time warping alignments in r: The dtw package. *Journal of Statistical Software*, 31(7).

Declan Groves and Dag Schmidtke. 2009. Identification and analysis of post-editing patterns for MT. In *Proceedings of Machine Translation Summit XII: Commercial MT User Program*, Ottawa, Canada.

Francisco Guzman, Hassan Sajjad, Stephan Vogel, and Ahmed Abdelali. 2013. The AMARA corpus: building resources for translating the web's educational content. In *Proceedings of the 10th International Workshop on Spoken Language Translation: Papers*, Heidelberg, Germany.

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1546, Vancouver, Canada. Association for Computational Linguistics.

J. Edward Hu, Huda Khayrallah, Ryan Culkin, Patrick Xia, Tongfei Chen, Matt Post, and Benjamin Van Durme. 2019. Improved lexically constrained decoding for translation and monolingual rewriting. 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 839–850, Minneapolis, Minnesota. Association for Computational Linguistics.

J. Iranzo-Sánchez, J. A. Silvestre-Cerdà, J. Jorge, N. Roselló, A. Giménez, A. Sanchis, J. Civera, and A. Juan. 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)*, pages 8229–8233.

Yiping Jin, Min-Yen Kan, Jun-Ping Ng, and Xiangnan He. 2013. Mining scientific terms and their definitions: A study of the ACL Anthology. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 780–790, Seattle, Washington, USA. Association for Computational Linguistics.

Allison Koenecke, Andrew Joo Hun Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences of the United States of America*, 117:7684 – 7689.

Jérôme Louradour. 2023. whisper-timestamped. https://github.com/linto-ai/whisper-timestamped.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020. Results of the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online. Association for Computational Linguistics.

Evgeny Matusov, Gregor Leusch, Oliver Bender, and Hermann Ney. 2005. Evaluating machine translation output with automatic sentence segmentation. In *Proceedings of the Second International Workshop on Spoken Language Translation*, Pittsburgh, Pennsylvania, USA.

Sharon O'Brien. 2007. An empirical investigation of temporal and technical post-editing effort. *The Information Society*, 2:83–136.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Silvio Picinini and Nicola Ueffing. 2017. A detailed investigation of bias errors in post-editing of MT output. In *Proceedings of Machine Translation Summit XVI: Commercial MT Users and Translators Track*, pages 79–90, Nagoya Japan.

Marcis Pinnis, Rihards Kalnins, Raivis Skadins, and Inguna Skadina. 2016. What can we really learn from post-editing? In *Conferences of the Association for Machine Translation in the Americas: MT Users' Track*, pages 86–91, Austin, TX, USA. The Association for Machine Translation in the Americas.

Mirko Plitt and François Masselot. 2010. A productivity test of statistical machine translation post-editing in a typical localisation context. In *Prague Bulletin of Mathematical Linguistics*.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1314–1324, New Orleans, Louisiana. Association for Computational Linguistics.

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

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ramon Sanabria, Nikolay Bogoychev, Nina Markl, Andrea Carmantini, Ondrej Klejch, and Peter Bell. 2023. The edinburgh international accents of english corpus: Towards the democratization of english asr.

Scarton Scarton, Mikel L. Forcada, Miquel Esplà-Gomis, and Lucia Specia. 2019. Estimating post-editing effort: a study on human judgements, task-based and reference-based metrics of MT quality. In *Proceedings of the 16th International Conference on Spoken Language Translation*, Hong Kong. Association for Computational Linguistics.

Anne-Kathrin Schumann and Héctor Martínez Alonso. 2018. Automatic annotation of semantic term types in the complete ACL Anthology reference corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Sukanta Sen, Ondřej Bojar, and Barry Haddow. 2022. Simultaneous translation for unsegmented input: A sliding window approach.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Rachael Tatman and Conner Kasten. 2017. Effects of talker dialect, gender & race on accuracy of bing speech and youtube automatic captions. In *Interspeech*.

Midori Tatsumi. 2009. Correlation between automatic evaluation metric scores, post-editing speed, and some other factors. In *Proceedings of Machine Translation Summit XII: Posters*, Ottawa, Canada.

Ioannis Tsiamas, Gerard I. Gállego, José A. R. Fonollosa, and Marta Ruiz Costa-jussà. 2022. Shas: Approaching optimal segmentation for end-to-end speech translation. In *Interspeech*.

Changhan Wang, Juan Pino, Anne Wu, and Jiatao Gu. 2020. CoVoST: A diverse multilingual speech-to-text translation corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4197–4203, Marseille, France. European Language Resources Association.

Vilém Zouhar, Martin Popel, Ondřej Bojar, and Aleš Tamchyna. 2021. Neural machine translation quality and post-editing performance. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10204–10214, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Appendix

A.1 Additional Metadata for ACL 60/60 Evaluation Sets

Below we list the duration for talks in the evaluation sets, along with additional demographic metadata about the presenting author (speaker) and content (conference track). Conference tracks are taken from the ACL 2022 handbook. Gender annotations were checked with speakers' listed pronouns[13] and validated by speakers where available. For speaker demographics and accent we list L1 and native country where available, as well as country of affiliation as a rough proxy.

Gender	L1	Country	Affiliation	Time	Track
M	Kinyarwanda	Rwanda	USA	0:11:35	Theme: Language Diversity (Best Paper)
M	—	—	USA	0:11:35	Dialogue and Interactive Systems
F	Spanish	Spain	Spain	0:12:17	Resources and Evaluation
F	Marathi	India	USA	0:12:09	Question Answering
M	Polish	Poland	Poland	0:09:37	Machine Learning for NLP
				0:57:13	*Total development set duration*
M	Chinese	China	China	0:12:03	NLP Applications
M	—	Belgium	Netherlands	0:12:02	Resources and Evaluation
F	Romanian	Romania	Germany	0:09:22	Language Grounding, Speech and Multimodality
M	Japanese	Japan	Japan	0:14:02	NLP Applications
M	Hebrew	Israel	Israel	0:11:53	NLP Applications
				0:59:22	*Total evaluation set duration*

Table 3: Additional metadata for talks in the evaluation sets.

A.2 ACL 2022 Conference Participation Statistics

Aggregate statistics for self-identified gender as listed on conference registrations were provided by ACL.

Gender	#	%
Woman	909	28.7
Man	2164	68.3
Non-binary / Genderqueer / Third gender	14	<1
Genderfluid / Gender non-confirming	<10	<1
Prefer not to say	77	2.4
Specify your own	<10	<1
TOTAL	3170	100

Table 4: Aggregate statistics on gender of ACL 2022 conference participants.

[13]Though we note pronouns do not always indicate gender.

A.3 Publicly Available Corpora

Below are the current publicly available multi-way parallel speech translation corpora with English as the speech source. We note that for MuST-C not all target languages are available in all versions of the corpus as successive versions added additional language coverage. For full coverage v1.2 or above is required.

Corpus	Src	Tgt	
MuST-C (Di Gangi et al., 2019)	en	all (10)	ar, de, fa, fr, ja, nl, pt, ru, tr, zh
CoVoST (Wang et al., 2020)	en	all (10)	ar, de, fa, fr, ja, nl, pt, ru, tr, zh
Europarl-ST (Iranzo-Sánchez et al., 2020)	en	some (4)	de, fr, pt, tr

Table 5: Current publicly available aligned speech translation corpora covering the ACL 60/60 language pairs. Target languages are abbreviated using ISO 639-1 codes as follows – *Arabic: ar, German: de, Farsi: fa, French: fr, Japanese: ja, Dutch: nl, Portuguese: pt, Russian: ru, Turkish: tr, Mandarin Chinese: zh.*

A.4 Transcription Post-editing Guidelines and Interface

The following guidelines were used for transcription post-editing by aiXplain. The acceptance criterion was word accuracy >95%.

- Accuracy. Only type the words that are spoken in the audio file. Phrases or words you don't understand should NOT be omitted. Instead, they should be annotated using the label "#Unclear".
- Keep everything verbatim. Include every utterance and sound exactly as you hear. All filler words should be included (ex. #ah, #hmm). If the user corrects his/her self, all the utterances should be transcribed and corrected words need to preceded with a # mark (ex. She says #said that).
- Do not paraphrase. Do not correct the speaker's grammar nor rearrange words. Also, do not cut words that you think are off-topic or irrelevant. Any words not spoken should not be included. Type the actual words spoken. If the speaker makes a grammatical mistake, the transcript must reflect the mistake (ex. If the speaker says: "he were", it should be transcribed as is without correction).
- Repeat repeated words in the transcript. For example, if the user says: I I said, you must include both instances of I.
- Do not add additional information such as page numbers, job numbers, titles or your comments in your submission.
- Foreign words should be transliterated using Latin letters.
- All abbreviations need to be spelled out. For example, doctor should NOT be spelled as Dr. Similarly, percent should NOT be spelled as %.
- All numbers and special symbols (ex.: %, $, +, @, =, etc.), or combinations of both must be spelled out as words, and must match what the speaker says exactly.
- All proper names (ex. Google, NATO, Paris) should be transliterated in English.
- Proper punctuation needs to be placed in the text (ex. He, the boy, .). Please pay special attention and do not miss/omit these punctuation marks: , . ? ! :)(
- Personally identifiable information (like phone number, address, IDs) should be marked in the text as <PII></PII>. For example: My address is <PII>address</PII>
- Use double dashes "--" to indicate truncated words, attached whether at the beginning or the end of the word (ex. transfor–).

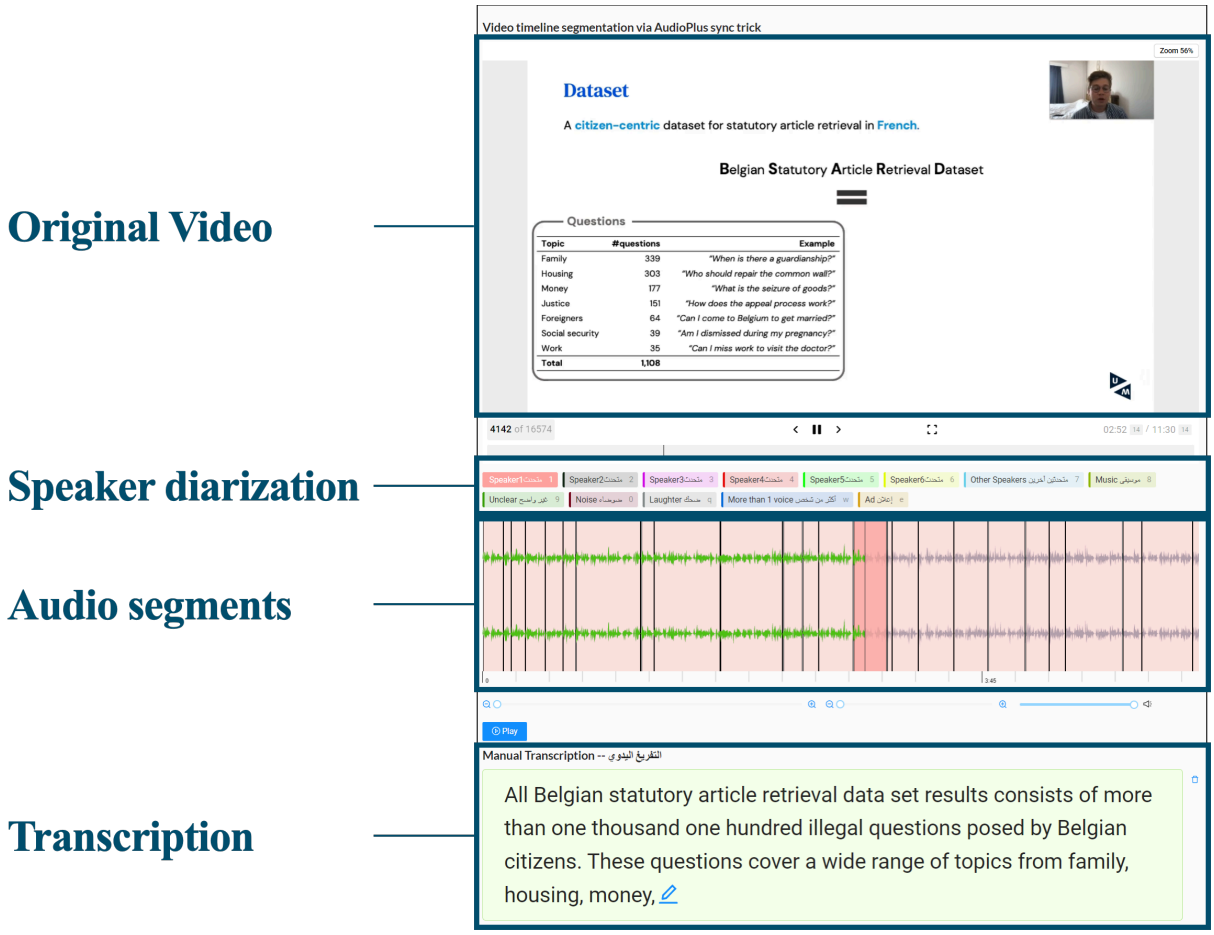


Figure 10: LabelStudio interface for transcription post-editing.

A.5 Translation Post-editing Instructions and Interface

The translation post-editing task was carried out in Matecat[14], an open-source CAT tool that allows annotators to collaborate and get suggestions from ModernMT in real-time. Matecat also offers an embedded glossary feature that ensures effective and consistent terminology management (as shown in the interface image in Figure 11 below, featuring Matecat glossary suggestions).

The following guidelines were used for translation post-editing:

- Any term found in the 60-60 terminologies list, should be translated using the translation in the terminologies list.
- Any abbreviation if not found in the terminologies list, should be kept it in the English form
- The terms in the terminologies list may contain one or more translation for each term separated by ':::'. The translator should pick the proper one based on the context
- If the translator thinks that none of the given translations for a specific term makes sense in the given context, the translators can use a better translation if they are very confident. If not very confident, keep the word in the English form

[14] https://site.matecat.com/

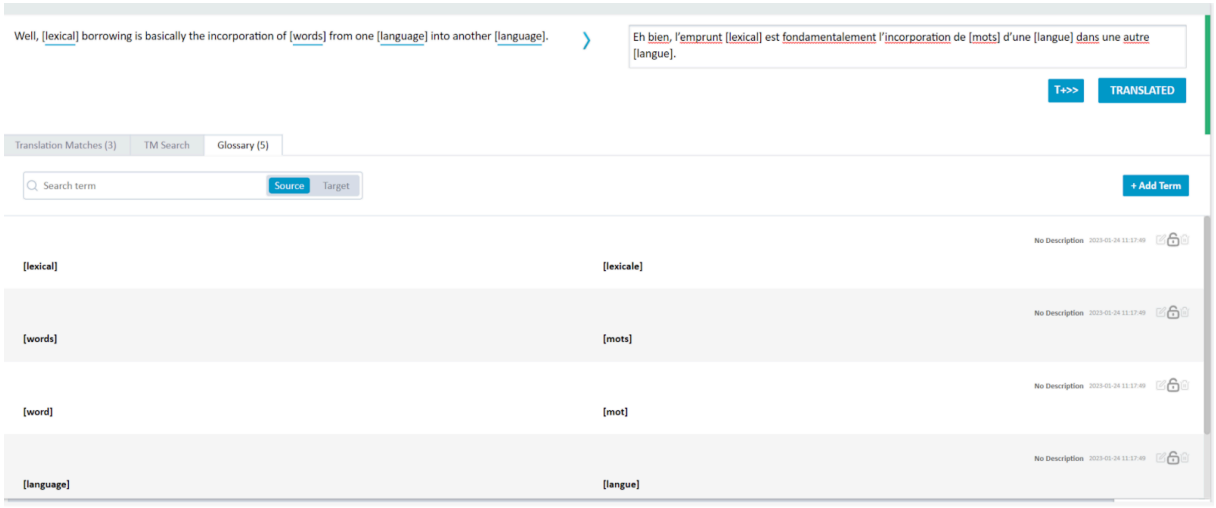


Figure 11: Matecat interface for translation post-editing.

A.6 Segmentation Comparison

Set	Segmentation	*ar*	*de*	*fa*	*fr*	*ja*	*nl*	*pt*	*ru*	*tr*	*zh*	Avg.
dev	Sentences	66.9	68.7	53.4	73.9	47.8	74.3	74.0	55.0	62.4	50.4	62.7
	Commercial VAD	66.6	68.5	52.7	74.1	46.2	73.6	73.7	53.9	60.6	49.8	62.0
	SHAS	66.5	68.6	52.8	73.7	46.9	73.8	73.5	54.3	59.9	49.7	62.0
eval	Sentences	64.0	66.1	51.3	69.0	43.9	71.0	71.9	55.8	63.8	46.0	60.3
	Commercial VAD	63.5	66.3	51.1	69.0	43.7	70.4	72.0	55.1	62.9	47.1	60.1
	SHAS	64.4	66.4	51.5	69.6	42.0	71.4	72.4	55.7	63.1	45.4	60.2

Table 6: Cascaded ST by language for different source speech segmentations, resegmented and scored with chrF.

A.7 Subtitle Guidelines

Subtitle guidelines following industry standards, see for example Netflix[15] and TED[16]:

- No one segment is allowed to be longer than 30 seconds.
- Each line can not be longer than 42 characters.
- A maximum of 2 lines of text can be shown on screen at once.
- The subtitle reading speed should kept to a maximum of ∼20 characters per second.[17]

If one of the segments created by the VAD does not adhere to the above guidelines, an English model is used to force alignment the long audio segment and its transcript to get the timestamp of each token, and then the segment is split into shorter subsegments. Note that these guidelines are automatically applied; the above means that if a VAD segment conforms to these guidelines it will not be resegmented, and subtitle segments may differ from manually created subtitles were semantic coherence may be prioritized over longer segments within these guidelines, or text may be lightly changed from what is spoken to optimize subtitle quality (here not allowed).

[15] https://partnerhelp.netflixstudios.com/hc/en-us/articles/217350977-English-Timed-Text-Style-Guide
[16] https://www.ted.com/participate/translate/subtitling-tips
[17] Varies by program audience, commonly between 17 and 21.

A.8 Segmentation Examples

Examples of each transcript segmentation approach discussed (VAD, subtitles, and sentences) for sample data from the development set. Examples were chosen to show segments from the longest and shortest VAD quartiles, and the resulting subtitles following subtitle guidelines from § A.7.

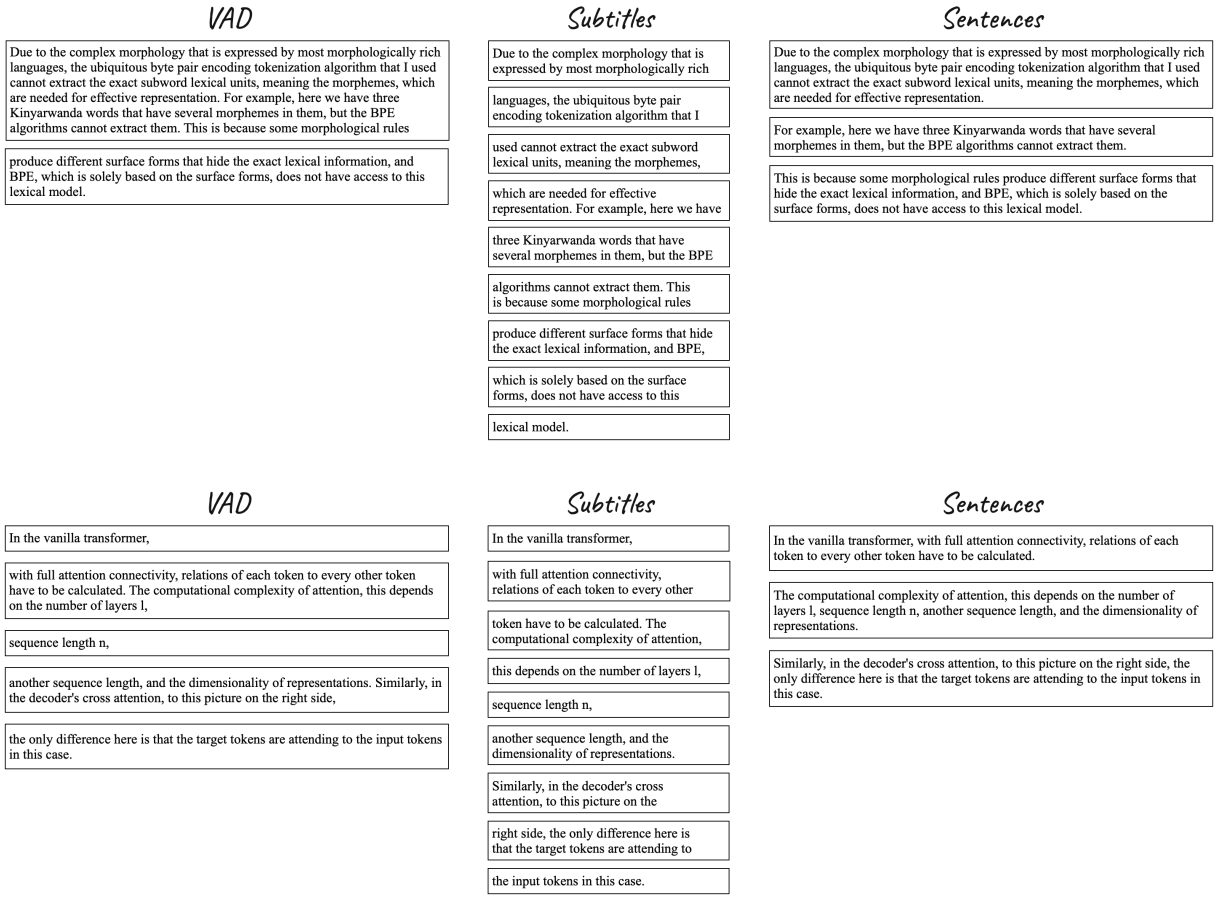


Figure 12: Examples of each discussed transcript segmentation approach for sample data from the development set.