Highland Puebla Nahuatl–Spanish Speech Translation Corpus for Endangered Language Documentation

Jiatong Shi
The Johns Hopkins University
jiatong_shi@jhu.edu

Jonathan D. Amith
Gettysburg College
jonamith@gmail.com

Xuankai Chang, Siddharth Dalmia, Brian Yan, and Shinji Watanabe
Carnegie Mellon University
shinjiw@ieee.org

Abstract

Documentation of endangered languages (ELs) has become increasingly urgent as thousands of languages are on the verge of disappearing by the end of the 21st century. One challenging aspect of documentation is to develop machine learning tools to automate the processing of EL audio via automatic speech recognition (ASR), machine translation (MT), or speech translation (ST). This paper presents an open-access speech translation corpus of Highland Puebla Nahuatl (glottocode high1278), an EL spoken in central Mexico. It then addresses machine learning contributions to endangered language documentation and argues for the importance of speech translation as a key element in the documentation process. In our experiments, we observed that state-of-the-art end-to-end ST models could outperform a cascaded ST (ASR > MT) pipeline when translating endangered language documentation materials.

1 Introduction

Due to the need for global communication, computational technologies such as automatic speech recognition (ASR), machine translation (MT: text-to-text), and speech translation (ST: speech-to-text) have focused their efforts on languages spoken by major population groups (Henrich et al., 2010). Many other languages that are spoken today will probably disappear by the end of the 21st century (Grenoble et al., 2011). For this reason, until very recently they have not been targeted for machine learning technologies. This is changing, however, as increasing attention has been paid to language loss and the need for preservation and, in best-case scenarios, revitalization of these languages.

This paper presents an open-access speech translation corpus from Highland Puebla Nahuatl to Spanish and discusses our initial effort on ST over the corresponding corpus. The following of this paper is organized as follows: in Section 2, we discuss the benefits of speech translation for EL documentation and pioneer-suggest it as the first step in the documentation process. In Section 3, we compare the strategies (i.e., cascaded model and end-to-end models) that can be used to automate ST for ELs. In Section 4 we introduce the Highland Puebla Nahuatl-to-Spanish corpus. Initial experimental efforts in building ST models are elaborated in Section 5. The conclusion is presented in in Section 6.

2 Benefits of speech-to-text translation as a first step in language documentation

The present article suggests that speech translation (ST) could be a viable and valuable tool for EL documentation efforts for three reasons (Anastasopoulos, 2019). First, the transcription of native language recordings may become particularly problematic and time-consuming (the “transcription bottleneck") when the remaining speakers are elderly, and the younger generation has at best a passive knowledge of the language, a common situation of ELs. Second, in many cases ST may be more accurate than MT for target language translation. Finally, many EL documentation projects suffer from a lack of human resources with the skills and time to transcribe and analyze recordings (for similar points about a "translation before transcription workflow", see Bird, 2020, section 2.2.2).

By beginning with ST, semi- and passive speakers can better contribute to EL documentation of their native languages with a level of effort far lower than needed for transcription and analysis. Bilingual native speakers or researchers with incomplete knowledge of the source language structure can quickly produce highly informative free translations even if the original text is never, or only much later, segmented and glossed. A free translation in audio and subsequent capture by typing or using ASR systems for the major target L2 lan-

Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas, pages 53–63

guage (that are more accurate for major as opposed to minor and endangered languages) may take 4–5 hours of effort per hour of audio, whereas transcription (without analysis) may take 30–100 hours for the same unit. Starting with free translation, then, increases the pool of potential native speaker participants and quickly adds value to an audio corpus that may languish if the first step is always fixed as transcription and segmentation (morphological parsing and glossing).

In general, EL documentation proceeds in a fairly set sequence: (1) record; (2) transcribe in time-coded format; (3a) analyze by parsing (morphological segmentation) and glossing; and (3b) freely translate into a dominant, often colonial, language. It may be that some projects prioritize free translation (3b) over morphological segmentation and glossing. Given that each procedure adds a certain, often significant, amount of time to the processing pipeline, there is an increasing scarcity of resources as one proceeds from (1) to (3a/b). If the standard sequence is followed, there are invariably more recordings than transcriptions, more transcriptions than analyses, and (if the sequence is 3a > 3b) more analyses than free translations or (if the sequence is 3b > 3a) more free translations than analyses (see Bird, 2020, Table 3, p. 720).

The argument presented here is that the easiest data to obtain are the recordings followed by free translations into a major language. It may be beneficial to reorder the workflow so that an ST corpus, i.e., free translation of the recording, is prioritized. Only later would transcription and analysis (morphological segmentation and glossing) be inserted into the pipeline. To facilitate computational support for speech-to-text production, we would recommend a targeted number of recordings (e.g., 50 hours), followed by division into utterances with time stamps and free translation of the utterances into a major language. This corpus (or perhaps one even larger) would be used to train an end-to-end neural network in speech-to-text production. The trained ST system would then be used to process additional recordings, thus generating a very extensive freely translated corpus. Our hope would be that instead of basing ASR on an acoustic signal alone, using two coupled inputs—the speech signal and the free translation—might well lower ASR error rates from those obtained from the speech signal alone. The extent of improved accuracy is at this point simply a hypothesis. It would have to be empirically researched, something we hope to do in the near future (see Anastasopoulos, 2019, chap. 4). In this scenario for EL documentation, transcription and analysis proceed forward, but only after an extensive ST training/validation/test corpus has been developed. The resultant ST system would then be used to freely translate additional recordings as they are made.

Speech translation (ST) is very challenging, particularly for resource-scarce endangered languages. The degree of challenge might well be reduced if corpus creation focused from the beginning on translation without intermediate steps (transcription and analysis, which would take documentation in the direction of MT). Moreover, translation itself is a challenging art complicated by the lexical and morphosyntactic intricacies of languages and, more often than not, the discrepancies in vision and structure between source and target language (cf. Sapir, 1921, chap. 5). Extremely large corpora might smooth out the edges, but if free translations are created only after transcription, then the "transcription bottleneck" will also limit the availability of free translations. Limited EL free translation resources, in turn, creates the danger that idiosyncratic or literal translations might dominate the training set. This is another reason to position free translation directly from a recording *before* transcription and analysis.

Free translation and textual meaning: Even when a transcription has been produced and then morphologically segmented and glossed, free translations are beneficial, either generated from the transcription or directly from the speech signal. For example, although multiple sense glossing (i.e., choosing from multiple senses or functions in glossing a morpheme) clarifies ambiguous meanings, it is time-consuming for a human and challenging to automate. The semantic ambiguity of single morphemes will be mitigated if not resolved, however, if accompanied by free translations. Note the following interlinearization, in which, in isolation, the meaning of the gloss line is confusing. The free translations clarifies the meaning and offers a secondary sense to the verb root *koto:ni*.

Ko:koto:nis a:t komo a:mo kiowis.
0-ko:-koto:ni-s a:-t komo a:mo kiowi-s
3sgS-rdpl-to.snap-irreal.sg water-abs if not rain-irreal.sg

The stream will dry up into little ponds if it doesn't rain.

Note also that multi-word lemmas and idiomatic expressions are in many cases opaque in word-by-word (or, even more challenging, morpheme-by-morpheme) glossing. Again a gloss and parallel free translation preserve literal meaning while clarifying the actual meaning to target language speakers.

3 Strategies for automate speech-to-text translation: Cascaded model vs. end-to-end model

One intuitive solution to automating free translation is the cascaded model. But this is difficult to implement since it relies on a pipeline from automatic speech recognition (ASR) to machine translation (MT). Most ELs, however, lack the material and data necessary to robustly train both ASR and MT systems (Do et al., 2014; Matsuura et al., 2020; Shi et al., 2021).

End-to-end ST has received much attention from the NLP research community because of its simpler implementation and computational efficiency (Bérard et al., 2016; Weiss et al., 2017; Inaguma et al., 2019; Wu et al., 2020). In addition, it can also avoid propagating errors from ASR components by directly processing the speech signal. However, as with ASR and MT, ST also often suffers from limited training data and resultant difficulties in training a robust system, which makes the task challenging. There are few available examples of ST applied to endangered languages.

Indeed, most speech translation efforts are between major languages (Di Gangi et al., 2019a; Cattoni et al., 2021; Kocabiyikoglu et al., 2018; Salesky et al., 2021). In these corpora, both source and target languages usually have a standardized writing system and ample training data, a situation generally absent for ELs. A well-known low-resource ST corpus is the Mboshi-French corpus (Godard et al., 2018). However, it is based on the reading of written texts, which does not present the difficulties encountered in conversational speech scenarios. In EL documentation projects, it is these latter scenarios that are most common.

4 Corpus Description

4.1 Characteristics of Highland Puebla Nahuatl (glottocode high1278)

In this paper, we release a Highland Puebla Nahuatl (HPN; glottocode high1278) speech translation corpus for EL documentation. The corpus is governed by a Creative Commons BY-NC-SA 3.0 license and can be downloaded from `http://www.openslr.org/92`. We have analyzed the corpus and explored different ST models and corresponding open-source training recipes in ESPNet (Watanabe et al., 2018).

Nahuatl languages are polysynthetic, agglutinative, head-marking languages with relatively productive derivational morphology, reduplication, and noun incorporation. A rich set of affixes creates the basis for a high number of potential words from any given lemma. As illustrated in Table 1, a transitive verb may contain half a dozen affixes; up to eight in a single word is not uncommon. Suffixes (not represented in Table 1) include tense/aspect/mood markings as well as "associated motion" (*ti-cho:ka-ti-nemi-ya-h* 1plS-cry-ligature-walk-imperf-pl 'we used to go around crying' and directionals (*ti-mits-ih-ita-to-h* 1plS-2sgO-rdpl-see-extraverse.dir-pl 'we went to visit you').

Noun incorporation is not reflected in Table 1 as verbs with incorporated nouns may be treated as lexicalized stems with a compound internal structure. The function of the nominal stem can be highly varied (Tuggy, 1986) as it may lower valency (object incorporation) or leave valency unaffected, as with subject incorporation (not common), as well as both possessor raising (*ni-kone:-miki-k* 1sgS-child-die-perfective.sg 'My child died on me') and modification (*ni-kone:-tsahtsi-0* 1sgS-child-shout-pres.sg 'I shout like a child'). Though noun incorporation is not fully productive (Mithun, 1984), it does increase the number of lemmas. It complicates patterns and meaning of reduplication, which may be at the left edge of the compound (transitive *ma:teki* > *ma:ma:teki* 'to cut repeatedly on the arm') or stem internal (e.g., *ma:tehteki* 'to harvest by hand'). It also complicates automatic translation, particularly in the case of out of vocabulary compounds in which there is no precedent for any of the possible interpretations of the incorporated noun stem.

The main challenge to developing machine translation algorithms for HPN is its morphological complexity, large numbers of words with a low

A	B	C	D	E		F	G	H
subj.	referential obj.	directional prefix	reflexive	non-referential obj. +human	non-referential obj. -human	adverbials (na:l-, ye:k-)	redupli-cation	verb stem

Table 1: Transitional verb morphology: General overview of prefixation

Language	#Tokens	#Types	Ratio (Tokens/Type)	% Corpus in top 100 types
HPN	476,108	96,890	11.39	58.9
Yoloxóchitl Mixtec	955,602	26,445	36.14	59.0
English	783,555	9,601	81.61	63.0

Table 2: Comparative impact of morphological complexity on type-to-token ratios (the English statistics are from DARPA Transtac; the Mixtec statistics are from corpus presented in (Amith and García, 2020; Shi et al., 2021))

token-to-type ratio, and significant occurrences of both noun incorporation and reduplication accompanied by considerable variation in the semantic implications of incorporated noun stems and reduplicants. Table 2 lists type/token ratios in sample texts for three languages, including HPN. While the most frequent 100-word types cover roughly the same portion of text in all three languages, the remaining word types are represented in much lower frequency in HPN than in Yoloxóchitl Mixtec (glottologyolo1241, another EL spoken in Mexico) or English. As a corollary, this means that the remaining 41.1% of tokens (195,680) in the HPN corpus represents 41,718 types, a type-to-token ratio of 1:4.7. The equivalent ratio for English is 1:30.5.

Finally, HPN word order is relatively flexible, which may pose an additional challenge to free translation as neither case marking or word order unambiguously serves to indicate grammatical function. The degree to which MT or ST can handle this relative variability in word order, even with relatively abundant resources, It is not clear.

4.2 Corpus Transcription

Recording: The HPN corpus was developed with speakers from the municipality of Cuetzalan del Progreso, in the northeastern sierra of the state. Most speakers were from San Miguel Tzinacapan and neighboring communities. Recordings use a 48 kHz sampling rate at 16-bits. To facilitate transcription of overlapping speech, each speaker was miked separately into one of two channels with a head-worn Shure SM-10a dynamic mic. A total of 954 recordings were made in a variety of genres. The principal topic, with 591 separate conversations, was plant nomenclature, classification, and use.

Transcription: The workflow commenced with recording sessions in relatively isolated environments. The original transcription was done in Transcriber (Barras et al., 2001) by one of four native speaker members of the research team: Amelia Domínguez Alcántara, Hermelindo Salazar Osollo, Ceferino Salgado Castañeda, and Eleuterio Gorostiza Salazar. Amith then reviewed each transcription, checking any doubts with a native speaker, before importing the finalized Transcriber file into ELAN (Wittenburg et al., 2006). In import, each speaker was assigned a separate tier, and then an additional dependent tier for the free translation was created for each speaker.

Spanish influence: Endangered languages are often spoken in a (neo-)colonial context in which the impact of a dominant language (often but not always non-Indigenous) is felt in many spheres (McConvell and Meakins, 2005). HPN, particularly from the municipality of Cuetzalan, is striking for manifesting two perhaps contrary tendencies: (1) a puristic ideology that has motivated the creation of many neologisms along with (2) morphosyntactic shift under the subtle and covert influence of Spanish.[1] It is probably the case that neither neologisms nor morphosyntactic change poses much of a problem for machine translation; Spanish loans and code-switching into Spanish would undoubtedly be even less problematic. Indeed, it may well be that Spanish impact in many domains of HPN poses minimal problems for machine translation, particularly if the translation is text-to-text. One potential area of difficulty would be in speech translation, in which the Spanish translation is produced directly from a Nahuatl recording. In the conventions

[1]Details of two patterns are discussed in Appendix A.

for HPN transcription, a Spanish loan with distinct meanings in Spanish vs. Nahuatl contexts is distinguished orthographically. It might be difficult to disambiguate the two if the translation is direct from audio. Thus note the following: *āmo nikmati* como *tikchīwas* ('I don't know *how* you will do it') vs. *āmo nikmati komo tikchīwas* ('I don't know if you will do it'). Spanish *como* ('how') may retain its Spanish meaning in a Nahuatl narrative (in which case it is written as if Spanish), or it may be used as a conditional ('if'), in which case it is conventionally written in Nahuatl orthography (*komo*). Even though the decision to orthographically distinguish [komo] / <como> meaning 'how' from [komo] / <komo> meaning 'if' is a particular feature of HPN transcription conventions, the ambiguity in meaning (i.e., translation) would persist even if the orthographies of the two senses were to be different.

In sum, then, it may be that the Spanish impact on Nahuatl is less problematic for MT than for ASR. The most problematic situation for ST is when a Spanish word is used in a Nahuatl-speaking community with both its original Spanish meaning or an innovative Nahuatl meaning. In this case, working via MT from a written transcription may have an advantage if the orthography used for each different meaning (original Spanish vs. innovated) is represented differently based on orthographic convention (as with *como*). But in other cases of Spanish language impact, it is not clear that the cascaded ST (ASR > MT) pipeline enjoys advantages over the direct end-to-end ST system.

4.3 Standardized Splits

The HPN corpus includes corpora for two tasks: ASR and ST(MT). The statistics and the partition information are shown in Table 3. The ASR corpus contains high-quality speech with phone-level transcription. The ST corpus is a subset of the ASR corpus in that it comprises the subset of the ASR corpus that includes time-aligned free translation of the HPN transcription.

5 Experiments

In this section, we present our initial effort on building an automatic ST model for EL documentation. Following the discussion in Section 3, we compare the cascaded model with end-to-end models. To construct the cascaded model, we first conduct experiments on ASR and MT, respectively. Next, we compare different ST models. All the models are constructed with ESPNet, while all the training recipes are available at the ESPNet GitHub repository.[2]

Corpus	Subset	#Utts	Dur (h)
ASR	Train	96,890	123.67
	Validation	7,742	11.48
	Test	16,348	20.97
ST & MT	Train	30,414	36.17
	Validation	2,181	3.13
	Test	5,386	6.65

Table 3: Corpus partition for HPN-ASR and for HPN-ST/HPN-MT

5.1 Automatic Speech Recognition (ASR)

In many open-data tasks, end-to-end ASR compares favorably to traditional hidden Markov model–based ASR systems. The same trend is also shown in ASR for another endangered language, Yoloxóchitl Mixtec as presented in Shi et al. (2021), Table 2. Following a methodology similar to that used for ASR of Yoloxóchitl Mixtec, we have constructed a baseline system based on end-to-end ASR, specifically the transformer-based encoder-decoder architecture with hybrid CTC/attention loss (Watanabe et al., 2017; Karita et al., 2019). We have employed the exact same network configurations as the ESPNet MuST-C recipe.[3] The target of the system is 150 BPE units trained from the unigram language model. For decoding, we integrate the recurrent neural network language model with the ASR model. Specaugmentation is adopted for data augmentation (Park et al., 2019).

The results in character error rate (CER) and word error rate (WER) are shown in Table 4. The experiments show that ASR improves only slightly as the result of increasing the data size from 45 to 156 hours.

5.2 Machine Translation (MT)

The MT experiments are conducted over the ST corpus with ground truth HPN transcription by native-speaker transcribers. We also adopt ESPNet to train the MT model with encoder-decoder architecture (Inaguma et al., 2020). The settings

[2] https://github.com/espnet/espnet/tree/master/egs/puebla_nahuatl

[3] https://github.com/espnet/espnet/tree/master/egs/must_c/asr1

	% CER		% WER	
Corpus	dev	test	dev	test
ASR(156h)	8.8	8.5	23.9	22.4
ST (45h)	9.9	11.2	23.7	25.5

Table 4: ASR results for the HPN-ASR(156h) and HPN-ST(45h) corpora. ASR is directly used for cascaded model and applied for pre-training for end-to-end ST

Model	Val.	Test
MT	14.81	14.10
Cascaded-ST (ASR > MT)	14.72	13.26
E2E-ST w/ ASR-MTL	9.84	9.38
E2E-ST w/ ASR-SI	**15.22**	**15.41**

Table 5: MT and ST BLEU on different models: MTL is the system with multi-task learning; SI is the system with searchable intermediates.

exactly follow the settings for the ESPNet Must-C recipe.[4] The MT result on validation and test sets is shown in Table 5. As discussed in Section 3, the recordings are all of the conversational speech. For text-to-text machine translation the Nahuat inputs are native speaker transcriptions. For the cascading ST model, the Nahuat inputs are outputs from ASR, which have in built-in error rate. Due to the factor, the ASR transcription as a source text may not be an ideal candidate for cascaded ST translation, as it introduces additional noise from conversational transcription.

[4] https://github.com/espnet/espnet/tree/master/egs/must_c/asr1

Model	Val.	Test
E2E-ST w/ ASR-MTL	9.84	9.38
+ ASR encoder init.	14.77	14.05
+ MT decoder init.	11.06	11.03
+ ASR & MT init.	**15.08**	**14.24**

Table 6: Mitigating low resource ST by initializing encoders and decoders with pre-trained models. The ASR model is pre-trained using the 123.67 hours of HPN-ASR corpus, and the MT model is trained on the 30,414 text utterances from the HPN-ST corpus.

5.3 Speech Translation (ST)

While the traditional cascading approach to automating free translations (using two models, ASR and MT) shows strong results on many datasets, recent works have also shown competitive results using end-to-end systems that directly output translations from speech using a single model (Jan et al., 2019; Sperber and Paulik, 2020; Ansari et al., 2020). For low-resource settings, in particular, the data efficiencies of different methodologies become key performance factors (Bansal et al., 2018; Sperber et al., 2019). In this paper, we compare the performance of our dataset of both cascaded and single ST end-to-end systems. Both our cascaded and end-to-end systems are based on the encoder-decoder architecture (Bérard et al., 2016; Weiss et al., 2017) and the transformer-based model (Di Gangi et al., 2019b; Inaguma et al., 2019).

(a) Cascaded ST Model (ASR > MT Pipeline): The cascaded model consists of an ASR module and an MT module, each optimized separately during training. Each module is pre-trained with the same method as presented in Sections 5.1 and 5.2. During inference, the 1-best hypothesis from the ASR module is obtained via beam search with a beam size of 10, and this decoded transcription is passed to the subsequent MT module that finally outputs translated text. Results are shown in Table 5.

(b) End-to-end ST Model: In our experiments, we adopt the transformer-based encoder-decoder architecture with Specaugmentation. In addition, we default train the current system with the combination of ASR CTC-based loss from the encoder and ST translation loss from the decoder; this is referred to as E2E-ST with ASR-MTL. We also evaluate the Searchable Intermediates (SI) based ST model (E2E-ST with ASR-SI) introduced in Dalmia et al. (2021), where the ASR intermediates are found using the same decoding parameters as the ASR models of the cascade model. The detailed hyper-parameters follow the configuration of the ESPNet Must-C recipes.[5]

ST results are shown in Table 5. While the performance of the Cascaded-ST system is close to that of the MT system, the E2E-ST with ASR-MTL system shows a significantly worse result. Since E2E-ST with ASR-MTL jointly optimizes a speech

[5] https://github.com/espnet/espnet/tree/master/egs/must_c/asr1

encoder with an ASR decoder that is not included in the final inference network, this subnet waste is likely causing data inefficiency that is evident in our low-resource dataset (Sperber et al., 2019). In contrast, E2E-ST with SI actually outperforms both the MT and cascaded-ST systems, suggesting that it is less degraded by the low-resource constraint (Anastasopoulos and Chiang, 2018; Wang et al., 2020; Dalmia et al., 2021). Furthermore, this result shows that Nahuatl is more easily translated with a methodology that can consider both speech and transcript sequences as inputs.

(c) Pre-training for end-to-end ST: To investigate the pre-training effect for HPN, we adopt the models trained from Sections 5.1 and 5.2. The ASR model in Section 5.1 was used for initialization of the ST encoder, while the MT model in Section 5.2 was used for initialization of the ST decoder.

As shown in Table 6, the best performance is reached with initialization from both ASR encoder and MT decoder. Pre-training encoder and decoder could help better ST modeling, while using the pre-trained ASR encoder could contribute to more performance improvements.

Some examples with the best model in Table 6 are shown in Appendix B. Based on the analysis, it generally indicates that the current ST system can translate some essential information into Spanish. However, it still cannot fully replace the human effort on the task. And the translation still needs significant correction from a human annotator.

6 Conclusions

In this paper, we release the Highland Puebla Nahuatl corpus for ASR, MT, and ST tasks. The corpus, related baseline models, and training recipes are open source under the CC BY-NC-ND 3.0 license. We expect the corpus to facilitate all three tasks for EL documentation. We also discuss and present three specific reasons for prioritizing ST as an initial step in the endangered language documentation sequence after the recording has taken place. Finally, we explore different technologies for ST of Highland Puebla Nahuatl and compare these to results obtained by processing through the cascaded ST pipeline.

As discussed in Section 2, we suggest that prioritizing free translation as a first, not final, step in documentation should be considered as: (1) it can rapidly make a corpus valuable to potential users even if transcription, morphlogical segmentation, and morpheme glossing is incomplete; (2) it enables semi-, passive and heritage speakers to participate in documentation of their languages; (3) it provides an alternative process for ASR in which the ASR target is not a transcription but a translation into a Western language; and (4) it creates a scenario in which the acoustic signal and free translation may be coupled as inputs into an end-to-end ASR system. Therefore, our future works will focus on how the human effort could be reduced via ST models and on how to incorporate ST to improve the ASR performances.

Acknowledgements

The authors gratefully acknowledge the following support for documenting and studying Highland Puebla Nahuat: National Science Foundation, Documenting Endangered Languages (DEL): Awards 1401178, 0756536 (Amith, PI on both awards); National Endowment for the Humanities, Preservation and Access: PD-50031-14 (Amith, PI); Endangered Language Documentation Programme: Award MDP0272 (Amith, PI). The native speaker documentation team responsible for transcription and translation included Amelia Domínguez Alcántara. Ceferino Salgado Castañeda, Hermelindo Salazar Osollo, and Eleuterio Gorostiza Salazar. Yoloxóchitl Mixtec documentation was supported by the following grants: NSF-DEL: Awards 1761421, 1500595, 0966462 (Amith, PI on all three awards; the second was a collaborative project with SRI International, Award 1500738, Andreas Kathol, PI); Endangered Language Documentation Programme: Awards MDP0201, PPG0048 (Amith, PI on both awards). Rey Castillo García has been responsible for all transcriptions.

References

Jonathan D. Amith and Rey Castillo García. 2020. Audio corpus of Yoloxóchitl Mixtec with accompanying time-coded transcriptons in ELAN. http://www.openslr.org/89/. Accessed: 2021-03-05.

Antonios Anastasopoulos. 2019. *Computational tools for endangered language documentation*. Ph.D. thesis, University of Notre Dame, Computer Science and Engineering.

Antonios Anastasopoulos and David Chiang. 2018. Tied multitask learning for neural speech 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 82–91, New Orleans, Louisiana. Association for Computational Linguistics.

Ebrahim Ansari, Nguyen Bach, Ondřej Bojar, Roldano Cattoni, Fahim Dalvi, Nadir Durrani, Marcello Federico, Christian Federmann, Jiatao Gu, Fei Huang, et al. 2020. Findings of the IWSLT 2020 evaluation campaign. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 1–34.

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. 2018. Low-resource speech-to-text translation. *Computing Research Repository (CoRR)*, abs/1803.09164.

Claude Barras, Edouard Geoffrois, Zhibiao Wu, and Mark Liberman. 2001. Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communication*, 33(1-2):5–22.

Alexandre Bérard, Olivier Pietquin, Laurent Besacier, and Christophe Servan. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. In *NIPS Workshop on end-to-end learning for speech and audio processing*.

Steven Bird. 2020. Sparse transcription. *Computational Linguistics*, 46(4):713–744.

Roldano Cattoni, Mattia Antonino Di Gangi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. MuST-C: A multilingual corpus for end-to-end speech translation. *Computer Speech & Language*, 66:101–155.

Siddharth Dalmia, Brian Yan, Vikas Raunak, Florian Metze, and Shinji Watanabe. 2021. Searchable hidden intermediates for end-to-end models of decomposable sequence tasks. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019a. 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.

Mattia A Di Gangi, Matteo Negri, and Marco Turchi. 2019b. Adapting transformer to end-to-end spoken language translation. *Proc. Interspeech 2019*, pages 1133–1137.

Thi-Ngoc-Diep Do, Alexis Michaud, and Eric Castelli. 2014. Towards the automatic processing of Yongning Na (Sino-Tibetan): Developing a 'light' acoustic model of the target language and testing 'heavyweight' models from five national languages. In *Spoken Language Technologies for Under-Resourced Languages*.

P Godard, G Adda, Martine Adda-Decker, J Benjumea, Laurent Besacier, J Cooper-Leavitt, GN Kouarata, L Lamel, H Maynard, M Müller, et al. 2018. A very low resource language speech corpus for computational language documentation experiments. In *Language Resources and Evaluation Conference (LREC)*.

LA Grenoble, Peter K Austin, and Julia Sallabank. 2011. *Handbook of Endangered Languages*. Cambridge University Press.

Joseph Henrich, Steven J Heine, and Ara Norenzayan. 2010. Most people are not weird. *Nature*, 466(7302):29–29.

Hirofumi Inaguma, Kevin Duh, Tatsuya Kawahara, and Shinji Watanabe. 2019. Multilingual end-to-end speech translation. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 570–577. IEEE.

Hirofumi Inaguma, Shun Kiyono, Kevin Duh, Shigeki Karita, Nelson Yalta, Tomoki Hayashi, and Shinji Watanabe. 2020. ESPNet-ST: All-in-one speech translation toolkit. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 302–311.

Niehues Jan, Roldano Cattoni, Stuker Sebastian, Matteo Negri, Marco Turchi, Salesky Elizabeth, Sanabria Ramon, Barrault Loic, Specia Lucia, and Marcello Federico. 2019. The IWSLT 2019 evaluation campaign. In *16th International Workshop on Spoken Language Translation 2019*.

Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson Enrique Yalta Soplin, Ryuichi Yamamoto, Xiaofei Wang, et al. 2019. A comparative study on transformer vs RNN in speech applications. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 449–456. IEEE.

Ali Can Kocabiyikoglu, Laurent Besacier, and Olivier Kraif. 2018. Augmenting LibriSpeech with French translations: A multimodal corpus for direct speech translation evaluation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Kohei Matsuura, Sei Ueno, Masato Mimura, Shinsuke Sakai, and Tatsuya Kawahara. 2020. Speech corpus of Ainu folklore and end-to-end speech recognition for the Ainu language. In *Proceedings of*

The 12th Language Resources and Evaluation Conference, pages 2622–2628.

Patrick McConvell and Felicity Meakins. 2005. Gurindji Kriol: A mixed language emerges from code-switching. *Australian Journal of Linguistics*, 25(1):9–30.

Marianne Mithun. 1984. The evolution of noun incorporation. *Language*, 60(4):847–894.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. *Proc. Interspeech 2019*, pages 2613–2617.

Elizabeth Salesky, Matthew Wiesner, Jacob Bremerman, Roldano Cattoni, Matteo Negri, Marco Turchi, Douglas W Oard, and Matt Post. 2021. The multilingual TEDx corpus for speech recognition and translation. *arXiv preprint arXiv:2102.01757*.

Edward Sapir. 1921. *An Introduction to the Study of Speech*. Harcourt, Brace & World, New York:.

Jiatong Shi, D Amith, Jonathan, Rey Castillo García, Esteban Guadalupe Sierra, Kevin Duh, and Shinji Watanabe. 2021. Leveraging end-to-end asr for endangered language documentation: An empirical study on Yoloxóchitl Mixtec. *arXiv preprint arXiv:2101.10877*.

Matthias Sperber, Graham Neubig, Jan Niehues, and Alex Waibel. 2019. Attention-passing models for robust and data-efficient end-to-end speech translation. *Transactions of the Association for Computational Linguistics*, 7:313–325.

Matthias Sperber and Matthias Paulik. 2020. Speech translation and the end-to-end promise: Taking stock of where we are. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

David Tuggy. 1986. Noun incorporations in nahuatl. In *Proceedings of the Annual Meeting of the Pacific Linguistics Conference*, volume 2, pages 455–470.

Chengyi Wang, Yu Wu, Shujie Liu, Zhenglu Yang, and Ming Zhou. 2020. Bridging the gap between pretraining and fine-tuning for end-to-end speech translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9161–9168.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson-Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, et al. 2018. ESPnet: End-to-end speech processing toolkit. *Proc. Interspeech 2018*, pages 2207–2211.

Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R Hershey, and Tomoki Hayashi. 2017. Hybrid CTC/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1240–1253.

Ron J Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. 2017. Sequence-to-sequence models can directly translate foreign speech. *Proc. Interspeech 2017*, pages 2625–2629.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: A professional framework for multimodality research. In *5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1556–1559.

Anne Wu, Changhan Wang, Juan Pino, and Jiatao Gu. 2020. Self-supervised representations improve end-to-end speech translation. *Proc. Interspeech 2020*, pages 1491–1495.

A Spanish language impact on Highland Puebla Nahuat

HPN, particularly from the municipality of Cuetzalan, is striking for manifesting two seemingly contrary tendencies: neologisms and morphosyntactic. The first is a puristic ideology that values the native language as an expression of Indigenous identity. The second is a very strong influence of Spanish syntax that has led to a significant number of calques that are not only direct translations of Spanish, but that yield expressions that violate basic grammatical constraints of Nahuatl. Puristic ideology motivates many neologisms, many of which are nouns, that provide an alternative to Spanish loans. Spanish impact on morphosyntax is also prevalent. For example, with very few exceptions, the valency of Nahuatl verbs is fixed as either intransitive, transitive, or ditransitive. Thus to accept an object, an intransitive must undergo valency increase through an overt morphological process. But Spanish influence has created situations in which intransitive Nahuat verbs mark two arguments (subject and object) on the erstwhile intransitive stem. Under Spanish influence, the intransitive verbs *kīsa* 'to emerge' (Spanish 'salir')' and *tikwi* 'to light up' (Spanish 'prenderse') manifest otherwise ungrammatical forms: (a) *āmo nēchkīsa* (*Ø-nēch-kīsa-Ø*; 3sgS-1sgO-to.emerge-pres.sg) is a calque from Spanish 'no me sale' ('it doesn't turn out right for me'); (b) *motikwi* (*Ø-mo-tikwi-Ø*; 'it lights up') uses an unnecessary and ungrammatical reflexive marker influenced by the reflexive Spanish term 'se prende'.

B Speech translation examples

This appendix shows five examples of our ST hypothesis (i.e., HYP) with speech transcription (i.e., HPN) and Spanish translation reference (i.e., REF). We indicate the corresponding utterance IDs in the parenthesis of each example.[6]

EG1 (AND308_2011-07-13-p-R_000105):
HPN: wa:n se: taman de n' kwaltia yo:n ista:k xo:no:t yo:n se: kehke:wa wa:n k..., se: kitsa:la:nwia
REF: y el segundo es el ista:k xo:no:t que se se puede usar este se le extrae la corteza se mete en un tronco partido a la mitad se jala con fuerza para separar la corteza con la fibra que es el jonote
HYP: y el otro tipo de árbol que se le puede extraer la fibra del jonote blanco se desprende la corteza y se jala con fuerza para separar la fibra
Notes: This is part of a recording about a tree called "jonote", which covers several species of *Heliocarpus*. The bark of one of these trees, the white jonote, is particularly good: It is stripped from the tree and then pulled through a wedge cut into the trunk so that the meaty part of the bark is stripped away and the fiber is left somewhat clean. The E2E translation captures the better quality of the white jonote and presents the process of stripping fiber. Note that the gist of the Nahuatl is accurately captured, from the fact that the narrator is now commenting on a second type of jonote (white) that is the most used, to the mechanism for cleaning the stripped bark. The word *kitsa:la:nwia* ('to put in the middle') is also correctly translated as the action of pulling the bark through a wedge to clean the fiber.

EG2 (AND308_2011-07-13-p-R_000133):
HPN: pero a:mo nochi kokotoka n' chi:chi:lxo:no:t
REF: pero no todo se revienta el chi:chi:lxo:no:t
HYP: pero no todos se revientan los frutos de chi:chi:lxo:no:t
Notes: Here the reference is to the red jonote, *chi:chi:lxo:no:t*, which is less commonly used than the white jonote. This is because the bark of the red jonote is more brittle and tends to snap into pieces as it is pulled off the trunk. Here, however, the narrator is commenting that not all red jonote trees have bark that pulls apart into pieces. The Nahuatl text, however, makes no mention of "fruit" ('fruto') as the computer generated translation suggests.

EG3 (AND308_2011-07-13-p-R_000186):
HPN: ke:mah wa:n no: kwali tine:chili:s compadrito de que
REF: sí y también puedes platicarme compadrito de que
HYP: sí y también puedes decirme compadrito que
Notes: The alternation between 'platicarme' and 'decirme' is basically a case of synonymy. The computer translation is accurate.

EG4 (AND308_2011-07-29-i-R_000077):
HPN: nextik ihkó:n pero yo:n a:mo
REF: es de color gris pero ésa planta no
HYP: es gris pero ese no
Notes: Here the computer translation is closer to the Nahuatl and fully understandable. The human translator inserted "color", which was not necessary to the understanding of 'grey' as a color.. The term *yo:n* is a pronoun, "that". The human translator added the anteceding referent, "flower", to make the text clearer. In English the difference would be: (1) it is a grey color, but that plant isn't" (Human) vs. (2) it is grey but that one isn't" (Computer). The computer translation is in fact closer to the original.

[6]The utterance IDs can be obtained from corresponding recipes at ESPNet

EG5 (RMM302_2011-07-27-l-L_000106):
HPN: oksé: no: n'neki nimitstahtani:s n' tehwa:tsi:n komo yehwa n' tein neh niki:mxattok ¿ke:niwki n' ixo:chiotsi:n
REF: otra coso quiero preguntarte si acaso sea el mismo que yo conozco ¿cómo es la flor
HYP: otra cosa que quiero preguntarte si es que yo conozco como es su flor
Notes: In the Nahuatl text the narrator is basically saying that he has another question to ask ('otra cosa que quiero preguntarte') and that this question is whether 'the flower that I know ..., what is its flower like?' (¿cómo es su flor?).