Speech-to-text recognition for multilingual spoken data in language documentation

Lorena Martín Rodríguez and **Christopher Cox**
School of Linguistics and Language Studies / Carleton University
lorenamartinrodrigue@cmail.carleton.ca / christopher.cox@carleton.ca

Abstract

More than 85% of the languages spoken in Canada are deemed as vulnerable (Lewis, 2009). Efforts for language revitalization and maintenance are essential to maintain both the Indigenous knowledge and cultural diversity in the country. In this context, it is relevant to develop computational tools that may aid language communities and linguists involved in the process of language documentation and revitalization. One challenge of the documentation process is the existence of more than one language in spoken language materials. We propose a pipeline to handle multilingual spoken data using both custom-trained and commercial speech-to-text services. The main advantage of our approach is its user-friendly format as a recognizer for the audiovisual annotation application ELAN (Brugman and Russel, 2004), which facilitates its use in both community and university-based language work.

1 Introduction

Canada is the home of a wide language diversity, with almost 200 languages comprised in 11 Indigenous language families and isolates (Lewis, 2009). Many of these languages are spoken by First Nations, Inuit, and Métis peoples and are deemed as "fundamental to the identities, cultures, spirituality, relationships to the land, world views and self-determination of Indigenous peoples" (Branch, 2020). However, this linguistic diversity is endangered. More than 85% of the languages spoken in Canada are deemed as vulnerable (Lewis, 2009). Efforts for language revitalization and maintenance are essential to maintain Indigenous knowledge and cultural diversity in the country. In this context, it is relevant to develop computational tools that may aid language communities and linguists involved in the process of language documentation and revitalization.

One challenge of the documentation process is the existence of more than one language in the data. In many language documentation and revitalization initiatives, audiovisual recordings frequently include interactions that involve multiple varieties of languages. In projects focused on grammatical or lexical description, for instance, it is not uncommon for consultation sessions to feature bilingual (or, in some cases, multilingual) elicitation, interpretation, and/or translation involving both local languages and one or more languages of wider communication (e.g., lingue franche). In other cases, documentation of linguistic knowledge and practices may also feature multiple languages as a consequence of patterns of multilingualism that are part of the fabric of the local language ecology. Bilingual or multilingual documentation such as this poses additional challenges for annotation efforts that aim to make such resources more accessible for use in advocacy, revitalization, and research, beyond what has already commonly been reported as part of the broader "transcription bottleneck" (Reiman, 2010; Boerger, 2011) that current speech technologies applied in documentary linguistic contexts have generally sought to address (cf. Thieberger, 2016; Adams et al., 2018, inter alia).

We propose a solution for handling spoken multilingual data involving well-documented and under-resourced languages. We develop a modular speech-to-text pipeline to handle spoken multilingual data in English and Tsuut'ina (ISO 639-3: srs, Glottocode: sars1236), a Na-Dene language spoken by members of the Tsuut'ina Nation (Treaty 7, southern Alberta, Canada). Our approach makes use of existing speech-to-text services to provide automatic transcriptions of English utterances in tandem with a fine-tuned XLS-R model (Babu et al., 2022) to provide automatic transcriptions of Tsuut'ina utterances. These models are integrated into ELAN (Brugman and Russel, 2004), an open-source desktop application for audiovisual annotation widely used in language documentation and revitalization, to provide a user-friendly interface that may be employed by language workers without requiring extensive knowledge of computer science.

2 Literature Review

Extensive use of multilingual practices has been documented for speakers of minority languages belonging to a variety of language families and geographical areas. Research on under-researched languages has explored these code-mixing and code-switching practices in languages spoken in South America belonging to the Matacoan family (Campbell and Grondona, 2010) and the Eastern Tukanoan family (Silva, 2020); in Austronesian languages spoken in Vanuatu (Lindstrom, 2007); in Tibeto-Burman language spoken in Eastern Nepal (Stoll et al., 2015), as well as in Inuit languages in Canada (Allen et al., 2009). These multilingual practices do not conform to a homogeneous group of practices, but are realized differently for each language community and are shaped by language ideologies and cultural practices (Pakendorf et al., 2021). The documentation of these multilingual practices offers valuable insights into the linguistic practices and language ideologies of the communities, as well as into the status and vitality of languages by documenting possible cases of language-shifting. Therefore, it is relevant to develop computational tools to address multilingual practices present in the documentation.

A variety of approaches have been proposed for automatic speech recognition (ASR) of multilingual utterances in spoken data. Lin et al. (2009) defined a universal phone set across alphabets to improve their ASR model performance in multilingual contexts. In recent years, efforts have increased in developing quality multilingual models for ASR. Yadav and Sitaram (2022)'s survey of state-of-the-art models indicates that in certain models, multilingual pre-trained models outperform models pre-trained with monolingual data. Their survey highlights the relevance of the features selected in the models' general performance, as well as the choice of training languages. Although cross-lingual transfer methods generally perform better across languages of the same family, recent methods have achieved accurate results in languages that do not belong to the same language family as those in the training data, which indicates that the features selected may have more impact on the performance of the models than the choice of training languages.

One problem with several of the existing approaches for multilingual speech recognition is that they are not implemented in a user-friendly interface and require users with specialized knowledge in computer science. Recent efforts have started implementing state-of-the-art systems in user-friendly interfaces to promote wider access to these resources by non-specialized users and language communities. Foley et al. (2018) propose Elpis, a user-friendly pipeline that allows language documentation workers and language communities to develop their own speech recognition models. Other work has focused on incorporating resources into existing language documentation tools in the form of plugins. Adams et al. (2021) integrate a speech recognition toolkit, ESPNet, into Elpis to provide user-friendly access to readily available speech recognition technologies. Cox (2019) integrates an automatic phoneme transcription tool, Persephone (Adams et al., 2018), into ELAN, with Partanen et al. (2020) applying this extension to automatic phoneme recognition for Samoyedic languages. Our approach continues with these efforts of implementing speech recognition systems in user-friendly interfaces by integrating existing commercial and open-source models as extensions to ELAN.

3 Integrating speech technology into ELAN

This paper introduces the implementation of different models for speech detection and automatic speech recognition as extensions to ELAN. ELAN is widely adopted in language work involving audio and video recordings, serving as the current *de facto* standard for audiovisual annotation in documentary linguistics (Carreau et al., 2018). Its popularity among users involved in language documentation and revitalization, in particular, motivates the creation of tools that contribute to speed the task of language annotation, such as those presented in the current paper. By integrating existing models into the software, our proposal aims to make state-of-the-art speech recognition systems more widely available to language communities.

Since ELAN is open-source software, it is possible to incorporate new features by editing and recompiling the Java source code for the application itself. This has been the approach taken to date with import and export options within ELAN, for instance, with each supported format implemented as hard-coded Java classes that provide both the application logic and necessary user interface components. While this method of integration offers ready access to data structures internal to ELAN that may be helpful for conversion between file formats, it requires a degree of familiarity with both the ELAN code base and with Java software development practices, and may not always be amenable to integrating features from third-party packages that have not been developed in Java.

As an alternative to this approach, the AVATecH project (Tschöpel et al., 2011) developed an API within ELAN that aimed to facilitate the integration of free-standing audio and video analysis components as extensions, or 'recognizers'. Within this framework, recognizers could be developed either in Java, thereby having the ability to access data structures internal to ELAN and implement custom user interface components; or as an external, executable application developed in any programming language, interacting with ELAN via an XML-based protocol and exposing any user-defined input and output parameters via the ELAN user interface through a static, CMDI-based metadata definitions. Both 'native' (i.e., Java-based) and 'local' (non-Java-based) recognizers allow for external services to interact with ELAN without requiring changes to the ELAN source code itself. Importantly, from the perspective of ELAN users who may not have a background in language technology or software development, recognizers such as these provide a straightforward means of applying speech and language technologies to textual annotations and audiovisual materials represented in their ELAN documents from directly within the ELAN user interface, without having to learn how to use another application or service and import its results into ELAN.

The current paper presents a set of ELAN recognizers publicly available that automatizes parts of the annotation process for multilingual data. These recognizers can be used as stand-alone services or in combination with each other to generate linguistic outputs as ELAN tiers. The collection of multilingual audio data in Tsuut'ina and English provides a case study of how these recognizers are used in real linguistic documentation work and how they handle multilingual data in the context of elicitation sessions.

3.1 Voxseg-ELAN and SileroVAD-ELAN

As Hjortnæs et al. (2020, 36) note, applying speech recognition and other commonly targeted forms of language technology to new audiovisual recordings in language documentation also requires segmentation (and, for recordings involving more than one recorded participant, often speaker diarization, as well). Fundamental annotation tasks such as these are often still accomplished manually in language documentation, and contribute to the challenge of working with such materials.

As an initial step towards addressing this aspect of the annotation bottleneck beyond the silence vs. non-silence audio event detection provided by the Silence Recognizer bundled with ELAN[1], we have sought to integrate two current, DNN-based voice activity detection (VAD) models into ELAN, which distinguish speech from non-speech segments in unannotated audio recordings. This includes both the Voxseg voice activity detection (VAD) package (Wilkinson and Niesler, 2021), which provides a full VAD pipeline and pre-trained VAD model implemented in Python; and Silero-VAD (Silero Team, 2021), a model for voice activity detection pre-trained on extensive multilingual corpora.

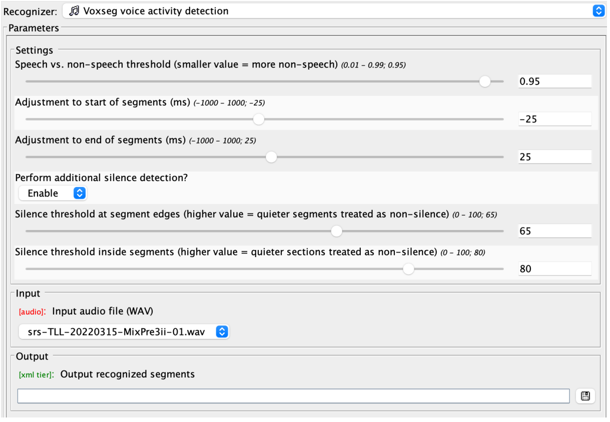
Voxseg-ELAN[2] (Cox, 2022) and SileroVAD-ELAN[3] are local ELAN recognizers that allow users to apply VAD to a selected audio recording in the current ELAN transcript, returning a tier containing annotations representing any sections of speech detected by the underlying model. Figure 1 shows the Voxseg-ELAN recognizer user interface in ELAN, exposing both parameters to the VAD model itself (e.g., the speech vs. non-speech threshold value, which determines the sensitivity of VAD to possible speech events) and post-hoc adjustments made by the recognizer (e.g., fixed adjustments applied to the start and end times of all annotations returned by the model).

Both VAD models succeed in differentiating background noise from voice activity in most cases. However, it has been observed that sibilants placed at the beginning of words are not generally detected by the model. Therefore, the recognizer allows users to manually specify milliseconds of audio that should be added to the start and end of

[1] https://www.mpi.nl/corpus/html/elan/ch05s04s03.html

[2] https://github.com/coxchristopher/voxseg-elan

[3] https://github.com/l12maro/SileroVAD-Elan



Recognizer: Voxseg voice activity detection

Parameters

Settings

Speech vs. non-speech threshold (smaller value = more non-speech) *(0.01 - 0.99; 0.95)*: 0.95

Adjustment to start of segments (ms) *(-1000 - 1000; -25)*: -25

Adjustment to end of segments (ms) *(-1000 - 1000; 25)*: 25

Perform additional silence detection? Enable

Silence threshold at segment edges (higher value = quieter segments treated as non-silence) *(0 - 100; 65)*: 65

Silence threshold inside segments (higher value = quieter sections treated as non-silence) *(0 - 100; 80)*: 80

Input

[audio]: Input audio file (WAV)

srs-TLL-20220315-MixPre3ii-01.wav

Output

[xml tier]: Output recognized segments

Figure 1: Parameters for Voxseg-ELAN

the outputted segments, and/or to apply silence vs. non-silence detection at the start and end of outputted segments to capture sibilants that may have been missed by the VAD model. This allows users to customize the output of the model according to the needs of their annotations. When applied to Tsuut'ina audio recordings in recent, community-based documentation, both VAD models generally produce acceptable results, although we note that some phonemes not found in the languages represented in these models' training data (e.g., ejective stops and affricates) are often not included in the output segments when appearing in utterance-initial position. It may be possible to address this through fine-tuning with Tsuut'ina training data, although this remains to be investigated in future work.

3.2 XLS-R-ELAN

XLS-R-ELAN[4] (Cox, 2023) presents a local recognizer that allows users to apply a fine-tuned XLS-R automatic speech recognition model (Babu et al., 2022) to an audio recording that has been segmented in ELAN. Input segments are taken from a user-specified tier, and may thus be manually created annotations or the output of one of the voice activity detection recognizers mentioned above. Internally, the recognizer converts the input segments into individual audio clips, applies the fine-tuned XLS-R to each one, and returns a new tier containing the recognized text.[5] In the case of Tsuut'ina, using XLS-R-ELAN to apply a XLS-R model that has been fine-tuned on a small audio corpus (14,144 tokens, 8,717 audio clips; 3h52m26s audio total) to segments produced automatically by one of the above VAD recognizers is generally sufficient to create usable first-pass transcriptions of Tsuut'ina speech (CER < 0.2), significantly reducing the overall workload required to annotate new recordings in ELAN.

3.3 Eng-ELAN

Eng-ELAN[6] comprises different state-of-the-art Speech-to-text models for the automatic transcription of English. Three models are made easily available through this recognizer: two commercial (provided by Amazon Web Services and Google Cloud) and one open-source model (Whisper AI). The recognizer incorporates different features, which allow the selection of an existing tier to annotate only certain audio segments or to provide annotations at the utterance or word level. Although the case study focuses on the transcription of English, all of the models provided have been trained with extensive multilingual data and provide support for several languages.

We integrate three existing speech-to-text models for the automatic transcription of well-resourced languages in multilingual corpora as an ELAN recognizer. These models include two commercial models (AWS, Google Cloud), and one open-source model (Whisper AI). Commercial models are chosen for their ready availability and support of quality transcription services for several languages (31 supported languages in AWS, 80 for Google Cloud). However, they provide barriers to user-friendly applications, since they require users to previously create an account and connect their profiles to their desktops using command-line instructions. Moreover, they are not free of cost and charge per transcribed audio second, which may affect the cost of documentation projects which include large quantities of audio data.

On the other hand, Whisper AI is an open-source model that provides quality transcriptions free of cost and does not require the usage of command-line instructions prior to its use. One disadvantage of this model over the commercial one is that it only

[4] https://github.com/coxchristopher/xls-r-elan

[5] Optionally, XLS-R-ELAN is also able to apply a CTC-based word beam search (Scheidl et al., 2018) to decode the predictions of the XLS-R model, drawing on a user-provided text corpus to derive a dictionary and language model that may help refine the XLS-R model's results. While not necessary for basic uses of XLS-R-ELAN, this feature proved useful in cases where written text was already available for a given audio recording (e.g., when the recording represented a reading of an existing written resource), improving the accuracy of the XLS-R model's results.

[6] https://github.com/l12maro/Eng-ELAN

provides utterance-level transcriptions. Therefore, word-level annotations are currently unavailable using this model.

The possibility of specifying an existing ELAN tier as input facilitates the implementation of this recognizer along with the voice activity detection models previously described. In the event that a tier is provided as input, the existing segments are used for utterance-level annotation, and a transcription is provided only for segments with no existing annotations. Research suggests that the segmentation of multilingual audio in monolingual segments improves the performance of ASR models (Ramabhadran et al., 2003). Therefore, when a tier is provided as input, the original audio is split based on the existing spans within the tier.

4 Further improvements

The incorporation of state-of-the-art models into user-friendly interfaces includes numerous challenges. One of them is how to provide a user-friendly installation of the recognizers for their use by language documentation workers. The current installation process requires users to manually retrieve the recognizers from their respective GitHub repositories and perform a manual installation within the corresponding ELAN folder in the user's system. Moreover, additional libraries need to be installed by the user. In order to improve the accessibility of these resources, other options should be explored that do not place the responsibility of installing additional recognizers and libraries on the user side.

Similarly, the implementation of existing models presents new challenges for user-friendly interfaces. Commercial models such as Amazon Web Services or Google Cloud require a prior synchronization of the user accounts with the computer used for the annotations using the command line, which may not be accessible to users not familiar with this type of technology. Moreover, in order to use existing models that require fine-tuning, such as XLS-R for Tsuut'ina, prior knowledge of data management is needed. While other efforts in the area of language documentation technology have succeeded in easing the training process (cf. Foley et al., 2018), some of the recognizers presented in this article still need to overcome these difficulties.

In order to provide a service that can automatically handle multilingual data, a future improvement of this project is to train a language diarization model that may automatically classify the linguistic utterances according to the language spoken. This automatic classification could then be used to feed the segments to a transcription model tailored to the specific language.

In the implementation of models as ELAN recognizers, it is necessary to acknowledge the limitations of the software in terms of what can be automatized. While ELAN allows the creation of hierarchical tier structures and the modification of existing tiers, this must be done manually. Due to limitations in the current XML schema for tiers, hierarchical relations between tier structures cannot be automatically generated by recognizers. Similarly, while some of the current recognizers presented in this work allow the selection of a specific tier as input, it is not possible to return modifications within the specified tiers. Instead, a new tier is returned, which preserves the existing annotations. It is then still necessary that the user manually manipulates the tiers to reflect the desired hierarchical relations or modifications.

Another potential improvement of the present work would be to 'daisy-chain' recognizers, that is, for the output of one to be fed directly in as the input to the next. To our knowledge, this has not been done in ELAN before, but it could be potentially implemented with the creation of a Java-based recognizer in ELAN that connects the different recognizers linearly. Implementing this would automatize the language documentation process from the detection of voice activity to the transcription of linguistic data.

Moreover, language documentation projects generally comprise sets of recordings. However, recognizers in ELAN currently focus on individual recordings. Exploring options for batch processing outside of the software could extend the use of the recognizers to sets of transcripts and recordings, which would therefore speed the process of language documentation.

5 Conclusions

This paper has explored how existing speech technologies can be integrated into one of the most widely used annotation tools in language documentation. The aim of this contribution is to widen access to speech technologies by providing user-friendly interfaces to existing models. This would facilitate the contribution to tasks in language documentation by community members, language workers, and linguists alike.

Moreover, the recognizers shared in this paper provide a framework for the integration of models into ELAN. This framework can be easily adapted to other models and uses in ELAN, as illustrated by Partanen et al. (2020). With this contribution, we hope to encourage other initiatives for the implementation of speech technologies in user-friendly interfaces.

Acknowledgements

We gratefully acknowledge the ongoing contributions of the Tsuut'ina Nation to this work, in particular through the leadership of the Office of the Tsuut'ina Language Commissioner and the Tsuut'ina Gunaha Institute. This research has been funded by a Partnership Grant (895-2019-1012) from the Social Sciences and Humanities Research Council (SSHRC) of Canada.

References

Oliver Adams, Trevor Cohn, Graham Neubig, Hilaria Cruz, Steven Bird, and Alexis Michaud. 2018. Evaluation phonemic transcription of low-resource tonal languages for language documentation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Oliver Adams, Benjamin Galliot, Guillaume Wisniewski, Nicholas Lambourne, Ben Foley, Rahasya Sanders-Dwyer, Janet Wiles, Alexis Michaud, Séverine Guillaume, Laurent Besacier, Christopher Cox, Katya Aplonova, Guillaume Jacques, and Nathan Hill. 2021. User-friendly automatic transcription of low-resource languages: Plugging ESPnet into Elpis. *Proceedings of the Workshop on Computational Methods for Endangered Languages*, 1(2).

Shanley Allen, Fred Genesee, Sarah Fish, and Martha Crago. 2009. *Typological constraints on code mixing in Inuktitut–English bilingual adults*, volume 86, page 273–306. John Benjamins Publishing Company, Amsterdam.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. XLS-R: Self-supervised cross-lingual speech representation learning at scale. In *Interspeech 2022*, page 2278–2282. ISCA.

Brenda H. Boerger. 2011. To BOLDly go where no one has gone before. *Language Documentation and Conservation*, 5:208–233.

Legislative Services Branch. 2020. Consolidated federal laws of Canada, Indigenous Languages Act.

Hennie Brugman and Albert Russel. 2004. Annotating multi-media/multi-modal resources with ELAN. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Lyle Campbell and Verónica Grondona. 2010. Who speaks what to whom? multilingualism and language choice in Misión La Paz. *Language in Society*, 39(5):617–646.

Kathryn Carreau, Melissa Dane, Kat Klassen, Joanne Mitchell, and Christopher Cox. 2018. Integrating collaboration into the classroom: Connecting community service learning to language documentation training. In Wilson de Lima Silva and Katherine J. Riestenberg, editors, *Collaborative Approaches to the Challenges of Language Documentation and Conservation: Selected papers from the 2018 Symposium on American Indian Languages (SAIL)*, volume 20 of *Language Documentation & Conservation Special Publication*, pages 6–19. University of Hawai'i Press, Honolulu, HI.

Christopher Cox. 2019. *Persephone-ELAN: Automatic phoneme recognition for ELAN users*. Version 0.1.2.

Christopher Cox. 2022. *Voxseg-ELAN: Voice activity detection for ELAN users*. Version 0.1.1.

Christopher Cox. 2023. *XLS-R-ELAN: An implementation of XLS-R automatic speech recognition as a recognizer for ELAN*. Version 0.1.0.

Ben Foley, Joshua T. Arnold, Rolando Coto-Solano, Gautier Durantin, T. Mark Ellison, Daan van Esch, Scott Heath, František Kratochvil, Zara Maxwell-Smith, David Nash, Ola Olsson, Mark Richards, Nay San, Hywel M. Stoakes, N. Thieberger, and Janet Wiles. 2018. Building speech recognition systems for language documentation: The CoEDL Endangered Language Pipeline and Inference System (ELPIS). In *Workshop on Spoken Language Technologies for Under-resourced Languages*.

Nils Hjortnæs, Niko Partanen, Michael Rießler, and Francis M. Tyers. 2020. Towards a speech recognizer for Komi, an endangered and low-resource Uralic language. In *Proceedings of the Sixth International Workshop on Computational Linguistics of Uralic Languages*, pages 31–37, Wien, Austria. Association for Computational Linguistics.

M. Paul Lewis. 2009. *Ethnologue: Languages of the world*, 16. ed edition. SIL International, Dallas, Tex.

Hui Lin, Li Deng, Dong Yu, Yi-fan Gong, Alex Acero, and Chin-Hui Lee. 2009. A study on multilingual acoustic modeling for large vocabulary ASR. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4333–4336.

Lamont Lindstrom. 2007. Bislama into Kwamera: Code-mixing and language change on Tanna (Vanuatu).

Brigitte Pakendorf, Nina Dobrushina, and Olesya Khanina. 2021. A typology of small-scale multilingualism. *International Journal of Bilingualism*, 25(4):835–859.

Niko Partanen, Mika Hämäläinen, and Tiina Klooster. 2020. Speech recognition for endangered and extinct Samoyedic languages. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, pages 523–533, Hanoi, Vietnam. Association for Computational Linguistics.

B. Ramabhadran, Jing Huang, and M. Picheny. 2003. Towards automatic transcription of large spoken archives - english ASR for the MALACH project. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, volume 1, pages I–216–I–219, Hong Kong, China. IEEE.

D. Will Reiman. 2010. Basic oral language documentation. *Language Documentation and Conservation*, 4:254–268.

H. Scheidl, S. Fiel, and R. Sablatnig. 2018. Word beam search: A Connectionist Temporal Classification decoding algorithm. In *16th International Conference on Frontiers in Handwriting Recognition*, pages 253–258. IEEE.

Silero Team. 2021. Silero VAD: pre-trained enterprise-grade Voice Activity Detector (VAD), number detector and language classifier. `https://github.com/snakers4/silero-vad`.

Wilson de Lima Silva. 2020. Multilingual interactions and code-mixing in northwest Amazonia. *International Journal of American Linguistics*, 86(1):133–154.

Sabine Stoll, Taras Zakharko, Steven Moran, Robert Schikowski, and Balthasar Bickel. 2015. Syntactic mixing across generations in an environment of community-wide bilingualism. *Frontiers in Psychology*, 6.

Nick Thieberger. 2016. Documentary linguistics: Methodological challenges and innovatory responses. *Applied Linguistics*, 37(1):88–99.

Sebastian Tschöpel, Daniel Schneider, Rolf Bardeli, Oliver Schreer, Stefano Masneri, Peter Wittenburg, Han Sloetjes, Przemek Lenkiewicz, and Eric Auer. 2011. AVATecH: Audio/Video Technology for Humanities Research. In *Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage*, pages 86–89, Hissar, Bulgaria. Association for Computational Linguistics.

Nicholas Wilkinson and Thomas Niesler. 2021. A hybrid CNN-BiLSTM voice activity detector. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6803–6807. ISSN: 2379-190X.

Hemant Yadav and Sunayana Sitaram. 2022. A survey of multilingual models for automatic speech recognition.