Constructing a Multimodal, Multilingual Translation and Interpreting Corpus: A Modular Pipeline and an Evaluation of ASR for Verbatim Transcription

Alice Fedotova[1], Adriano Ferraresi[1,*], Maja Miličević Petrović[1] and Alberto Barrón-Cedeño[1]

[1]DIT, Università di Bologna, Corso della Repubblica 136, 47121, Forlì, Italy

Abstract

This paper presents a novel pipeline for constructing multimodal and multilingual parallel corpora, with a focus on evaluating state-of-the-art automatic speech recognition tools for verbatim transcription. The pipeline was developed during the process of updating the European Parliament Translation and Interpreting Corpus (EPTIC), leveraging recent NLP advancements to automate challenging tasks like multilingual alignment and speech recognition. Our findings indicate that current technologies can streamline corpus construction, with fine-tuning showing promising results in terms of transcription quality compared to out-of-the-box Whisper models. The lowest overall WER achieved for English was 0.180, using a fine-tuned Whisper-small model. As for Italian, the lowest WER (0.152) was obtained by the Whisper Large-v2 model, with the fine-tuned Whisper-small model still outperforming the baseline (0.201 vs. 0.219).

Keywords

multimodal corpora construction, translation and interpreting corpora, verbatim automatic speech recognition

1. Introduction

The present paper introduces a pipeline for the construction of multimodal and multilingual parallel corpora that could be used for translation and interpreting studies (TIS), among others. The construction of such resources has been acknowledged as a "formidable task" [1], which if automated —as we propose— involves a number of subtasks such as automatic speech recognition (ASR), multilingual sentence alignment, and forced alignment, each of which poses its own challenges. Yet tackling these subtasks also offers a unique way to evaluate state-of-the-art natural language processing (NLP) tools against a unique, multilingual benchmark. In this paper we discuss the development of a modular pipeline adaptable for each of these subtasks and address the issue of whether performing ASR with OpenAI's Whisper [2] could be suitable for verbatim transcription.

We showcase the utility of this pipeline by expanding the European Parliament Translation and Interpreting Corpus (EPTIC), a multimodal parallel corpus comprising speeches delivered at the European Parliament along with their official interpretations and translations [1, 3]. The transcription conventions adopted for the compilation of EPTIC were developed *ad hoc* and aim at reproducing minimal prosodic features, but can still be considered an instance of verbatim transcription [3, 1]; the issue of what truly constitutes *verbatimness* is still an object of debate and will be further discussed. There is fairly widespread agreement on the statement that every transcription system reflects a certain methodological approach [4, 5], and that by "choosing not to transcribe a particular dimension, the researcher has implicitly decided that the dimension plays no role in the phenomenon in question" [4]. To investigate the characteristics of Whisper's [2] transcriptions in English and Italian, we formulate the following two research questions: **RQ1** Is it possible to use fine-tuning to adapt the transcription style to the one of an expert annotator? **RQ2** What is the impact of speech type (native, non-native, interpreted) on transcription quality?

We find that satisfactory results can be achieved with automatic speech recognition, although challenges remain, especially with regards to the verbatimness of the transcription —a crucial factor in corpora intended for TIS. Fine-tuning Whisper-small on English data obtains a lower word error rate (WER) of 0.180 compared to Whisper-large v2 (0.194), potentially indicating that fine-tuning Whisper models holds promise for improving their performance in terms of adhering to a certain transcription style. However, this was not the case when considering the experiments based on Italian. In the Italian scenario, Whisper-large-v2 obtained a WER of 0.152 compared to a WER of 0.201 obtained by the fine-tuned Whisper-small model. It should be noted, however, that this constituted an improvement over the baseline Whisper-small model, which obtained a higher WER of

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy

*Corresponding author.

alice.fedotova2@unibo.it (A. Fedotova); adriano.ferraresi@unibo.it (A. Ferraresi); maja.milicevic2@unibo.it (M. Miličević Petrović); a.barron@unibo.it (A. Barrón-Cedeño)

0009-0001-4850-0974 (A. Fedotova); 0000-0002-6957-0605 (A. Ferraresi); 0000-0003-4137-1898 (M. Miličević Petrović); 0000-0003-4719-3420 (A. Barrón-Cedeño)

0.219. A significant limitation in the case of fine-tuning in Italian was constituted by the smaller amount of data available for tuning compared to English. Lastly, we find that sentence alignment can be facilitated through state-of-the-art embedding-based tools, whereas forced alignment can be considered a largely solved problem. This makes the construction of corpora such as EPTIC more streamlined and requiring less human intervention, with wider implications for multilingual corpus construction in the field of TIS and beyond.

2. Related Work

Recent advancements in the field of corpus linguistics have led to a multitude of complex multilingual and multimodal corpora, as well as novel approaches to corpus construction. Transcribing spoken data, identifying prosodic features, and aligning parallel texts are some of the tasks that are commonly involved. In this sense, a particularly representative case in point is constituted by interpreting corpora, such as EPIC [6], DIRSI [7], and EPTIC [3, 1], the latter also including translated texts. Based on data obtained from the European Parliament, these complex corpora require multi-step approaches for gathering and processing parallel, multilingual texts and multimodal data. Though the construction of translation and interpreting corpora has been largely carried out manually, it can also constitute a unique opportunity for developing new tools and benchmarking recent advancements in the fields of NLP and ASR. ASR, in particular, has garnered increasing attention due to the time-consuming nature of spoken data transcription.

A related research strand in the field of ASR concerns the level of detail of the transcriptions produced by ASR systems, as the task is usually not only to transcribe the speech but to make sure that prosodic features, such as disfluencies, are maintained. [8] conducted a comprehensive comparison of different ASR systems and acoustic models for disfluency detection and categorization, examining Wav2Vec [9], HuBERT [10], WavLM [11], Whisper [2], and Azure [12]. Their findings indicate that fine-tuned models generally outperform their off-the-shelf counterparts. [13] evaluated pre-trained models, revealing that Whisper-Large achieved the best overall WER and chrF (character n-gram F-measure [14]) scores. [15] demonstrated the potential of Whisper for adaptation in spoken language assessment with limited training data. In the realm of commercial ASR services, [16] explored IBM's offering for transcribing English source speeches and their interpretation, reporting an impressively low error rate of 4.7%. [17] conducted a systematic comparison of automatic transcription tools, evaluating factors such as data protection, accuracy, time efficiency, and costs for English and German interviews, and found that Whisper performs best overall among the tools considered.

Despite these advancements, several limitations persist in the current research. First, most studies focus primarily on English, with only some including other languages such as Chinese [16]. Furthermore, the field of speech disfluency research faces challenges due to the scarcity of publicly available benchmarking datasets, attributed to high annotation costs, the clinical nature of some tasks, and the use of proprietary datasets [18]. The choice between Wav2Vec and Whisper remains a point of debate, with [8] finding similar results for both after fine-tuning, while Azure off-the-shelf performed best, followed by Whisper off-the-shelf. Still, [17] did not explore fine-tuning, and [8] suggests that fine-tuned models generally perform better. The requirement for punctuation marks in some corpora, such as EPTIC, introduces another consideration in model selection. Wav2Vec does not output punctuation, while Whisper does, potentially influencing its suitability for certain applications. Additionally, while [13] used a large corpus, [15] indicated that Whisper can perform well with less data, highlighting the need for further investigation into optimal data requirements.

3. Corpus Construction

The present work is based on the European Parliament Translation and Interpreting Corpus (EPTIC), a multimodal parallel corpus comprising speeches delivered at the European Parliament (EP) along with their official interpretations and translations.[1] Within EPTIC, the corpus construction process revolves around individual speech events, where edited *verbatim* reports published by the EP and transcriptions of the speeches are accompanied by transcriptions of interpretations and official translations into other languages. These components form a multi-parallel corpus, i.e. a corpus containing verbatim transcriptions of source speeches, official verbatim reports and corresponding target translations and interpretations (quasi parallel at the intermodal level [3]). The English partition consists of source English texts and their translations into various languages. Corpora containing translations in both possible directions (e.g., from English to French and vice versa) are referred to as bidirectional, while those with translations in only one direction are referred to as unidirectional. Table 1 shows the languages included and the size of the latest version, EPTIC v2, planned for release by the end of 2024.

Our approach to corpus expansion began with a review of previous guidelines for developing EPTIC [1, 19]. The former procedure first involved obtaining data by either scraping texts from the EP website[2] or by man-

[1] https://corpora.dipintra.it/eptic/

[2] https://www.europarl.europa.eu/plenary/en/debates-video.html

Table 1
Token counts, by language, of the latest version of EPTIC.

	Sources		Targets	
Language	Spoken	Written	Interpr.	Transl.
English	43,138	41,047	55,109	58,651
French	35,648	34,063	31,935	35,566
Italian	21,208	20,646	27,329	31,816
Polish	9,458	9,193	–	–
Slovene	–	–	19,717	22,476
German	–	–	18,258	19,822
Finnish	–	–	11,624	12,045

ually downloading videos and then transcribing them. Transcripts of the original speeches and interpretations were manually adapted following editing conventions to annotate features of orality such as disfluencies and timestamped using Aegisub.[3] Then, the texts were automatically segmented into sentences and aligned across languages and modalities, for instance between transcriptions and verbatim reports, with the help of the Intertext Editor alignment tool.[4]

The creation of the new workflow started with the previous procedure as a basis. It was first subdivided into separate tasks, the main ones being automatic speech recognition, multilingual sentence alignment, and forced alignment. Software selection was based on criteria such as ease of use and setup, compatibility with the Python programming language, linguistic coverage, and compatibility with Sketch Engine, an established corpus query tool for teaching and research [20, 21]. Python v. 3.11.5 was used along with the Poetry[5] package manager for portability.[6] Next, we discuss the tasks and the considerations made when designing the pipeline.

Automatic Speech Recognition has seen recent advancements, with the introduction of Whisper [2] and Wav2Vec 2.0 [9]. However, achieving a reasonable level of transcription quality is complex and context-dependent, as it can be interpreted and evaluated differently depending on the domain, task, and application [22]. We decided to employ the WhisperX[7] variant of Whisper, given its documented reliable performance for long-form transcription, which is oftentimes needed when dealing with parliamentary speech [23].

Sentence Alignment involves identifying and aligning parallel sentences, both mono- and multilingually. For this task, we use Bertalign [24]. Unlike predecessors such as Hunalign[8] that rely on lexical translation probabilities, Bertalign employs sentence embeddings to identify parallel sentences, providing a more robust approach for handling semantic similarities. We used a version of the tool that has been extended to produce outputs in the Sketch Engine format for corpus indexing [20, 21].

Forced Alignment, the task of automatically aligning audio with transcriptions, is the most mature task for spoken corpora. Although WhisperX performs timestamping during transcription, we experimented with forced alignment on an existing portion of spoken EPTIC data, using the aeneas library, which supports more than thirty languages.[9]

The pipeline is structured in a modular fashion so as to maximize reusability. The process begins with the extraction of text and video data from the EP website, using ad-hoc scripts which partially automate scraping of the EP website. Transcription is then performed using WhisperX. To remove mistranscriptions and to ensure adherence to the transcription guidelines, the transcripts undergo manual review to incorporate disfluencies and rectify potential mistranscriptions. Once the texts have been transcribed, they undergo sentence splitting and sentence alignment using Bertalign. Relevant metadata, encompassing session topics, are automatically retrieved from the EP website. The only item requiring manual input is the speech type, which can be defined as impromptu, read out, or mixed. After exporting the alignments in the Intertext format and performing part-of-speech tagging with Sketch Engine, the texts and metadata are converted to the vertical format required for indexing in Sketch Engine [20, 21].

4. ASR for Verbatim Transcription: Evaluating Whisper

We require an ASR system to produce a verbatim transcription where all words are transcribed, along with disfluencies and extra-linguistic information. However, *verbatimness* is a broad concept, given the variety of transcription conventions existing in linguistics [17]. Whisper has been observed to produce transcripts "often almost comparable to the final read through of a manual (verbatim to gisted) transcript" [17], where gisted refers to a transcription that "omits non-essential information (e.g., filler words, word fragments, repetition of words), and summarizes or grammatically correctly rephrases the audio content" [17]. Hereby, we define a verbatim

[3] https://aegisub.org
[4] https://wanthalf.saga.cz/intertext
[5] https://python-poetry.org
[6] The code is available at https://github.com/TinfFoil/eptic_v2_pipeline
[7] https://github.com/m-bain/whisperX
[8] https://github.com/danielvarga/hunalign
[9] https://www.readbeyond.it/aeneas/

Table 2
Performance of Whisper by language, expressed in WER.

Model	English	Italian	French	Slovenian
Small	0.212	0.219	0.162	0.463
Small-FT	**0.180**	0.201	–	–
Medium	0.196	0.173	0.213	0.327
Large-v2	0.194	**0.152**	**0.118**	**0.262**

Table 3
Speech performance across types, expressed in WER.

Speech Type	English	Italian
Native	0.104	0.131
Non-native	0.110	–
Interpreted	0.222	0.188

transcription as a transcription where "all words are transcribed without additional grammatical corrections [and] word repetitions, utterances, word interruptions, and elisions are kept" along with some rudimentary extra-linguistic contextual information, such as applauses [17].

As part of our experiments, we tested the HuggingFace release[10] of the Whisper models. The test set included English, Italian, French, and Slovenian, though further experiments were conducted exclusively with English and Italian due to dataset limitations. We used 7 hours of audio for English, 5 for Italian, 1.5 hours for French and 1.5 hours for Slovenian. Besides evaluating the models on the whole set of held-out data, we computed word error rates (WERs) for different speech types: native speech, non-native speech, and interpreted speech.[11] In addition to experimenting with the out-of-the-box versions of Whisper, we explored fine-tuning Whisper-small for English and Italian. To train and test the models, we used 80% of the data for training, 10% for validation, and 10% for testing. The training parameters for the Whisper-small model were set to a batch size of 16, a learning rate of 1e-5, mixed-precision training enabled, and a maximum of 5,000 training steps. Evaluation and saving checkpoints were enabled every 1,000 steps, optimizing for WER.

The experimented Whisper models showed a robust performance across languages and speech types. Our findings suggest that satisfactory results can be achieved for Italian, which exhibits a low WER of 0.152, and English, with a WER as low as 0.194. The full set of results is presented in Table 2, where the fine-tuned model is referenced as Small-FT. This fine-tuned model obtained the lowest WER for English, performing better than Whisper-large-v2, which could indicate that the model is learning to produce a more verbatim transcription. In the case of Italian, the fine-tuned model obtains a lower WER compared to the baseline Whisper-small model (0.201 for the fine-tuned model compared to the WER of 0.219 obtained by the baseline Whisper-small). However, the lowest WER of 0.152 is obtained by Whisper-large-v2, which could be attributed to the lower amount of data available for fine-tuning compared to English.

Lastly, to address **RQ2**, we evaluated whether factors such as *nativeness* influenced the WER. Findings for these experiments are presented in Table 3, and indicate a WER of 0.104 for native English speakers, 0.110 for non-native speakers, and a notably higher WER of 0.222 for interpreted speech. Similar results were also obtained for Italian, with a WER of 0.131 for native speakers and 0.188 for interpreted speech, which provides further evidence for the finding of interpreted speech being more challenging to transcribe [16].

To further explore the claim that fine-tuning improves the performance of the model by steering its output towards a more verbatim transcription, we now present the results of a qualitative error analysis. We consider a set of "markers of verbatimness" based on the definition in [17]: contractions, truncated words, discourse markers, repetitions, filled pauses and empty pauses. The following paragraphs present results that emerge from the analysis, with examples provided in Table 4. Following [15], we furthermore report the recall metric for each category.

As for contractions, they are sometimes incorrectly resolved by the standard Whisper-large-v2 model; fine-tuning results in improvements. For instance, in the example shown in Table 4, the fine-tuned version of Whisper-small maintains the contraction while the large model does not. Generally, however, Whisper-large-v2 shows acceptable performance even when fine-tuning is not performed, as Whisper was trained with unnormalized transcripts including contractions, punctuation and capitalization [2].

Truncations are not transcribed by the Whisper models out-of-the-box. Fine-tuning shows some promising results, though truncations are not always transcribed reliably and transcription errors are sometimes introduced, as illustrated in Table 4. This is possibly due to the observation in [15] that, being largely trained on speech data with a high level of inverse text normalization (ITN), a process including disfluency removal, Whisper tends to omit features of orality in favor of readability, which is unfavorable for the purpose of verbatim transcription.

Discourse markers are mostly transcribed in English, even by the baseline Whisper-large-v2. In Italian, discourse markers are omitted considerably more often. An example of this is provided in Table 4. This could be attributed to the fact that, even though Whisper models have been trained to produce transcriptions without any significant standardization [2], the amount and qual-

[10] https://huggingface.co/docs/transformers/en/model_doc/whisper

[11] Which can be both into the interpreter's A or B language.

Table 4
Transcription examples by disfluency type. For each example, we include (a) the reference transcription, (b) the transcription produced by Whisper-**small-FT** and (c) by Whisper-**large-v2**.

Example Transcription	Rec EN	Rec IT
Contractions		
(a) **I'm** encouraged that the interim leadership …	100.00	–
(b) **I'm** encouraged that the interim leadership …	95.40	–
(c) **I am** encouraged that the interim leadership …	86.30	–
Truncations		
(a) …foreign **direct in**- ehm investment …	100.00	100.00
(b) …foreign **directin**- ehm investment …	58.20	60.00
(c) …foreign **direct investment** …	0.00	0.00
Discourse markers		
(a) …la conduzione della famiglia regnante **diciamo**.	100.00	100.00
(b) …la conduzione della ehm famiglia regnante, **diciamo**.	97.50	90.40
(c) …la conduzione della famiglia regnante.	97.40	66.60
Repetitions		
(a) …**but I w- I would** urge you, if you're interested …	100.00	100.00
(b) …**but I w- I would** urge you, if you're interested …	90.40	90.90
(c) …**but I would** urge you if you're interested …	0.00	0.00
Empty pauses		
(a) …azioni che **…** rivelano il volto opprimente …	100.00	100.00
(b) …azioni che **…** rivellano il volto frimente …	84.40	78.20
(c) …azioni che rivelano il volto frimente …	0.00	0.00
Filled pauses		
(a) …azioni che **…** rivelano il volto opprimente …	100.00	100.00
(b) …azioni che **…** rivellano il volto frimente …	56.50	88.20
(c) …azioni che rivelano il volto frimente …	0.00	0.00

ity of training data for English are likely more extensive and varied compared to Italian, especially when it comes to examples of spontaneous speech. As for repetitions, the example in Table 4 shows both a repetition and a truncation, a common occurrence due to disfluent speech often comprising a combination of both. In the example, the fine-tuned Whisper-small model accurately transcribes both disfluencies, while Whisper-large-v2 rephrases them into a corrected transcription. Overall, the baseline Whisper-large-v2 model always omitted repetitions both in English and Italian. This could be due to the powerful language model used by Whisper, which has been observed to correct such errors [13].

The last examples in Table 4 illustrate transcriptions of empty and filled pauses. Whereas Whisper-small-FT often captures them, the baseline model does not. However, the fine-tuned model's performance is not consistent, and occasionally non-existent empty pauses are transcribed by the model. As in the case of truncations, pauses are never transcribed by Whisper-large-v2, likely due to the models having been trained on data processed with ITN.

5. Conclusions and Future Work

This paper presented a novel pipeline for constructing multimodal and multilingual parallel corpora, with a focus on evaluating state-of-the-art automatic speech recognition tools for verbatim transcription. Experiments with Whisper models on EPTIC revealed robust performance across languages and speech types, particularly for English and Italian. However, some limitations remain regarding ASR performance and achieving verbatim transcriptions. Fine-tuning Whisper showed promising reductions in WER, particularly for English, indicating the potential of adapting the model to use a more verbatim style. Yet qualitative analysis revealed inconsistencies in handling disfluencies, truncations, and discourse markers. Furthermore, higher WERs for non-native and interpreted speech underscore remaining challenges.

Future research efforts could explore incorporating additional metrics beyond WER to better capture the degree of *verbatimness* in the transcriptions, and expanding the Italian dataset to potentially improve the performance of the fine-tuned model. Another avenue for research could include augmenting the dataset with external data containing pairs of audio and verbatim transcripts, most notably the Switchboard corpus introduced in [25]. Other methods besides fine-tuning could be explored to enhance the quality of transcriptions, for instance by leveraging the official verbatim reports on the European Parliament's website. Lastly, a model could be developed for detecting the metadata item relative to the speech type, i.e. impromptu, read out, or mixed, based on textual or multimodal features.

Acknowledgments

The work of A. Fedotova is supported by the NextGeneration EU programme, ALMArie CURIE 2021 - Linea SUpER, Ref. CUPJ45F21001470005.

References

[1] S. Bernardini, A. Ferraresi, M. Russo, C. Collard, B. Defrancq, Building interpreting and intermodal corpora: A how-to for a formidable task, in: C. B. Mariachiara Russo, B. Defrancq (Eds.), Making Way in Corpus-Based Interpreting Studies, Springer, Singapore, 2018, pp. 21–42. doi:10.1007/978-981-10-6199-8_2.

[2] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, I. Sutskever, Robust speech recognition via large-scale weak supervision, in: Proceedings of the 40th International Conference on Machine Learning, Proceedings of Machine Learning Research, 2023, pp. 28492–28518. URL: https://proceedings.mlr.press/v202/radford23a.html, retrieved May 13, 2024.

[3] S. Bernardini, A. Ferraresi, M. Miličević, From epic to eptic — exploring simplification in interpreting and translation from an intermodal perspective, Target. International Journal of Translation Studies 28 (2016) 61–86. doi:https://doi.org/10.1075/target.28.1.03ber.

[4] R. J. Kreuz, M. A. Riordan, The transcription of face-to-face interaction, in: W. Bublitz, N. R. Norrick (Eds.), Foundations of Pragmatics, De Gruyter, Berlin, 2011, pp. 657–679. doi:10.1515/9783110214260.657.

[5] J. C. Lapadat, A. C. Lindsay, Transcription in research and practice: From standardization of technique to interpretive positionings, Qualitative Inquiry 5 (1999) 64–86. doi:10.1177/107780049900500104.

[6] M. Russo, C. Bendazzoli, A. Sandrelli, N. Spinolo, The european parliament interpreting corpus (epic): Implementation and developments, in: Breaking Ground in Corpus-Based Interpreting Studies, Springer, 2012, pp. 53–90.

[7] C. Bendazzoli, From international conferences to machine-readable corpora and back: An ethnographic approach to simultaneous interpreter-mediated communicative events, in: Breaking Ground in Corpus-Based Interpreting Studies, volume 147, Springer, 2012, pp. 91–117.

[8] A. Romana, K. Koishida, E. M. Provost, Automatic disfluency detection from untranscribed speech, arXiv preprint arXiv:2311.00867 (2023).

[9] A. Baevski, Y. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, Advances in Neural Information Processing Systems 33 (2020) 12449–12460. URL: https://10.48550/arXiv.2006.11477, retrieved May 19, 2024.

[10] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, A. Mohamed, Hubert: Self-supervised speech representation learning by masked prediction of hidden units, IEEE/ACM transactions on audio, speech, and language processing 29 (2021) 3451–3460.

[11] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, et al., Wavlm: Large-scale self-supervised pre-training for full stack speech processing, IEEE Journal of Selected Topics in Signal Processing 16 (2022) 1505–1518.

[12] Microsoft Azure, Speech to text – audio to text translation | microsoft azure, https://azure.microsoft.com/en-us/products/ai-services/speech-to-text, 2024. Accessed: 2024-07-17.

[13] J. Michot, M. Hürlimann, J. Deriu, L. Sauer, K. Mlynchyk, M. Cieliebak, Error-preserving automatic speech recognition of young english learners' language, arXiv preprint arXiv:2406.03235 (2024).

[14] M. Popović, chrF: character n-gram F-score for automatic MT evaluation, in: O. Bojar, R. Chatterjee, C. Federmann, B. Haddow, C. Hokamp, M. Huck, V. Logacheva, P. Pecina (Eds.), Proceedings of the Tenth Workshop on Statistical Machine Translation, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 392–395. URL: https://aclanthology.org/W15-3049. doi:10.18653/v1/w15-3049.

[15] R. Ma, M. Qian, M. Gales, K. Knill, Adapting an asr foundation model for spoken language assessment, arXiv preprint arXiv:2307.09378 (2023).

[16] X. Wang, B. Wang, Exploring automatic methods for the construction of multimodal interpreting corpora. how to transcribe linguistic information and identify paralinguistic properties?, Across Languages and Cultures 25 (2024) 48–70. URL: https://eprints.whiterose.ac.uk/212127/.

[17] S. Wollin-Giering, M. Hoffmann, J. Höfting, C. Ventzke, Automatic transcription of qualitative interviews, Forum Qualitative Sozialforschung Forum: Qualitative Social Research 25 (2023). doi:https://doi.org/10.17169/fqs-25.1.4129.

[18] P. Mohapatra, S. Likhite, S. Biswas, B. Islam, Q. Zhu, Missingness-resilient video-enhanced multimodal disfluency detection, arXiv preprint arXiv:2406.06964 (2024).

[19] M. Kajzer-Wietrzny, A. Ferraresi, Guidelines for EPTIC collaborators, 2020. Unpublished manuscript.

[20] P. Rychlý, Manatee/bonito-a modular corpus manager, RASLAN (2007) 65–70. URL: https://www.sketchengine.eu/wp-content/uploads/Manatee-Bonito_2007.pdf, retrieved May 14, 2024.

[21] A. Kilgarriff, V. Baisa, J. Bušta, M. Jakubíček, V. Kovář, J. Michelfeit, P. Rychlý, V. Suchomel,

The sketch engine: Ten years on, Lexicography 1 (2014) 7–36. doi:https://doi.org/10.1007/s40607-014-0009-9.

[22] K. Kuhn, V. Kersken, B. Reuter, N. Egger, G. Zimmermann, Measuring the accuracy of automatic speech recognition solutions, ACM Transactions on Accessible Computing 16 (2024) 1–23. doi:https://doi.org/10.1145/3636513.

[23] M. Bain, J. Huh, T. Han, A. Zisserman, Whisperx: Time-accurate speech transcription of long-form audio, arXiv preprint (2023). URL: https://arxiv.org/pdf/2303.00747, retrieved May 20, 2024.

[24] L. Lei, M. Zhu, Bertalign: Improved word embedding-based sentence alignment for chinese–english parallel corpora of literary texts, Digital Scholarship in the Humanities 38 (2022) 621–634. doi:https://doi.org/10.1093/llc/fqac089.

[25] J. J. Godfrey, E. C. Holliman, J. McDaniel, Switchboard: Telephone speech corpus for research and development, in: Acoustics, Speech, and Signal Processing, IEEE International Conference on, volume 1, IEEE Computer Society, 1992, pp. 517–520. doi:10.1109/ICASSP.1992.225858.