End-to-End Automatic Speech Recognition: Its Impact on the Workflow for Documenting Yoloxóchitl Mixtec

Jonathan D. Amith[1] **Jiatong Shi**[2] **Rey Castillo García**[3]
[1]Department of Anthropology, Gettysburg College, Gettysburg, Pennsylvania
[2]The Johns Hopkins University, Baltimore, Maryland
[3]Secretaría de Educación Pública, Estado de Guerrero, Mexico
(jonamith@gmail.com, jiatong_shi@jhu.edu, reyyoloxochitl@gmail.com)

Abstract

This paper describes three open access Yoloxóchitl Mixtec corpora and presents the results and implications of end-to-end automatic speech recognition for endangered language documentation. Two issues are addressed. First, the advantage for ASR accuracy of targeting informational (BPE) units in addition to, or in substitution of, linguistic units (word, morpheme, morae) and then using ROVER for system combination. BPE units consistently outperform linguistic units although the best results are obtained by system combination of different BPE targets. Second, a case is made that for endangered language documentation, ASR contributions should be evaluated according to extrinsic criteria (e.g., positive impact on downstream tasks) and not simply intrinsic metrics (e.g., CER and WER). The extrinsic metric chosen is the level of reduction in the human effort needed to produce high-quality transcriptions for permanent archiving.

1 Introduction: Endangered language documentation history and context

Endangered language (EL) documentation emerged as a field of linguistic activity in the 1990s, as reflected in several seminal moments. In 1991 the Linguistic Society of America held a symposium entitled “Endangered Languages and their Preservation"; in 1992 Hale et al. (1992) published a seminal article on endangered languages in *Language*, the LSA’s flagship journal. In 1998, Himmelmann (1998) argued for the development of documentary linguistics as an endeavor separate from and complementary to descriptive linguistics. By the early years of the present millennium, infrastructure efforts were being developed: metadata standards and best practices for archiving (Bird and Simons, 2003); tools for lexicography and corpus developments such as Shoebox, Transcriber (Barras et al., 1998), and ELAN (Wittenburg et al., 2006), and financial support for endangered language documentation (the Volkswagen Foundation, the NSF Documenting Endangered Language Program, and the SOAS Endangered Language Documentation Programme). Recent retrospectives on the impact of Hale et al. (1992) and Himmelmann (1998) have been published by Seifart et al. (2018) and McDonnell et al. (2018). Within the last decade, the National Science Foundation supported a series of three workshops, under the acronym AARDVARC (Automatically Annotated Repository of Digital Audio and Video Resources Community) to bring together field linguists working on endangered languages and computational linguists working on automatic annotation—particularly automatic speech recognition (ASR)—to address the impact of what has been called the "transcription bottleneck" (Whalen and Damir, 2012). Interest in applying machine learning to endangered language documentation is also manifested in four biennial workshops on this topic, the first in 2014 (Good et al., 2021). Finally, articles directly referencing ASR of *endangered languages* have become increasingly common over the last five years (Adams et al., 2018, 2020; Ćavar et al., 2016; Foley et al., 2018, 2019; Gupta and Boulianne, 2020; Jimerson and Prud’hommeaux, 2018; Jimerson et al., 2018; Michaud et al., 2018; Mitra et al., 2016; Shi et al., 2021).

This article continues work on Yoloxóchitl Mixtec ASR (Mitra et al., 2016; Shi et al., 2021). The most recent efforts (2020 and 2021) have adopted the ESPNet toolkit for end-to-end automatic speech recognition (E2E ASR). This approach has proven to be very efficient in terms of time needed to develop the ASR recipe (Shi et al., 2021) and in yielding ASR hypotheses of an accuracy capable of significantly reducing the extent of human effort needed to finalize accurate transcribed audio for permanent archiving as here demonstrated. Section 2 discusses the Yoloxóchitl Mixtec corpora,

Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas, pages 64–80

and Section 3 explores the general goals of EL documentation. Section 4 reviews the E2E ASR and corresponding results using ESPNet. The conclusion is offered in Section 5.

2 Yoloxóchitl Mixtec: Corpus characteristics and development

2.1 The language

Much work on computer-assisted EL documentation is closely related to work on low-resource languages, for the obvious reason that most ELs have limited resources, be they time-coded transcriptions, interlinearized texts, or corpora in parallel translation. The resources for Yoloxóchitl Mixtec, the language targeted in this present study, are, however, relatively abundant by EL standards (119.32 hours over three corpora), the result of over a decade of linguistic and anthropological research by Amith and Castillo García (2020).

Yoloxóchitl Mixtec (henceforth YM), an endangered Mixtecan language spoken in the municipality of San Luis Acatlán, Guerrero, Mexico, is one of some 50 languages in the Mixtec language family, which is within a larger unit, Otomanguean, that Suárez (1983) considers a hyper-family or stock. Mixtec languages (spoken in Oaxaca, Guerrero, and Puebla) are highly varied, the result of approximately 2,000 years of diversification. YM is spoken in four communities: Yoloxóchitl, Cuanacaxtitlan, Arroyo Cumiapa, and Buena Vista. Mutual intelligibility among the four communities is high despite differences in phonology, morphology, and syntax.

All villages have a simple common segmental inventory but apparently significant though still undocumented variation in tonal phonology; only Cuanacaxtitlan manifests tone sandhi. YMC (referring only to the Mixtec of the community of Yoloxóchitl [16.81602, -98.68597]) manifests 28 distinct tonal patterns on 1,451 to-date identified bimoraic lexical stems. The tonal patterns carry a significant functional load regarding the lexicon and inflection (Palancar et al., 2016). For example, 24 distinct tonal patterns on the bimoraic segmental sequence [nama] yield 30 words (including five homophones). The three principal aspectual forms (irrealis, incompletive, and completive) are almost invariably marked by a tonal variation on the first mora of the verbal stem (1 or 3 for the irrealis, 4 for the incompletive, and 13 for the completive; in addition 14 on the initial mora almost always indicates negation of the irrealis[1]). In a not-insignificant number of cases, suppletive stems exist, generally manifesting variation in a stem-initial consonant and often the stem-initial vowel.

The ample tonal inventory of YMC presents obstacles to native speaker literacy and an ASR system learning to convert an acoustic signal to text. It also complicates the construction of a language lexicon for HMM-based systems, a lexicon that is not required in E2E ASR. The phonological and morphological differences between YMC and the Mixtec of the three other YM communities create challenges for transcription and, by extension, for applying YMC ASR to speech recordings from these other villages. To accomplish this, it will be necessary first to learn the phonology and morphology of these variants and then use this as input into a transfer learning scenario. Intralanguage variation among distinct communities (see Hildebrandt et al., 2017b and other articles in Hildebrandt et al., 2017a) is an additional factor that can negatively impact computer-assisted EL documentation efforts in both intra- and intercommunity contexts.

2.2 The three corpora

YMC-Exp: The corpus originally available to develop E2E ASR, here titled YMC-Exp (Expert transcription), comprises 98.99 hours of time-coded transcription divided as follows for initial ASR development: Training: 92.46 hours (52,763 utterances); Validation: 4.01 hours (2,470 utterances); and Test: 2.52 hours (1,577 utterances).

The size of this initial YM corpus (505 files, 32 speakers, 98.99 hours) sets it apart from other ASR initiatives for endangered languages (Adams et al., 2018; Ćavar et al., 2016; Jimerson et al., 2018; Jimerson and Prud'hommeaux, 2018). This ample size has yielded lower character (CER) and word (WER) error rates than would usually occur with truly low-resource EL documentation projects.

Amith and Castillo García recorded the corpus at a 48KHz sampling rate and 16-bits (usually with a Marantz PMD 671 recorder, Shure SM-10a dynamic headset mics, and separate channels for each speaker). The entire corpus was transcribed by Castillo, a native speaker linguist (García, 2007).

YMC-FB: A second YMC corpus (YMC-FB; for 'field botany') was developed during ethno-

[1] Tones are V^1 low to V^4 high, with V^{13} and V^{14} indicating two of several contour tones; see also fn. 2.

botanical fieldwork. Kenia Velasco Gutiérrez (a Spanish-speaking botanist) and Esteban Guadalupe Sierra (a native speaker from Yoloxóchitl) led 105 days of fieldwork that yielded 888 distinct plant collections. A total of 584 recordings were made in all four YM communities; only 452 were in Yoloxóchitl, and of these, 435, totaling 15.17 hours with only three speakers, were used as a second test case for E2E ASR. Recordings were done outdoors at the plant collection site with a Zoom H4n handheld digital recorder. The Zoom H4n internal mic was used; recordings were 48KHz, 16-bit, a single channel with one speaker talking after another (no overlap). Each recording has a short introduction by Velasco describing, in Spanish, the plant being collected. This Spanish section has not been factored into the duration of the YMC-FB corpus, nor has it been evaluated for character and word error rates at this time (pending future implementation of a multilingual model). The processing of the 435 recordings falls into two groups.

- 257 recordings (8.36 hours) were first transcribed by a novice trainee (Esteban Guadalupe) as part of transcription training. They were corrected in a separate ELAN tier by Castillo García and then the acoustic signals were processed by E2E ASR trained on the YMC-Exp corpus. The ASR CER and WER were obtained by comparing the ASR hypotheses to Castillo's transcriptions; Guadalupe's skill level (also measured in CER and WER) was obtained by comparing his transcription to that of Castillo. The results are discussed in Table 9 of Shi et al. (2021).
- 178 recordings (6.81 hours) were processed by E2E ASR, then corrected by Castillo. This set was not used to teach or evaluate novice trainee transcription skills but only to determine CER and WER for E2E ASR with the YMC-FB corpus.

No training or validation sets were created from this YMC-FB corpus, which for this present paper was used solely to test E2E ASR efficiency using the recipe developed from YMC-Exp corpus. CER and WER scores for YMC-FB were only produced after Castillo used the ELAN interface to correct the ASR hypotheses for this corpus (see Appendix A for an example ASR output).

YMC-VN: The final corpus is a set of 24 narratives made to provide background information and off-camera voice for a documentary video. The recordings involved some speakers not represented in the YMC-Exp corpus. All recordings (5.16 hours) were made at 44.1kHz, 16-bit with a boom-held microphone and a Tascam portable digital recorder in a hotel room. This environment may have introduced reverb or other effects that might have negatively affected ASR CER and WER.

Accessibility: All three corpora (119.32 hours) are available at the OpenSLR data portal (Amith and Castillo García, 2020)

3 Goals and challenges of corpora-based endangered language documentation

3.1 Overview

The oft-cited Boasian trilogy of grammar, dictionaries, and texts is a common foundation for EL documentation. Good (2018, p. 14) parallels this classic conception with a "Himmelmannian" trilogy of recordings, metadata, and annotations (see Himmelmann 2018). For the purpose of the definition proposed here, EL documentation is considered to be based on the Boasian trilogy of (1) corpus, (2) lexicon (in the sense of dictionary), and (3) grammar. In turn, each element in the trilogy is molded by a series of expectations and best practices. An audio corpus, for example, would best be presented interlinearized with (a) lines corresponding to the transcription (often in a practical orthography or IPA transcription), (b) morphological segmentation (often called a 'parse'), (c) parallel glossing of each morpheme, (d) a free translation into a target, often colonial language, and (e) metadata about recording conditions and participants. This is effectively the Himmelmannian trilogy referenced by Good. A dictionary should contain certain minimum fields (e.g., part of speech, etymology, illustrative sentences). Grammatical descriptions (books and articles) are more openly defined (e.g., a reference vs. a pedagogical grammar) and may treat only parts of the language (e.g., verb morphology).

In a best-case scenario, these three elements of the Boasian trilogy are interdependent. Corpus-based lexicography clearly requires ample interlinearized transcriptions (IGT) of natural speech that can be used to (a) develop concordances mapped to lemmas (not word forms); (b) enrich a dictionary by finding lemmas in the corpus that are absent from an extant set of dictionary headwords; and (c) discover patterns in the corpus suggestive of multiword lemmas (e.g., ku^3-na^3a^4 followed by i^3ni^2 (lit., 'darken heart' but meaning 'to faint'). A grammar will inform decisions about morphological segmentation used in the IGT as well as part-of-speech tags and other glosses. And a grammar itself would benefit greatly from a large set of annotated natural speech recordings not simply to provide examples of particular structures but to facilitate a statistical analysis of speech patterns (e.g., for YMC, the relative frequency of completive verbs marked solely by tone vs. those marked by the prefix ni^1-). This integration of elements into one "hypertextual" documentation effort is proposed by Musgrave and Thieberger (2021), who note the importance of spontaneous text (i.e., corpora, which they separate into two elements, media, and text) and comment that "all examples [in the dictionary and grammar] should come from the spontaneous text and should be viewed in context" (p. 6).

Documentation of YMC has proceeded on the assumption that the hypertextual integration suggested by Musgrave and Thieberger is central to effective endangered language documentation based on natural speech and that textual transcription of multimedia recordings of natural speech is, therefore, the foundation for a dictionary and grammar based on actual language use. End-to-end ASR is used to rapidly increase corpus size while offering the opportunity to target certain genres (such as expert conversations on the nomenclature, classification, and use of local flora and fauna; ritual discourse; material cultural production; techniques for fishing and hunting) that are of ethnographic interest but are often insufficiently covered in EL documentation projects that struggle to produce large and varied corpora. With the human effort–reducing advances in ASR for YMC presented in this paper, such extensive targeted recording of endangered cultural knowledge can now easily be included in the documentation effort.

The present paper focuses on end-to-end automatic speech recognition using the ESPNet toolkit (Guo et al., 2020; Shi et al., 2021; Watanabe et al., 2020, 2017, 2018). The basic goal is simple: To develop computational tools that reduce the amount of human effort required to produce accurate transcriptions in time-coded interlinearized format that will serve a wide range of potential stakeholders, from native and heritage speakers to specialized academics in institutions of higher learning, in the present and future generations. The evaluation metric, therefore, is not intrinsic (e.g., reduced CER and WER) but rather extrinsic: the impact of ASR on the downstream task of creating a large and varied corpus of Yoloxóchitl Mixtec.

3.2 Challenges to ASR of endangered languages

ASR for endangered languages is made difficult not simply because of limited resources for training a robust system but by a series of factors briefly discussed in this section.

Recording conditions: Noisy environments, including overlapping speech, reverberation in indoor recordings, natural sounds in outdoor recordings, less than optimal microphone placement (e.g., a boom mic in video recordings), and failure to separately mike speakers for multichannel recordings all negatively impact the accuracy of ASR output. Also to the point, field recordings are seldom made with an eye to seeding a corpus in ways that would specifically benefit ASR results (e.g., recording a large number of speakers for shorter durations, rather than fewer speakers for longer times). To date, then, processing a corpus through ASR techniques of any nature (HMM, end-to-end) has been more of an afterthought than planned at project beginning. Development of a corpus from the beginning with an eye to subsequent ASR potential would be immensely helpful to these computational efforts. It could, perhaps should, be increasingly considered in the initial project design. Indeed, just as funding agencies such as NSF require that projects address data management issues, it might be worth considering the suggested inclusion of how to make documentation materials more amenable to ASR and NLP processing as machine learning technologies are getting more robust.

Colonialization of language: Endangered languages do not die, to paraphrase Dorian (1978), with their "boots on." Rather, in the colonialized situation in which most ELs are immersed, there are multiple phonological, morphological, and syntactic influences from a dominant language. The incidence of a colonial language in native language recordings runs a gamut from multilanguage situations (e.g., each speaker using a distinct language, as often occurs in elicitation sessions: 'How would you translate ___ into Mixtec?'), to code-switching and borrowing or relexification in the speech of single individuals. In some languages (e.g., Nahuatl), a single word may easily combine stems from both native and colonial languages. Preliminary, though not quantified, CER analysis for YMC ASR suggests that “Spanish-origin" words provoke a significantly higher error rate than the YMC lexicon uninfluenced by Spanish. It is also not clear that a multilingual phone recognition system is the solution to character errors (such as ASR hypothesis ‘cereso’ for Spanish ‘cerezo’) that may derive from an orthographic system, such as that for Spanish, that is not designed, as many EL orthographies are, for consistency. Phonological shifts in borrowed terms also preclude the simple application of lexical tools to correct misspellings (as ‘agustu’ for the Spanish month ‘agosto’).

Orthographic conventions: The practical deep orthography developed by Amith and Castillo marks off boundaries of affixes (with a hyphen) and clitics (with an = sign). Tones are indicated by superscript numbers, from 1 low to 4 high, with five common rising and falling tones. Stem-final elided tones are enclosed in parentheses (e.g., underlying form *be'*^{3}*e*$^{(3)}$=^{2}; house=1sgPoss, 'my house'; surface form *be'*^{3}*e*^{2}). Tone-based inflectional morphology is not separated in any YMC transcriptions.[2]

The transcription strategy for YMC was unusual in that the practical orthography was a deep, underlying system that represented segmental morpheme boundaries and showed elided tones in parentheses. The original plans of Amith and Castillo were to use the transcribed audio as primary data for a corpus-based dictionary. A deep orthography facilitates discovery (without recourse to a morphological analyzer) of lemmas that may be altered in surface pronunciations by the effect of person-marking enclitics and certain common verbal prefixes (see Shi et al., 2021, §2.3).

Only after documentation (recording and time-coded transcriptions) was well advanced did work begin on a finite state transducer for the YMC corpus. this was made possible by collaboration with another NSF-DEL sponsored project.[3] The code was written by Jason Lilley in consultation with Amith and Castillo. As the FOMA FST was being built, FST output was repeatedly checked against expectations based on the morphological grammar until no discrepancies were noted. The FST, however, only generates surface forms consistent with Castillo's grammar. If speakers varied, for example, in the extent of vowel harmonization or regressive nasalization, the FST would yield only one surface form, that suggested by Castillo to be the most common. For example, underlying $be'^3e^{(3)}{=}an^4$ (house=3sgFem; ‘her house’) surfaces as $be'^3\tilde{a}^4$ even though for some speakers nasalization spreads to the stem initial vowel. Note, then, that the surface forms in the YMC-Exp corpus are based on FST generation from an underlying transcription as input and not from the direct transcription of the acoustic signal. It is occasionally the case that different speakers might extend vowel harmonization or nasalization leftward to different degrees. This could increase the CER and WER for ASR of surface forms, given that the reference for evaluation is not directly derived from the acoustic signal while the ASR hypothesis is so derived.

In an evaluation across the YMC-Exp development and test sets (total 6.53 hours) of the relative accuracy of ASR when using underlying versus surface orthography, it was found that training on underlying orthography produced slightly greater accuracy than training on surface forms: Underlying = 7.7/16.0 [CER/WER] compared to Surface = 7.8/16.5 [CER/WER] (Shi et al., 2021, see Table 4). The decision to use underlying representations in ASR training has, however, several more important advantages. First, for native speakers, the process of learning a deep practical orthography means that one learns segmental morphology as one learns to write. For the purposes of YMC language documentation, the ability of a neural network to directly learn segmental morphology as part of ASR training has resulted in a YMC ASR output across all three corpora with affixes and clitics separated and stem-final elided tones marked in parentheses. Semi- or un-supervised morphological learning as a separate NLP task is unnecessary when ASR training and testing was successfully carried out on a corpus with basic morphological segmentation. As the example in Appendix A demonstrates, ASR output includes basic segmentation at the morphological level.

[2] For example ka'^3an^4 ‘to have faith (irrealis)’; $ka'^{14}an^4$ ‘to not have faith (neg. irrealis)’, ka'^4an^4 ‘to have faith (incompletive)’; $ka'^{13}an^4$ ‘to have faith (completive). For now, the tonal inflection on the first mora is not parsed out from stems such as ka'^3an^4; see also fn. 1

[3] Award #1360670 (Christian DiCanio, PI; Understanding Prosody and Tone Interactions through Documentation of Two Endangered Languages).

Corpus	Intrinsic CER	Intrinsic WER	Extrinsic Correction Time
Reference	/	/	40 (estimated avg.)
Exp	7.6	14.7	(not measured)
FB	8.9	18.4	8.76
VN	6.1	15.8	10.28

Table 1: Intrinsic metrics vs. extrinsic metrics: Intrinsic metrics are based on Row I in Table 2. The extrinsic reference is the transcription time of an unaided human. The correction time for ASR output is measured in hours.

3.3 Intrinsic metrics: CER, WER, and consistency in transcriptions used as reference:

Although both CER and WER reference “error rate" in regards to character and word, respectively, the question of the accuracy of the *reference* itself is rarely explored (but cf. Saon et al., 2017). For YMC, only one speaker, Castillo García, is capable of accurate transcription, which in YMC is the sole gold standard for ASR training, validation, and testing. Thus there is a consistency to the transcription used as a reference.

In comparison, for Highland Puebla Nahuat (another language that the present team is exploring), the situation is distinct. Three native speaker experts have worked with Amith on transcription for over six years, but the reference for ASR development are native-speaker transcriptions carefully proofed by Amith, a process that both corrected simple errors and applied a single standard implemented by one researcher. When all three native speaker experts were asked to transcribe the same 90 minutes or recordings, and the results were compared, there was not an insignificant level of variation (9%).

The aforementioned scenario suggests the impact on ASR intrinsic metrics of variation in transcriptions across multiple annotators, or even inconsistencies of one skilled annotator in the context of incipient writing systems. This affects not only ASR output but also the evaluation of ASR accuracy via character and word error rates. It may be that rather than character and word *error* rate, it would be advisable to consider the character and word *discrepancy* rate a change in terminology that perhaps better communicates the idea that the differences between REF and HYP are often as much a matter of opinion as fact. The nature and value of utilizing intrinsic metrics (e.g., CER and WER) for evaluating ASR effectiveness for endangered language documentation merits rethinking.

An additional factor that has emerged in the YMC corpora, which contains very rapid speech, is what may be called “hypercorrection". This is not uncommon and may occur with lenited forms (e.g., writing $ndi^1ku^4chi^4$ when close examination of the acoustic signal reveals that the speaker used the fully acceptable lenited form $ndiu^{14}chi^4$) or when certain function words are reduced, at times effectively disappearing from the acoustic signal though not from the mind of a fluent speaker transcriber. In both cases, ASR "errors" might represent a more accurate representation of the acoustic signal than the transcription of even the most highly capable native speakers.

The above discussion also brings into question what it means to achieve human parity via an ASR system. Parity could perhaps best be considered as not based on CER and WER alone but on whether ASR output achieves a lower error rate in these two measurements as compared to what another skilled human transcriber might achieve.

3.4 Extrinsic metrics: Reduction of human effort as a goal for automatic speech recognition

Given the nature of EL documentation, which requires high levels of accuracy if the corpus is to be easily used for future linguistic research, it is essential that ASR-generated hypotheses be reviewed by an expert human annotator before permanent archiving. Certainly, audio can be archived with metadata alone or with unchecked ASR transcriptions (see Michaud et al., 2018, §4.3 and 4.4), but the workflow envisioned for YMC is to use ASR to reduce human effort while the archived corpus of audio and text maintains results equivalent to those that would be obtained by careful, and labor-intensive, expert transcription.

CER and WER were measured for YMC corpora with training sets of 10, 20, 50, and 92 hours. The CER/WER were as follows: 19.5/39.2 (10 hrs.), 12.7/26.2 (20 hrs.), 10.2/24.9 (50 hrs.), and 7.7/16.1 (92 hrs.); Table 5 in Shi et al. (2021). Measurement of human effort reduction suggests that with a corpus of 30–50 hours, even for a relatively challenging language such as YMC, E2E ASR can achieve the level of accuracy that allows a reduction of human effort by > 75 percent (e.g., from 40 to 10 hours, approximately).

Model	Unit	CER				WER			
		Exp(dev)	Exp(test)	FB	VN	Exp(dev)	Exp(test)	FB	VN
A	Morae	9.5	9.4	12.8	9.9	19.2	19.2	23.8	21.8
B	Morpheme	10.2	10.0	13.9	10.9	20.0	20.0	24.8	23.1
C	Word	12.0	11.9	14.0	11.4	19.3	19.3	21.2	20.2
D	BPE150	7.7	7.6	9.5	6.8	16.1	16.1	19.6	17.3
E	BPE500	7.6	7.7	9.3	6.6	15.8	16.0	19.1	16.7
F	BPE1000	7.9	7.7	9.8	6.8	16.1	15.9	19.5	16.9
G	BPE1500	7.9	7.8	10.1	6.9	16.3	16.1	19.8	16.9
H	ROVER (A-C)	9.2	9.2	12.5	9.4	21.8	22.0	27.0	23.6
I	ROVER(D-G)	7.5	7.6	**8.9**	**6.1**	14.6	**14.7**	**18.4**	**15.8**
J	ROVER(A-G)	**7.4**	**7.4**	9.0	**6.1**	**14.4**	14.8	18.6	15.9

Table 2: ASR results for different models with different units

Starting from the acoustic signal, Castillo García, a native speaker linguist, requires approximately 40 hours to transcribe 1 hour of YMC audio. Starting from initial ASR hypotheses incorporated into ELAN, this is reduced by approximately 75 percent to about 10 hours of effort to produce one finalized hour of time-coded transcription with marked segmentation of affixes and enclitics.

These totals are derived from measurements with the FB and VN corpora, the two corpora for which ASR provided the initial transcription, and Castillo subsequently corrected the output, keeping track of the time he spent. For the first corpus, Castillo required 58.20 hours to correct 6.65 hours of audio (from 173 of the 178 files that had not been first transcribed by a speaker trainee). This yields 8.76 hours of effort per hour of recording. The 5.16 hours (in 24 files) of the VN corpus required 53.07 hours to correct, a ratio of 10.28 hours of effort to finalize 1 hour of speech. Over the entire set of 197 files (11.81 hours), human effort was 111.27 hours, or 9.42 hours to correct 1 hour of audio. Given that the ASR system was trained on an underlying orthography, the final result of < 10 hours of human effort per hour of audio is a transcribed *and* partially parsed corpus. Table 3 presents an analysis of two lines of a recording that was first processed by E2E ASR and corrected by Castillo García. A fuller presentation and analysis are offered in the Appendix. This focus on extrinsic metrics reflects the realization that the ultimate goal of computational systems is not to achieve the lowest CER and WER but to help documentation initiatives more efficiently produce results that will benefit future stakeholders.

4 End-to-end ASR experiments

4.1 Experiment settings

Recently, E2E ASR has reached comparable or better performances than conventional Hidden-Markov-Model-based ASR (Graves and Jaitly, 2014; Chiu et al., 2018; Pham et al., 2019; Karita et al., 2019a; Shi et al., 2021). In practice, E2E ASR systems are less affected by linguistic constraints and are generally easier to train. The benefits of such systems are reflected in the recent trends of using end-to-end ASR for EL documentation (Adams et al., 2020; Thai et al., 2020; Matsuura et al., 2020; Hjortnaes et al., 2020; Shi et al., 2021).

In developing E2E ASR recipes for YMC, we have adopted transformer and conformer-based encoder-decoder networks with hybrid CTC/attention training (Karita et al., 2019b; Watanabe et al., 2017). We used the YMC-Exp (train-split) for training and other YMC corpora for evaluation. The hyper-parameters for the training and decoding follow Shi et al. (2021). Seven systems with different modeling units are examined in the experiments. Four systems employ the byte-pair encoding (BPE) method trained from unigram language models (Kudo and Richardson, 2018), with transcription alphabets limited to the 150, 500, 1000, and 1500 most frequent byte-pairs in the training set. The other three ASR systems adopt linguistic units, including word, morpheme, and mora. The YM word is defined as a stem with all prefixes (such as completetive ni^1-, causative sa^4-, and iterative nda^3-) separated from the stem by a hyphen; and all enclitics (particularly person markers for subjects, objects, and possessors, such as =yu^3, 1sg; =un^4, 2sg; =an^4, 3sgFem; =o^4, 1plIncl; as well as $= lu^3$, augmentive). Many vowel-initial enclitics have alternative vowels, and many encl-

ASR	yo’^{3}o^{4} xi^{13}i^{2} ba^{42} ndi^{4} ba’^{1}a^{3}=e^{2} ku^{3}-nu’^{3}ni^{2} tu^{3}tun^{4} kwi^{3}so$^{(3)}$=e^{4} mi^{4}i^{4} ti^{4} ba^{42} ko^{14}o^{3} yo’^{3}o^{4} kwa’^{1}an^{1} yo^{4}o^{4} xa^{14}ku’^{1}u^{1}
Exp	yo’^{3}o^{4} xi^{1}i^{32} ba^{42} ndi^{4} ba’^{1}a^{3}=e^{2} ku^{3}-nu’^{3}ni^{2} tu^{3}tun^{4} kwi^{3}so$^{(3)}$=e^{4} mi^{4}i^{4} ti^{4} ba^{42} ko^{14}o^{3} yo’^{3}o^{4} kwa’^{1}an^{1} ji’^{4}in$^{(4)}$=o^{4} xa^{14}ku’^{1}u^{1}
Note	ASR missed the word *ji’^{4}in^{4}* (‘with’, comitative) and as a result wrote the 1plInclusive as an independent pronoun and not an enclitic.
ASR	i^{3}ta$^{(2)}$=e^{2} ndi^{4} tan^{42} i^{4}in^{4} i^{3}ta^{2} tio^{3}o^{2} yu^{3}ku^{4} ya^{1} ba^{4}li^{4} coco nu^{14}u^{3} ñu’^{3}u^{4} sa^{3}kan^{4} i^{4}in^{4} i^{3}ta$^{(2)}$=e^{2}
Exp	i^{3}ta$^{(2)}$=e^{2} ndi^{4} tan^{42} i^{4}in^{4} i^{3}ta^{2} tio^{3}o^{2} yu^{3}ku^{4} ya^{1} ba^{4}li^{4} ko^{4}ko^{13} nu^{14}u^{3} ñu’^{3}u^{4} sa^{3}kan^{4} i^{4}in^{4} i^{3}ta$^{(2)}$=e^{2}
Note	ASR suggested Spanish ‘coco’ coconut for Mixtec *ko^{4}ko^{13}* (‘to be abundant[plants]’)

Table 3: Comparison of ASR and Expert transcription of two lines of recording (See Appendix A for full text).[4]

itics have alternative tones, depending on stem-final vowel and tone, respectively. Morphemes are stems, prefixes, and enclitics. The inflectional tone is not segmented out. The right boundary of a mora is a vowel or dipthong (with an optional <n> to indicate a nasalized vowel) followed by a tone. The left boundary is a preceding mora or word boundary. Thus the word ni^1-xa'^3nda^2=e^4 (completive-play(guitar)-1plIncl) would be divided into three morphemes ni^1-, xa'^3nda^2, =e^4 and into four morae given that xa'^3nda^2 would be segmented as xa'^3, nda^2.

We adopt recognizer output voting error reduction (ROVER) for the hypotheses combination (Fiscus, 1997). Three combinations have been evaluated: (1) ROVER among only linguistic units (i.e., morae, morpheme, and word), (2) ROVER among only sub-word units (in this case BPE); and (3) ROVER combination utilizing all seven systems.

4.2 Experimental results

Experimental results are presented in two subsections. The first addresses the performance of end-to-end ASR across three corpora, each with slightly different recording systems and content. As clear from the preceding discussion and illustrated in Table 2, in addition to training on the word unit, the YMC E2E ASR system was trained on six additional linguistic and informational sub-word units. ROVER was then used to produce composite systems in which the outputs of all seven systems were combined in three distinct manners. In all cases, ROVER combinations improved the result of any individual system, including the averages for either of the two types of units: linguistic and informational.

[4]Those interested in the recordings and associated ELAN files may visit Amith and Castillo García (2020).

ASR and ROVER across three YMC corpora: As evident in Table 2, across all corpora, informational units (BPE) are more efficient than linguistic units (word, morpheme, morae) in regards to ASR accuracy. The average CER/WER for linguistic units (rows A-C) was 10.4/19.5 (Exp[test]), 13.6/23.3 (FB), and 10.7/21.7 (VN). The corresponding figures for the BPE units (rows D–G) were 7.7/16.0 (Exp[test]), 9.7/19.5 (FB), and 6.8/16.8 (VN). In terms of percentage differences between the two types of units, the numbers are not insignificant. In regards to CER, performance improved from linguistic to informational units by 26.0, 28.7, and 36.4 percent across the Exp(Test), FB, and VN corpora. In regards to WER, performance improved by 17.9, 16.3, and 22.6 percent across the same three corpora.

The experiments also addressed two remaining questions: (1) does unweighted ROVER combination improve the accuracy of ASR results; (2) does adding linguistic unit performance units to the ROVER "voting pool" improve results over a combination of only BPE units. In regards to the first question: ROVER always improves results over any individual system (compare row H to rows A, B, and C, and row I to rows D, E, F, and G). The second question is addressed by comparing rows I (ROVER applied only to the four BPE results) to J (adding the ASR results for the three linguistic units into the combination). In only one of the six cases (CER of Exp[test]) does including word, morpheme, and morae lower the error rate from the results of a simple combination of the four BPE results (in this case from 7.6 [row I] to 7.4 [row J]). In one case, there is no change (CER for the VN corpus) and in four cases, including linguistic units slightly worsens the score from the combination of BPE units alone (row I with bold numbers). The implication of the preceding is that ASR using linguistic units yields significantly lower accuracy than ASR that uses informational (BPE) units. Combining the former with the latter in an unweighted ROVER system in most cases does not improve results. Whether a weighted combinatory system would do better is a question that will need to be explored.

5 Conclusion

A fundamental element of endangered language documentation is the creation of an extensive corpus of audio recordings accompanied by time-coded annotations in interlinear format. In the best of cases, such annotations include an accurate transcription aligned with morphological segmentation, glossing, and free translations. The degree to which such corpus creation is facilitated is the extrinsic metric by which ASR contributions to EL documentation should be considered. The project here discussed suggests a path to creating such corpora using end-to-end ASR technology to build up the resources (30–50 hours) necessary to train an ASR system with perhaps a 6–10 percent CER. Once this threshold is reached, it is unlikely that further improvement will significantly reduce the human effort needed to check the ASR output for accuracy. Indeed, even if there are no "errors" in the ASR output, confirmation of this through careful revision of the recording of the transcription would probably still take 3–4 hours. The effort reduction of 75 percent documented here for YMC is, therefore, approaching what may be considered the minimum amount of time to proofread transcription of natural speech in an endangered language.

This project has also demonstrated the advantage of using a practical orthography that separates affixes and clitics. In a relatively isolating language such as YM, such a system is not difficult for native speakers to write nor for ASR systems to learn. It has the advantage of creating a workflow in which parsed text is the direct output of E2E ASR. The error rate evaluations across the spectrum of corpora and CER/WER also demonstrate the advantage of using subword units such as BPE and subsequent processing by ROVER for system combination (see above and Table 2). The error rates could perhaps be lowered further as the corpus increases in size, as more care is placed on recording environments, and as normalization eliminates reported errors for minor discrepancies such as in transcription of back-channel cues. But such lower error rates will probably not significantly reduce the time for final revision.

A final question concerns additional steps once CER is reduced to 6–8 percent, and additional improvements to ASR would not significantly affect the human effort needed to produce a high-quality time-coded transcription and segmentation. Four topics are suggested: (1) address issues of noise, overlapping speech, and other challenging recording situations; (2) focus on transfer learning to related languages; (3) explore the impact of "colonialization" by a dominant language; and (4) focus additional ASR-supported corpus development on producing material for documentation of endangered cultural knowledge, a facet of documentation that is often absent from endangered language documentation projects.

Acknowledgments

The authors gratefully acknowledge the following support for documenting and studying Yoloxóchitl Mixtec: National Science Foundation, Documenting Endangered Languages (DEL): Awards 1761421, 1500595, 0966462 (Amith, PI on all three; 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). The following support is acknowledged for documenting and studying Highland Puebla Nahuat: NSF 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); and the Comisión Nacional para el Conocimiento y Uso de la Biodiversidad, Mexico (Gerardo Salazar, PI; Amith, co-PI). The FOMA FST for Yoloxóchitl Mixtec was built by Jason Lilley, Amith, and Castillo with support from NSF DEL Award 1360670 (Christian DiCanio, PI).

Finally, the authors thank Shinji Watanabe both for his advice and guidance and for the key role he played in bringing together a field linguist, a native speaker, and a computational linguist for this project.

References

Oliver Adams, Trevor Cohn, Graham Neubig, Hilaria Cruz, Steven Bird, and Alexis Michaud. 2018. Evaluating phonemic transcription of low-resource tonal languages for language documentation. In *LREC 2018 (Language Resources and Evaluation Conference)*, pages 3356–3365.

Oliver Adams, Benjamin Galliot, Guillaume Wisniewski, Nicholas Lambourne, Ben Foley, Rahasya Sanders-Dwyer, Janet Wiles, Alexis Michaud, Séverine Guillaume, Laurent Besacier, et al. 2020. User-friendly automatic transcription of low-resource languages: Plugging ESPnet into Elpis. In *ComputEL-4: Fourth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 51–62.

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.

Claude Barras, Edouard Geoffrois, Zhibiao Wu, and Mark Liberman. 1998. Transcriber: A free tool for segmenting, labeling and transcribing speech. In *First International Conference on Language Resources and Evaluation (LREC)*, pages 1373–1376.

Steven Bird and Gary Simons. 2003. Seven dimensions of portability for language documentation and description. *Language*, pages 557–582.

Malgorzata Ćavar, Damir Ćavar, and Hilaria Cruz. 2016. Endangered language documentation: Bootstrapping a Chatino speech corpus, forced aligner, ASR. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4004–4011.

Chung-Cheng Chiu, Tara N Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J Weiss, Kanishka Rao, Ekaterina Gonina, et al. 2018. State-of-the-art speech recognition with sequence-to-sequence models. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4774–4778.

Nancy C Dorian. 1978. The fate of morphological complexity in language death: Evidence from East Sutherland Gaelic. *Language*, 54(3):590–609.

Jonathan G Fiscus. 1997. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, pages 347–354.

Ben Foley, Josh Arnold, Rolando Coto-Solano, Gautier Durantin, E Mark, Daan van Esch, Scott Heath, Frantisek Kratochvil, Zara Maxwell-Smith, David Nash, et al. 2018. Building speech recognition systems for language documentation: The CoEDL Endangered Language Pipeline and Inference System (ELPIS). *The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*.

Ben Foley, Alina Rakhi, Nicholas Lambourne, Nicholas Buckeridge, and Janet Wiles. 2019. ELPIS: An accessible speech-to-text tool. *Proc. Interspeech 2019*, pages 4624–4625.

Rey Castillo García. 2007. La fonología tonal del mixteco de Yoloxóchitl, Guerrero. Master's thesis, Centro de Investigaciones y Estudios Superiores en Antropología Social, Mexico City, Mexico. MA thesis in Lingüística Indoamericana.

Jeff Good. 2018. Reflections on the scope of language documentation. *Reflections on Language Documentation 20 Years after Himmelmann 1998. Language Documentation & Conservation, Special Publication 15*, pages 13–21.

Jeff Good, Julia Hirschberg, and Rambow Owen, editors. 2021. *Proceedings of the Workshop on Computational Methods for Endangered Languages*, volume 1-4.

Alex Graves and Navdeep Jaitly. 2014. Towards end-to-end speech recognition with recurrent neural networks. In *International Conference on Machine Learning*, pages 1764–1772.

Pengcheng Guo, Florian Boyer, Xuankai Chang, Tomoki Hayashi, Yosuke Higuchi, Hirofumi Inaguma, Naoyuki Kamo, Chenda Li, Daniel Garcia-Romero, Jiatong Shi, et al. 2020. Recent developments on ESPNet toolkit boosted by conformer. *arXiv preprint arXiv:2010.13956*.

Vishwa Gupta and Gilles Boulianne. 2020. Speech transcription challenges for resource constrained indigenous language Cree. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 362–367.

Ken Hale, Michael Krauss, Lucille J Watahomigie, Akira Y Yamamoto, Colette Craig, LaVerne Masayesva Jeanne, and Nora C England. 1992. Endangered languages. *Language*, 68(1):1–42.

Kristine A Hildebrandt, Carmen Jany, and Wilson Silva. 2017a. *Documenting variation in endangered languages. Language Documentation & Conservation Special Publication 14*. University of Hawai'i Press.

Kristine A Hildebrandt, Carmen Jany, and Wilson Silva. 2017b. *Introduction: Documenting variation in endangered languages*, pages 1–7. University of Hawai'i Press.

Nikolaus P Himmelmann. 1998. Documentary and descriptive linguistics. *Linguistics*, 36:161–196.

Nikolaus P Himmelmann. 2018. Meeting the transcription challenge. *Reflections on Language Documentation 20 Years after Himmelmann 1998. Language Documentation & Conservation, Special Publication 15*, pages 33–40.

Nils Hjortnaes, 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.

Robbie Jimerson and Emily Prud'hommeaux. 2018. ASR for documenting acutely under-resourced indigenous languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Robert Jimerson, Kruthika Simha, Ray Ptucha, and Emily Prud'hommeaux. 2018. Improving ASR output for endangered language documentation. In *The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*.

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

Shigeki Karita, Nelson Enrique Yalta Soplin, Shinji Watanabe, Marc Delcroix, Atsunori Ogawa, and Tomohiro Nakatani. 2019b. Improving transformer-based end-to-end speech recognition with connectionist temporal classification and language model integration. *Proc. Interspeech 2019*, pages 1408–1412.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.

Kohei Matsuura, Sei Ueno, Masato Mimura, Shinsuke Sakai, and Tatsuya Kawahara. 2020. Speech corpus of Ainu folklore and end-to-end speech recognition for Ainu language. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2622–2628.

Bradley McDonnell, Andrea L Berez-Kroeker, and Gary Holton. 2018. *Reflections on Language Documentation 20 Years after Himmelmann 1998. Language Documentation & Conservation, Special Publication 15.* University of Hawai'i Press.

Alexis Michaud, Oliver Adams, Trevor Anthony Cohn, Graham Neubig, and Séverine Guillaume. 2018. Integrating automatic transcription into the language documentation workflow: Experiments with Na data and the Persephone toolkit. *Language Documentation & Conservation*, 12.

Vikramjit Mitra, Andreas Kathol, Jonathan D Amith, and Rey Castillo García. 2016. Automatic speech transcription for low-resource languages: The case of Yoloxóchitl Mixtec (Mexico). In *Proc. Interspeech 2016*, pages 3076–3080.

Simon Musgrave and Nicholas Thieberger. 2021. The language documentation quartet. In *Proceedings of the Workshop on Computational Methods for Endangered Languages*, volume 1, pages 6–12.

Enrique L Palancar, Jonathan D Amith, and Rey Castillo García. 2016. Verbal inflection in Yoloxóchitl Mixtec. *Tone and inflection: New facts and new perspectives*, pages 295–336.

Ngoc-Quan Pham, Thai-Son Nguyen, Jan Niehues, Markus Müller, and Alex Waibel. 2019. Very deep self-attention networks for end-to-end speech recognition. *Proceedings of Interspeech 2019*, pages 66–70.

George Saon, Gakuto Kurata, Tom Sercu, Kartik Audhkhasi, Samuel Thomas, Dimitrios Dimitriadis, Xiaodong Cui, Bhuvana Ramabhadran, Michael Picheny, Lynn-Li Lim, et al. 2017. English conversational telephone speech recognition by humans and machines. *Proc. Interspeech 2017*, pages 132–136.

Frank Seifart, Nicholas Evans, Harald Hammarström, and Stephen C Levinson. 2018. Language documentation twenty-five years on. *Language*, 94(4):e324–e345.

Jiatong Shi, Jonathan D. Amith, 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*.

Jorge A Suárez. 1983. *The Mesoamerican Indian languages*. Cambridge University Press.

Bao Thai, Robert Jimerson, Raymond Ptucha, and Emily Prud'hommeaux. 2020. Fully convolutional ASR for less-resourced endangered languages. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 126–130.

Shinji Watanabe, Florian Boyer, Xuankai Chang, Pengcheng Guo, Tomoki Hayashi, Yosuke Higuchi, Takaaki Hori, Wen-Chin Huang, Hirofumi Inaguma, Naoyuki Kamo, et al. 2020. The 2020 ESPNet update: New features, broadened applications, performance improvements, and future plans. *arXiv preprint arXiv:2012.13006*.

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.

Douglas Whalen and Ćavar Damir. 2012. Collaborative research: Automatically annotated repository of digital video and audio resources community (AARDVARC). https://nsf.gov/awardsearch/showAward?AWD_ID=1244713. Accessed: 2021-03-05.

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.

A Analysis of ASR errors in one recording from the FB corpus

Unique identifier: 2017-12-01-b
Speakers: Constantino Teodoro Bautista and Esteban Guadalupe Sierra
Spanish: The first 13 seconds (3 segments) of the recording were of a Spanish speaker describing the plant being collected (*Passiflora biflora* Lam.) and have not been included below.
Note: A total 16 out of 33 segments/utterances are without ASR error. These are marked with an asterisk.
Original recording and ELAN file: Download at `http://www.balsas-nahuatl.org/NLP`

4*. 00:00:13.442 –> 00:00:17.105
ASR constantino teodoro bautista
Exp Constantino Teodoro Bautista.
Notes: ASR does not output caps or punctuation.

5*. 00:00:17.105 –> 00:00:19.477
ASR ya^{1} mi^{4}i^{4} tu^{1}tu’^{4}un^{4} ku^{3}rra^{42}
Exp Ya^{1} mi^{4}i^{4} tu^{1}tu’^{4}un^{4} ku^{3}rra^{42}
Notes: No errors in the ASR hypothesis.

6. 00:00:19.477 –> 00:00:23.688
ASR ta^{1} mas^{4}tru^{2} tela ya^{1} i^{3}chi^{4} ya^{3}tin^{3} ye’^{1}4e^{4} ku^{3}rra^{42} ndi^{4} covalentín yo’^{4}o^{4}
Exp ta^{1} mas^{4}tru^{2} Tele ya^{1} i^{3}chi^{4} ya^{3}tin^{3} ye’^{1}4e^{4} ku^{3}rra^{42} Nicu Valentín yo’^{4}o^{4},
Notes: ASR missed the proper name, Nicu Valentín (short for Nicolás Valentín) but did get the accent on Valentín, while mistaking the first name Nicu for *ndi*^{4} co[valentín]

7*. 00:00:23.688 –> 00:00:31.086
ASR ya^{1} i^{3}chi^{4} kwa’^{1}an$^{(1)}$=e^{4} tan^{3} xa^{1}a$^{(1)}$=e^{4} ku^{3}rra^{42} chi^{4}ñu^{3} ka^{4}chi^{2}=na^{1} ya^{1} kwa’^{1}an^{1} ni^{1}nu^{3} yo’^{4}o^{4} ju^{13}ta’^{3}an^{2}=ndu^{1} ya^{1} ko^{4}ndo^{3} kwi^{1}yo’^{1}o^{4} ndi^{3}ku’^{3}un^{3}
Exp ya^{1} i^{3}chi^{4} kwa’^{1}an$^{(1)}$=e^{4} tan^{3} xa^{1}a$^{(1)}$=e^{4} ku^{3}rra^{42} chi^{4}ñu^{3} ka^{4}chi^{2}=na^{1} ya^{1} kwa’^{1}an^{1} ni^{1}nu^{3} yo’^{4}o^{4} ju^{13}ta’^{3}an^{2}=ndu^{1} ya^{1} ko^{4}ndo^{3} kwi^{1}yo’^{1}o^{4} ndi^{3}ku’^{3}un^{3}
Notes: No errors in the ASR hypothesis.

8*. 00:00:31.086 –> 00:00:37.318
ASR kwi^{1}yo’^{1}o^{4} ndi^{3}ku’^{3}un^{3} kwi^{4}i^{24} ka^{4}chi^{2}=na^{1} yo’^{4}o^{4} ndi^{4} ya^{1} yo’^{4}o^{4} ndi^{4} xa’^{4}nu^{3} su^{4}kun^{1} mi^{4}i^{4} ti^{4} ba^{42} i^{4}yo$^{(2)}$=a^{2} mi^{4}i^{4} bi^{1}xin^{3} tan^{3}
Exp kwi^{1}yo’^{1}o^{4} ndi^{3}ku’^{3}un^{3} kwi^{4}i^{24} ka^{4}chi^{2}=na^{1} yo’^{4}o^{4} ndi^{4} ya^{1} yo’^{4}o^{4} ndi^{4} xa’^{4}nu^{3} su^{4}kun$^{(1)}$=a^{1} mi^{4}i^{4} ti^{4} ba^{42} i^{4}yo$^{(2)}$=a^{2} mi^{4}i^{4} bi^{1}xin^{3} tan^{3}
Notes: The ASR hypothesis missed the inanimate enclitic after the verb *su*^{4}*kun*^{1} and as a result failed to mark the elision of the stem-final low tone as would occur before a following low-tone enclitic.

9. 00:00:37.318 –> 00:00:42.959
ASR yo’^{3}o^{4} xi^{13}i^{2} ba^{42} ndi^{4} ba’^{1}a^{3}=e^{2} ku^{3}-nu’^{3}ni^{2} tu^{3}tun^{4} kwi^{3}so$^{(3)}$=e^{4} mi^{4}i^{4} ti^{4} ba^{42} ko^{14}o^{3} yo’^{3}o^{4} kwa’^{1}an^{1} yo^{4}o^{4} xa^{14}ku’^{1}u^{1}
Exp yo’^{3}o^{4} xi^{1}i^{32} ba^{42} ndi^{4} ba’^{1}a^{3}=e^{2} ku^{3}-nu’^{3}ni^{2} tu^{3}tun^{4} kwi^{3}so$^{(3)}$=e^{4} mi^{4}i^{4} ti^{4} ba^{42} ko^{14}o^{3} yo’^{3}o^{4} kwa’^{1}an^{1} ji’^{4}in$^{(4)}$=o^{4} xa^{14}ku’^{1}u^{1},
Notes: ASR missed the word *ji’*^{4}*in*^{4} (’with’, comitative) and as a result wrote the 1plInclusive as an independent pronoun and not an enclitic.

10. 00:00:42.959 –> 00:00:49.142
ASR i^{3}ta$^{(2)}$=e^{2} ndi^{4} tan^{42} i^{4}in^{4} i^{3}ta^{2} tio^{3}o^{2} yu^{3}ku^{4} ya^{1} ba^{4}li^{4} coco nu^{14}u^{3} ñu’^{3}u^{4} sa^{3}kan^{4} i^{4}in^{4} i^{3}ta$^{(2)}$=e^{2}

Exp i^{3}ta$^{(2)}$=e^{2} ndi^{4} tan^{42} i^{4}in^{4} i^{3}ta^{2} tio^{3}o^{2} yu^{3}ku^{4} ya^{1} ba^{4}li^{4} ko^{4}ko^{13} nu^{14}u^{3} ñu'^{3}u^{4} sa^{3}kan^{4} i^{4}in^{4} i^{3}ta$^{(2)}$=e^{2},

Notes: ASR suggested Spanish 'coco' coconut for Mixtec *ko*^{4}*ko*^{13} ('to be abundant[plants]'). Note that 'coco' was spelled as it is in Spanish and no tones were included in the ASR output.

11. 00:00:49.142 –> 00:00:53.458

ASR la^{3}tun^{4}=ni^{42} ya^{3}a$^{(3)}$=e^{2} tan^{3} ti^{1}xin^{3}=a^{2} ndi^{4} ya^{1} nde'^{3}e^{4} ba^{42} tan^{3} o^{4}ra^{2} xi^{4}yo^{13} ndu^{1}u^{4}=a^{2} ndi^{4} ya^{1} kwi^{4}i^{24} ba^{43}

Exp la^{3}tun^{4}=ni^{42} ya^{3}a$^{(3)}$=e^{2} tan^{3} ti^{1}xin^{3}=a^{2} ndi^{4} ya^{1} nde'^{3}e^{4} ba^{42} tan^{3} o^{4}ra^{2} xi^{4}yo^{13} ndu^{1}u^{4}=a^{2} ndi^{4} ya^{1} kwi^{4}i^{24} ba^{42},

Notes: ASR missed tone 42, writing 43 instead. Note that the two tone patterns are alternate forms of the same word, the copula used in regards to objects.

12*. 00:00:53.458 –> 00:00:57.279

ASR tan^{3} o^{4}ra^{2} chi^{4}chi^{13}=a^{2} ndi^{4} ndu^{1}u^{4} nde'^{3}e^{4} ku^{4}u^{4} ndu^{1}u^{4}=a^{3}

Exp tan^{3} o^{4}ra^{2} chi^{4}chi^{13}=a^{2} ndi^{4} ndu^{1}u^{4} nde'^{3}e^{4} ku^{4}u^{4} ndu^{1}u^{4}=a^{3}.

Notes: No errors in the ASR hypothesis.

13*. 00:00:57.279 –> 00:01:02.728

ASR yu^{1}ku$^{(1)}$=a^{1} ndi^{4} tan^{42} i^{4}in$^{(4)}$=a^{2} ni^{1}-xa'^{3}nda^{2}=e^{4} tan^{42} i^{4}in^{4} yu^{1}ku^{1} tun^{4} si^{13}su^{2} kan^{4} sa^{3}kan^{4} i^{4}in^{4} yu^{1}ku$^{(1)}$=a^{1} tan^{3} ndi^{4}

Exp Yu^{1}ku$^{(1)}$=a^{1} ndi^{4} tan^{42} i^{4}in$^{(4)}$=a^{2} ni^{1}-xa'^{3}nda^{2}=e^{4} tan^{42} i^{4}in^{4} yu^{1}ku^{1} tun^{4} si^{13}su^{2} kan^{4} sa^{3}kan^{4} i^{4}in^{4} yu^{1}ku$^{(1)}$=a^{1} tan^{3} ndi^{4}

Notes: No errors in the ASR hypothesis.

14. 00:01:02.728 –> 00:01:06.296

ASR su^{14}u^{3} ya^{1} xa'^{4}nda^{2}=na^{1} ba^{42} ndi^{4} su^{14}u^{3} ki^{3}ti^{4} ja^{4}xi^{24}=ri^{4} sa^{3}kan^{4} i^{4}in^{4} yu^{1}ku^{1} mi^{4}i^{4} ba$^{(3)}$=e^{3}

Exp su^{14}u^{3} ya^{1} xa'^{4}nda^{2}=na^{1} ba^{42} tan^{3} ni^{4} su^{14}u^{3} ki^{3}ti^{4} ja^{4}xi^{24}=ri^{4}, sa^{3}kan^{4} i^{4}in^{4} yu^{1}ku^{1} mi^{4}i^{4} ba$^{(3)}$=e^{3},

Notes: ASR mistakenly proposed *ndi*^{4} for *tan*^{3} *ni*^{4}.

15*. 00:01:06.296 –> 00:01:10.981

ASR tan^{3} ya^{1} xa'^{4}nu^{3} su^{4}kun$^{(1)}$=a^{1} mi^{4}i^{4} ti^{4} ba^{42} sa^{3}ba^{3} xia^{4}an^{4} ku^{3}ta'^{3}an^{2}=e^{4}=e^{2} ndi^{4} xa'^{4}nu$^{(3)}$=a^{2} kwa^{1}nda^{3}a$^{(3)}$=e^{2} nda'^{3}a^{4} i^{3}tun^{4}

Exp tan^{3} ya^{1} xa'^{4}nu^{3} su^{4}kun$^{(1)}$=a^{1} mi^{4}i^{4} ti^{4} ba^{42} sa^{3}ba^{3} xia^{4}an^{4} ku^{3}ta'^{3}an^{2}=e^{4}=e^{2} ndi^{4} xa'^{4}nu$^{(3)}$=a^{2} kwa^{1}nda^{3}a$^{(3)}$=e^{2} nda'^{3}a^{4} i^{3}tun^{4}

Notes: No errors in the ASR hypothesis.

16. 00:01:10.981 –> 00:01:14.768

ASR u^{1}xi^{1} an^{4} nda^{1} xa'^{1}un^{1} metru ka^{1}a^{3} mi^{4}i^{4} i^{4}yo^{2} i^{3}tun^{4} ndo^{3}o^{3} tan^{3} ko^{4}ko^{13}=a^{2} kwa^{1}nde^{3}e^{3} ni^{1}nu^{3}

Exp u^{1}xi^{1} an^{4} nda^{1} xa'^{1}un^{1} metru ka^{1}a^{3} mi^{4}i^{4} i^{4}yo^{2} i^{3}tun^{4} ndo^{3}o^{3} tan^{3} ko^{4}ko^{13}=a^{2} kwa^{1}nda^{3}a$^{(3)}$=e^{2} ni^{1}nu^{3},

Notes: Not only did ASR recognize the Spanish *metru* borrowing but wrote it according to our conventions, without tone. Note that the correct underlying form *kwa*^{1}*nda*^{3}*a*$^{(3)}$*=e*^{2} (progressive of 'to climb [e.g., a vine]' with 3sg enclitic for inanimates =e^{2}) surfaces as *kwa*^{1}*nde*^{3}*e*^{2} quite close to the ASR hypothesis of *kwa*^{1}*nde*^{3}*e*^{3}, which exists, but as a distinct word (progressive of 'to enter[pl]').

17*. 00:01:14.768 –> 00:01:18.281

ASR mi^{4}i^{4} ba^{143} xa'^{4}nda^{2}=na$^{(1)}$=e^{1} ndi^{4} xa'^{4}nu^{3} su^{4}kun$^{(1)}$=a^{1}

Exp mi^{4}i^{4} ba^{143} xa'^{4}nda^{2}=na$^{(1)}$=e^{1} ndi^{4} xa'^{4}nu^{3} su^{4}kun$^{(1)}$=a^{1},

Notes: No errors in the ASR hypothesis.

18*. 00:01:18.281 –> 00:01:21.487
ASR ya^{1} kan^{4} ku^{4}u^{4} kwi^{1}yo’^{1}o^{4} ju^{13}ta’^{3}an^{2}=ndu^{1} i^{3}chi^{4} kwa’^{1}an^{1} ku^{3}rra^{42} chi^{4}ñu^{3} yo’^{4}o^{4}
Exp ya^{1} kan^{4} ku^{4}u^{4} kwi^{1}yo’^{1}o^{4} ju^{13}ta’^{3}an^{2}=ndu^{1} i^{3}chi^{4} kwa’^{1}an^{1} ku^{3}rra^{42} chi^{4}ñu^{3} yo’^{4}o^{4}.
Notes: No errors in the ASR hypothesis.

19*. 00:01:21.487 –> 00:01:24.658
ASR esteban guadalupe sierra
Exp Esteban Guadalupe Sierra.
Notes: ASR does not output caps or punctuation.

20. 00:01:24.658 –> 00:01:27.614
ASR ya^{1} ko^{4}ndo^{3} kwi^{1}yo’^{1}o^{4} ndi^{13}-kwi^{3}so^{3}=ndu^{2} ya^{1}
Exp ya^{1} ko^{4}ndo^{3} kwi^{1}yo’^{1}o^{4} ndi^{13}-kwi^{3}so^{3}=ndu^{2} ya^{1}
Notes: No errors in the ASR hypothesis.

21. 00:01:27.614 –> 00:01:33.096
ASR sa^{3}kan^{4} tan^{3} xa^{1}a$^{(1)}$=e^{4} ku^{3}rra^{42} chi^{4}ñu^{3} ya^{1} ja^{1}ta^{4} ku^{3}rra^{42} ta^{1} marspele yo’^{4}o^{4} ndi^{4}
Exp sa^{3}kan^{4} tan^{3} xa^{1}a$^{(1)}$=e^{4} ku^{3}rra^{42} chi^{4}ñu^{3} ya^{1} ja^{1}ta^{4} ku^{3}rra^{42} ta^{1} mas^{4}tru^{2} Tele yo’^{4}o^{4} ndi^{4}
Notes: ASR missed the Spanish *mas*^{4}*tru*^{2} Tele (teacher Tele(sforo)) and hypothesized a nonsense word in Spanish (note absence of tone as would be the case for Spanish loans).

22. 00:01:33.096 –> 00:01:39.611
ASR kwi^{1}yo’^{1}o^{4} ndi^{3}ku’^{3}un^{3} ba^{3} kwi^{1}yo’^{1}o^{4} ndi^{3}ku’^{3}un^{3} ka^{1}a^{3} ndi^{4} ko^{14}o^{3} u^{1}bi^{1} u^{1}ni^{1} nu^{14}u$^{(3)}$=a^{2} ña^{1}a^{4} ndi^{4} i^{3}nda^{14} nu^{14}u^{3} sa^{3}kan^{4} ba^{3} ba^{42}
Exp kwi^{1}yo’^{1}o^{4} ndi^{3}ku’^{3}un^{3} ba^{43}, kwi^{1}yo’^{1}o^{4} ndi^{3}ku’^{3}un^{3} ka^{1}a^{3} ndi^{4} ko^{14}o^{3} u^{1}bi^{1} u^{1}ni^{1} nu^{14}u$^{(3)}$=a^{2} ndi^{4} i^{3}nda^{14} nu^{14}u^{3} sa^{3}kan^{4} ba^{3} ba^{42},
Notes: ASR mistook the copula *ba*^{43} and instead hypothesized the modal *ba*^{3}. ASR also inserted a word not present in the signal, *ña*^{1}*a*^{4} (‘over there’).

23. 00:01:39.611 –> 00:01:43.781
ASR ya^{1} ka’^{4}an^{2}=na^{1} ji’^{4}in^{4} ku^{4}u^{4} kwi^{1}yo’^{1}o^{4} ndi^{3}ku’^{3}un^{3} kwi^{4}i$^{2(4)}$=o^{4} tan^{3}
Exp ya^{1} ka’^{4}an^{2}=na^{1} ji’^{4}in^{4} ku^{4}u^{4} kwi^{1}yo’^{1}o^{4} ndi^{3}ku’^{3}un^{3} kwi^{4}i^{24} yo’^{4}o^{4} tan^{3}
Notes: ASR mistook the adverbial *yo’*^{4}*o*^{4} (’here’) as the enclitic =*o*^{4} (1plIncl) and as a result also hypothesized stem final tone elision (4).

24. 00:01:43.781 –> 00:01:49.347
ASR ba^{14}3 bi^{4}xi^{1} i^{4}in$^{(4)}$=a^{2} ndi^{4} kwi^{1}yo’^{1}o^{4} kwa^{1}nda^{3}a^{3} nda’^{3}a^{4} i^{3}tun^{4} ba^{3} tan^{3} kwi^{1}yo’^{1}o^{4}
Exp ba^{14}3 bi^{4}xi^{1} i^{4}in$^{(4)}$=a^{2} ndi^{4} kwi^{1}yo’^{1}o^{4} kwa^{1}nda^{3}a^{3} nda’^{3}a^{4} i^{3}tun^{4} ba^{42} tan^{3} kwi^{1}yo’^{1}o^{4}
Notes: As in segment #22 above, ASR mistook the copula, here *ba*^{4}, and instead hypothesized the modal *ba*^{3}.

25. 00:01:49.347 –> 00:01:55.001
ASR ndi^{3}i^{4} ba^{42} ko^{14}o^{3} tu^{4}mi^{4} ja^{1}ta^{4}=e^{2} ya^{1} kan^{4} ndi^{4} i^{4}yo^{2} i^{4}yo^{2} xi^{1}ki^{4}=a^{2} i^{4}in^{4} tan^{3}
Exp ndi^{3}i^{4} ba^{42} ko^{14}o^{3} tu^{4}mi^{4} ja^{1}ta^{4}=e^{2} tan^{3} ndi^{4} i^{4}yo^{2} i^{4}yo^{2} xi^{1}ki^{4}=a^{2} i^{4}in^{4} tan^{3}
Notes: ASR missed the conjunction *tan*^{3} (’and’) and instead wrote *ya*^{1} *kan*^{4} (’that one’).

26*. 00:01:55.001 –> 00:02:00.110
ASR ya^{1} ba’^{1}a^{3}=e^{2} ndi^{4} ba’^{1}a^{3}=e^{2} ju^{4}-nu’^{3}ni^{2} tu^{3}tun^{4} i^{4}xa^{3}=na^{2}
Exp ya^{1} ba’^{1}a^{3}=e^{2} ndi^{4} ba’^{1}a^{3}=e^{2} ju^{4}-nu’^{3}ni^{2} tu^{3}tun^{4} i^{4}xa^{3}=na^{2},
Notes: No errors in the ASR hypothesis.

27*. 00:02:00.110 –> 00:02:04.380
ASR na^{1}kwa^{4}chi^{3} tu^{3} ndi^{4} chi^{3}ñu^{3}=ni^{42}=na^{1} ka^{3}ya^{2}=na$^{(1)}$=e^{1} su^{4}-kwe^{1}kun^{1}=na^{1} i^{3}na^{2} ju^{4}si^{4}ki^{24} ba^{3}=na^{3}
Exp na^{1}kwa^{4}chi^{3} tu^{3} ndi^{4} chi^{3}ñu^{3}=ni^{42}=na^{1} ka^{3}ya^{2}=na$^{(1)}$=e^{1} su^{4}-kwe^{1}kun^{1}=na^{1} i^{3}na^{2} ju^{4}si^{4}ki^{24} ba^{3}=na^{3},
Notes: No errors in the ASR hypothesis.

28*. 00:02:04.380 –> 00:02:06.242
ASR a^{1}chi^{1} kwi^{1}yo'^{1}o^{4} nde^{3}e^{4} ba^{43}
Exp a^{1}chi^{1} kwi^{1}yo'^{1}o^{4} nde^{3}e^{4} ba^{43},
Notes: No errors in the ASR hypothesis.

29*. 00:02:06.242 –> 00:02:08.865
ASR tan^{42} ka'^{4}an^{2} ta^{1} ta^{4}u^{3}ni^{2} constantino yo'^{4}o^{4} ndi^{4}
Exp tan^{42} ka'^{4}an^{2} ta^{1} ta^{4}u^{3}ni^{2} Constantino yo'^{4}o^{4} ndi^{4}
Notes: No errors in the ASR hypothesis.

30*. 00:02:08.865 –> 00:02:13.473
ASR i^{3}ta$^{(2)}$=e^{2} ndi^{4} tan^{42} i^{4}in^{4} i^{3}ta^{2} ya^{1}kan^{3} kwi^{1}yo'^{1}o^{4} ya^{1} i^{3}ta^{2} tio^{3}o^{2} kan^{4} sa^{3}kan^{4} i^{4}in^{4} i^{3}ta$^{(2)}$=e^{2} tan^{3}
Exp i^{3}ta$^{(2)}$=e^{2} ndi^{4} tan^{42} i^{4}in^{4} i^{3}ta^{2}, ya^{1}kan^{3}, kwi^{1}yo'^{1}o^{4} ya^{1} i^{3}ta^{2} tio^{3}o^{2} kan^{4} sa^{3}kan^{4} i^{4}in^{4} i^{3}ta$^{(2)}$=e^{2} tan^{3}
Notes: No errors in the ASR hypothesis, the fifth consecutive annotation without an ASR error.

31. 00:02:13.473 –> 00:02:17.927
ASR xi^{4}yo^{13} a^{1}su^{3} tan^{42} i^{4}in^{4} tio^{1}o^{32} i^{4}in$^{(4)}$=a^{2} ba^{4}li^{4} ko^{4}ndo^{3} ndu'^{1}u^{4}=a^{2} ya^{1} kwi^{4}i^{24} ba^{42} na^{4}
Exp xi^{4}yo^{13} a^{1}su^{3} tan^{42} i^{4}in^{4} tio^{3}o^{2} i^{4}in$^{(4)}$=a^{2} ba^{4}li^{4} ko^{4}ndo^{3} ndu^{1}u^{4}=a^{2}, ya^{1} kwi^{4}i^{24} ba^{42} na^{4}
Notes: ASR missed a word, writing $tio^{1}o^{32}$ (a word that does not exist) for $tio^{3}o^{2}$ (the passion fruit, *Passiflora* sp.). It also miswrote $ndu^{1}u^{4}$ (fruit) as $ndu'^{1}u^{4}$ a verb ('to fall from an upright position').

32*. 00:02:17.927 –> 00:02:21.014
ASR i'^{4}i$^{(3)}$=a^{2} tan^{3} na^{4} chi^{4}chi^{13}=a^{2} ndi^{4} ya^{1} nde'^{3}e^{4} ba^{42}
Exp i'^{4}i$^{(3)}$=a^{2} tan^{3} na^{4} chi^{4}chi^{13}=a^{2} ndi^{4} ya^{1} nde'^{3}e^{4} ba^{42},
Notes: No errors in the ASR hypothesis.

33. 00:02:21.014 –> 00:02:25.181
ASR ya^{1} mi^{4}i^{4} bi^{1}xin^{3} ya^{3}tin^{3} yu'^{3}u^{4} yu^{3}bi^{2} kan^{4} ba^{42} xi^{4}yo$^{1(3)}$=a^{3}
Exp ya^{1} mi^{4}i^{4} bi^{1}xin^{3} ya^{3}tin^{3} yu'^{3}u^{4} yu^{3}bi^{2} i^{3}kan^{4} ba^{42} xi^{4}yo$^{1(3)}$=a^{3}.
Notes: ASR missed the initial i^{3} in $i^{3}kan^{4}$ ('there'). It is to be noted that kan^{4} is an alternate form of $i^{3}kan^{4}$.

34*. 00:02:25.181 –> 00:02:27.790
ASR ya^{1} kan^{4} ba^{42} ndi^{13}-kwi^{3}so^{3}=ndu^{2} yo'^{4}o^{4}
Exp Ya^{1} kan^{4} ba^{42} ndi^{13}-kwi^{3}so^{3}=ndu^{2} yo'^{4}o^{4},
Notes: No errors in the ASR hypothesis.

35*. 00:02:27.790 –> 00:02:32.887
ASR tan^{3} ta^{1} ta^{4}u^{3}ni^{2} fernando yo'^{4}o^{4} ndi^{4} ji^{4}ni^{2}=ra$^{(1)}$=e^{1} ndi^{4} ji^{4}ni^{2}=ra$^{(1)}$=e^{1} ya^{1} sa^{3}kan^{4} i^{4}yo$^{(2)}$=a^{2} tan^{3}
Exp tan^{3} ta^{1} ta^{4}u^{3}ni^{2} Fernando yo'^{4}o^{4} ndi^{4} ji^{4}ni^{2}=ra$^{(1)}$=e^{1} ndi^{4} ji^{4}ni^{2}=ra$^{(1)}$=e^{1} ya^{1} sa^{3}kan^{4} i^{4}yo$^{(2)}$=a^{2} tan^{3}

Notes: No errors in the ASR hypothesis.

36. 00:02:32.887 –> 00:02:41.884
ASR ji^{14}ni^{2}=ra^{1} sa^{1}a^{3} na^{3}ni^{4}=a^{3} tan^{3} ni^{14}-ndi^{3}-kwi^{3}so^{3} ndu^{3}-ta^{1}chi^{4}=ra^{2} ji'^{4}in$^{(4)}$=a^{2} a^{1}chi^{1} ji^{14}ni^{2}=ra^{1} nda^{4}a$^{(2)}$=e^{2} ba'^{1}a$^{(3)}$=e^{3}
Exp ji^{14}ni^{2}=ra^{1} sa^{1}a^{3} na^{3}ni^{4}=a^{3}, tan^{3} ni^{14}-ndi^{3}-kwi^{3}so^{3}=ndu^{2} ta^{1}chi^{4}=ra^{2} ji'^{4}in$^{(4)}$=a^{2} a^{1}chi^{1} ji^{14}ni^{2}=ra^{1} nda^{4}a$^{(2)}$=e^{2} ba'^{1}a$^{(3)}$=e^{3}.
Notes: ASR hypothesized *ndu*^{3} as a verbal prefix instead of the correct interpretation as a person-marking enclitic (1plExcl) that is attached to the preceding verb.