


# Collaborative Research: DLI-DEL: Language Documentation with an AI Helper

Antonios Anastasopoulos<sup>†</sup>, David Chiang<sup>‡</sup>, Géraldine Walther<sup>†</sup>

<sup>†</sup>George Mason University, <sup>‡</sup>University of Notre Dame

## 1 Introduction

Language documentation practices have been developed and refined over several centuries and the materials they produce, including grammars, dictionaries, and annotated texts, constitute invaluable contributions to the conservation of humanity’s cultural heritage and the understanding of human language. However, traditional high-quality language documentation efforts take more time than can be afforded to keep up with current language extinction rates. According to Krauss [78], the current rate of language extinction could lead to the disappearance of 20–90% of today’s spoken languages by the end of the 21st century. The most constructive response to this crisis, as argued by Bird and collaborators [3, 33], is therefore to complement documentation efforts with data collection for as many languages as possible and use this data to document them later. Inexpensive recording, storage, and transmission technologies now make it possible to collect many hours of recordings in many languages. But in order for these data to be useful, they must be *interpretable*. Linguistics papers rarely present their data in the form of raw audio; they typically present them as

 *interlinear glossed text (IGT)*, which minimally includes a phonetic and/or orthographic transcription, a word-level or morpheme-level gloss, and a free translation. As manually annotating data is costly, both in time and human effort, we propose to leverage automatic annotation techniques that only rely on recorded and (partially) translated data. This will not only be highly valuable for preparing and speeding up full annotation of large quantities of collected data, but also offers the prospect of ultimately achieving annotation of greater consistency and quality.

We assume that it will be practical and relatively easy to obtain translations of these recordings into another, more widely-spoken language – for example, the mobile application Aikuma [35, 36] makes it possible to collect translations nearly in real time and has been used successfully in multiple data collection efforts [11, 36, 66, 2]. Furthermore, we suggest that translations contain enough information for computers to assist with or even to fully automate the process of converting recorded speech into IGT. In previous work [55, 19, 14, 16, 15], we developed methods for using translations to improve automatic speech transcription. But this previous work fell short of our ultimate goal in a number of ways. **It did not tackle the problem of languages that lack a standard writing system (or any writing system at all), and it did not address the problem of identifying and glossing morphemes. Moreover, it focused only on modeling single languages, missing the opportunity to use data from related (possibly higher-resource) languages.**

This proposal aims to get closer to the goal of automatically converting parallel audio to IGT, with three particular areas of focus. **We will model multiple levels of representation: not only audio and text, but also phones and morphemes. We will develop computational models of multiple languages, exploiting information from higher-resourced languages and improving processing of endangered ones. We will then package these models as components of an AI-helper toolbox that will integrate tightly with ELAN** and aid linguists in the language documentation process.



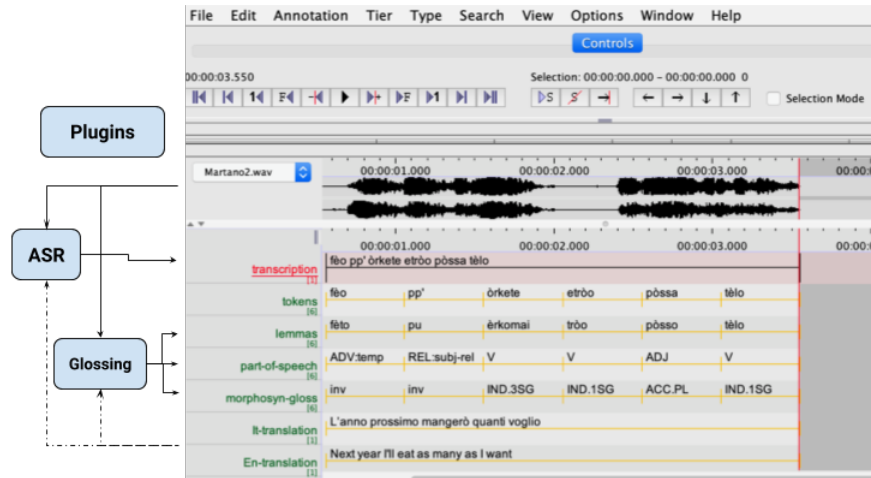


Figure 1: Example of enriched annotation in ELAN. Our AI Helper toolbox will automatically suggest utterance-level phonemic transcriptions, lemmas and morphological analyses.

## 1.1 Multiple levels

We want to explicitly model multiple levels of linguistic representation, for two (related) reasons. First, our training data will likely have multiple levels, and not necessarily uniformly complete; e.g., parts of the data might have only speech transcriptions and others only translations. Second, all levels of representation are relevant for linguistic study, with different analyses relying on information from different levels of annotation; **IGT typically records at least three.**

Recent advances in natural language processing (NLP) research make it opportune to build such multi-level models. **Whereas traditional NLP pipelines had sequences of models mapping from one level of representation to the next, current deep neural networks can learn multiple levels of representation at once:** for example, in a face recognition system, the lowest layer might learn features like lines, the next layer up parts like eyes and ears, and the last layer different types of faces. In natural language, these learned representations have been argued to be isomorphic to traditional linguistic representations [129], and the layers of a network can also be trained to fit explicit linguistic representations [132, 126].

## 1.2 Multiple languages

Leveraging data in multiple languages to train multilingual models has shown particular promise for improving the quality of NLP systems for under-resourced languages. **Successful approaches range from cross-lingual transfer between related languages to massively multilingual models trained on more than 100 languages. A current line of work investigates the extent to which such multilingual models capture truly general aspects of language representations** [114, 23, 75, 116, 143], with encouraging findings.

In this project, our goal is to reconcile two seemingly opposed requirements. The first is the development of machine learning approaches that will be able to *generalize* sufficiently to be applicable and useful for any of the world’s languages. The second is to design models that will capture the intricate *unique* characteristics that define individual languages. To do so, we will build upon recent multilingual NLP work [46, 86, 44, 23, 18, 77, *inter alia*]. Our models will be initialized using multilingual pre-training, but will have the capability to adapt to the language under study.

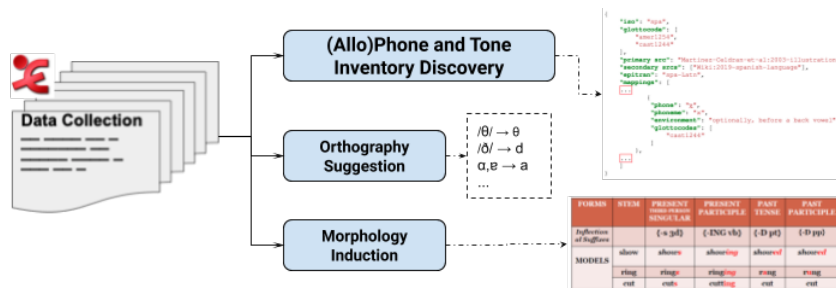


Figure 2: Schematic visualization of the corpus-level plugins of our AI Helper toolbox.

### 1.3 AI-helper toolbox

For machine learning models to be useful for language documentation, they will need to seamlessly integrate with linguists’ established workflows. To achieve this, we will package our models as plugins for popular annotation software, notably ELAN [142] and Toolbox. Our toolkit will include both *utterance-level* models for producing the different tiers of linguistic analysis and representation, as well as *corpus-level* models that will analyze the whole corpus to produce language-level analyses. Figure 1 shows the utterance-level tools alongside enriched annotations within ELAN, while Figure 2 shows the corpus-level models.

We emphasize that our **AI-helper toolkit is not intended as a replacement for linguists’ specialist analyses, but rather an assistant that will enable faster annotation and suggest higher-level groupings of system patterns**. Just as professional translators use computer-assisted translation tools to increase their speed and consistency without losing control over their work, so we hope to create tools that allow linguists to manually annotate far less data, instead simply correcting and refining the computer’s suggestions, which has been shown to significantly improve both annotation speed and accuracy [60]. Corrected data will in turn improve the AI-helper’s suggestions.

## 2 Speech/Phonology

A crucial step of linguistic annotation lies in a fine-grained representation of the speech signal that represents the language’s phonetic and phonological features in a coherent way. Typically, a linguist identifies phones, phonemes, and suprasegmental features such as tone (as well as graphemes, if the language has an established writing system) that are part of the language’s inventory, and then uses them to transcribe collected audio data.

Our AI-Helper toolkit will include two components to aid in this process: the **speech analysis component** will combine information from the whole data collection to produce tentative phone and tone inventories for the language, while the **speech recognition component** will operate at the utterance level and suggest transcriptions at different granularities (e.g., phone, phoneme, grapheme levels). Suggestions can then be refined by the linguist and the model updated and re-run.

### 2.1 Automatic speech recognition

**Tool and modeling approach** The automatic speech recognition (ASR) tool will be generally responsible for producing utterance-level annotations for every audio file in a collection, at multiple granularities (phones, phonemes, graphemes), using information from other representation levels and multiple languages. It will be implemented as an ELAN plug-in using the AV Recognizer API.

This tool will be similar to Persephone-ELAN,<sup>1</sup> which integrates the automatic phoneme recognition methods offered by the Persephone toolkit [9, 96] into ELAN. Our new tool’s capabilities will, however, go beyond the current capabilities of the Persephone toolkit, which still requires linguists to train the model on manual annotations before transcription. Our toolkit will include pre-packaged multilingually-pretrained models, which will be useful for producing first-pass phonemic transcriptions. It will also be able to fine-tune these pre-trained models on the linguist’s data collection, if some transcriptions are available. Finally, our underlying models will incorporate the advances discussed below, to further increase the tool’s accuracy and usefulness.

Our ASR models will be based on the current state of the art, which is the **transformer neural network [137] with hybrid attention–CTC loss [140, 41, 141, 98]**. In the following sections, we outline the extensions that we will make to such models to adapt them to the characteristics of endangered language data.

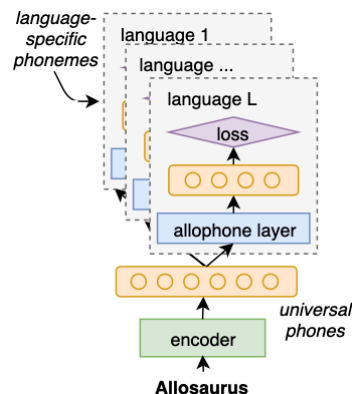
**Multiple languages** Adaptation of a high-resource language ASR system to work with a low-resource language has been shown to be a potentially viable path towards ASR for endangered languages [119, 131]. However, we propose to investigate what we consider to be an even more promising direction: the adaptation of **multilingual systems that leverage additional context (e.g., translations), two directions that are under-studied in the endangered languages setting.**

Multilingual ASR systems have typically been trained on a handful of the world’s best-resourced languages, relying on collections like GlobalPhone [120]. Only recently have models been trained on dozens of languages [e.g., 10] and deployed with successful results. Multilingual pre-training has also recently proven beneficial for text-based tasks [115, 125]. Building on the success of these models, we will adapt them to the particular characteristics of endangered languages. We will incorporate unsupervised pre-training similar to wav2vec [24], which has shown exceptional promise for low-resource settings, and extend it to multilingual pre-training.

A particular challenge of building ASR systems for endangered languages is that many such languages do not have a standardized orthography, or any orthography at all. **A more realistic alternative – and one that is often more useful to linguists – would be to build a system that produces a phonemic and/or phonetic transcription instead, using the International Phonetic Alphabet (IPA).**

In previous work, we took the first step towards building a generalized phonetic transcription system, Allosaurus, which was trained on 12 languages and produces IPA transcriptions for any audio input [84]. At the core of this model lies the idea that, although phonemic transcriptions, which glosses typically rely on, reflect language specific decisions about phonemic inventories, phones can function as neutral sound property representations. Allosaurus incorporates a language-specific allophonic mapping layer on top of a general phone representation that reinstates phonemic-like representations for suggested transcriptions (see figure at right).

Allosaurus has several desirable properties. First, it allows for multilingual training, leveraging existing data in high-resource languages. Second, it can produce outputs at several granularities (phones, phonemes). Third, it provides an interpretable mapping between these granularities (the allophone layer, in the case of phone-phoneme mappings) which can be learned in a language-specific manner. These properties make it ideal for



<sup>1</sup><https://github.com/coxchristopher/persephone-elan>

our speech transcription toolkit. In addition, we propose further orthogonal improvements to this system, which will make it even more effective as the first component of our annotation pipeline:

1. Broaden the system’s phone coverage to make it truly adapted for all languages.
2. Extend its output capabilities with suprasegmental information, notably with a focus on tone.
3. Add a phoneme to grapheme (p2g) layer for written languages.
4. Allow for *learning* the allophone and p2g layers, instead of requiring them in advance.
5. Improve accuracy and robustness to noise via extensions that take advantage of translations.

In the pilot study that introduced Allosaurus, the model was pre-trained on only 12 phonetically diverse languages. As a result, its phone inventory only covers about 82% of all inventoried phones [84]. In our work, we will take advantage of recent transcribed speech collection efforts such as the Mozilla Common Voice project [22], which collect audio and transcriptions in dozens of languages. We will combine these high-resource language data with publicly available small speech collections from linguistic archives like AILLA, ELAR, PARADISEC, and Pangloss. This will allow us to expand the phone inventory of the model to include missing phones and bring it as close as possible to a 100% coverage.

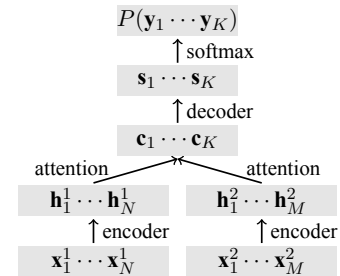
Utilizing linguistic archive data will also allow us to extend the transcription capabilities of our model to also include diacritics, e.g., mark (de)voicing or rhoticity, as well as suprasegmental information (stress, pitch accent, tone, etc.). Such narrow IPA representations are available in small amounts throughout linguistic archives. Our model’s truly multilingual learning capabilities will enable it to combine and leverage all available information from this collection of varied sources.

Incorporating tone recognition for low-resource tonal languages remains an important challenge [61]. According to the WALS database [90], over 40% of all languages are tonal; some estimates run as high as 70% [146]. Yet relatively little work has attempted ASR on tonal languages, with most attempts solely focusing on high-resource languages like Mandarin Chinese [13, 69, 150, *inter alia*] and Vietnamese [89]. Exceptions include Adams et al. [8] who developed ASR systems for Chatino and Yongning Na – both low-resource tonal languages – and obtained promising results; but no one, to our knowledge, has yet attempted to build on this work by applying more recent neural approaches that exploit cross-lingual signals or unsupervised pre-training as used in [24].

In our previous work, the allophone mapping layer was provided in advance, based on a resource, called Allovera [103], which we created in coordination with linguists, relying on language descriptions in the Illustrations of the IPA. In this project, we will also allow for this layer to be learned: we expect that random initialization along with a sparsity-encouraging loss will lead to interpretable learned mappings that will resemble the manually curated ones and can serve as usable suggestions for further linguistic analysis by specialists. We will additionally extend the model with a layer, similar to the allophone layer, that will learn a phoneme-to-grapheme mapping. This will allow the model to be trained on data with orthographic transcriptions, which, for some languages, tend to be significantly easier to obtain.

**Multiple Levels** Our proposed ASR model natively incorporates different levels for representing speech, from concrete (phones) to abstract (phonemes and potentially graphemes). It additionally does so in an interpretable manner, as it allows the incorporation of explicit mappings between those layers if they are available (e.g., manually defined), or automatic discovery of such mappings if not.

We will also explore the incorporation of information from





other representation levels. In particular, we will investigate how additional signals from existing translations of the audio utterance can be leveraged to boost downstream performance and robustness.

In previous work [17], we showed that translation information provided as additional input to a multi-source neural encoder-decoder model (see figure above for a visual representation of a multi-source model) can lead to improved performance. We evaluated our models on three languages (Ainu, Mboshi, and Spanish) and observed reductions in character error rate (listed in the Table) as well as improved robustness to ambient noise – the CER reductions were larger for noisier audio files. This study showed that we can improve ASR by utilizing additional signals. At the same time, it retained some deficiencies that we intend to address through our work in this project. First, our previous model requires the availability of translations both during training and inference. However, for most data collections, this is rarely available. We will modify the model’s architecture and training/inference process to relax this requirement and make it functional with incomplete data. Second, our previous model was trained from scratch. Since then, self-attentive approaches like the Transformer [137] that take advantage of pre-training on other languages or even unsupervised objectives have been shown to surpass previous ones; we will explore such methods.

Character Error Rate			
Model	Ainu	Mboshi	Spanish
ASR	40.7	29.8	52.0
+translation	40.6	28.6	37.6

## 2.2 Inventory Discovery

**Tool** This tool will operate at the level of complete collections. The model will receive all utterances and any annotations from other tiers, and produce a suggestion for a *structured* phone, phoneme, and tone inventory. The model will treat the inventory as a latent variable to be discovered, and will incorporate notions of *dual learning*. We expand further on the methodology below.

**Multiple levels** Despite a long line of work on word discovery directly from speech (manifested primarily in the Zero Resource Speech Challenges [138, 50, 51]), very little effort has focused on the discovery of discrete subword units. In addition, as Gutkin et al. [65] highlight, only *phonemic* multilingual structured inventories are available (e.g. Phoible [100] or Panphon [102]); the desired coupling with phonetic information is lacking.

It was only the Zero Resource Speech Challenge in 2020 [52] that shifted the focus to subword acoustic units [26, 79, 38]. A first approach is one that first segments audio into phoneme-like segments, and then clusters them using articulatory features [106, 105]. Other approaches employ an autoencoding scheme, with most methods using vector-quantized variational autoencoder (VQ-VAE) [135] and add inductive biases to tailor it to this task [71, 101, 136, 130, 64, 42], primarily to enforce sparsity. Other work adopts Bayesian approaches [147, 110] extended with neural components [58] or combine information from other modalities such as visual context [67, 68].

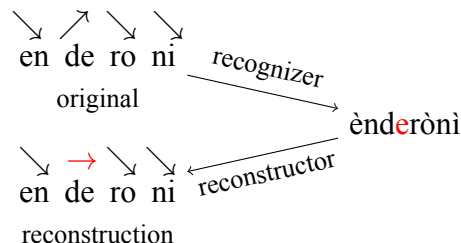
Most work on tone recognition relies on the *tilt* toolkit [128], which employs a combination of features and heuristics. Recent literature is very sparse: Lugosch and Tomar [87] use a convolutional encoder with CTC [63], while Li et al. [83] use an autoencoding scheme to cluster F1 contours. However, both only study Mandarin and Cantonese and rely on large amounts of data.

In this project, we will rely on our universal phone recognizers for an initial phone inventory discovery. We will then turn to discovering the tonal inventory of a language from raw data, by modeling how tones are realized in context. In particular, we will employ an auto-encoding scheme, similar to the VQ-VAE models for phone/subword unit discovery. For a given utterance, a recog-



nizer (an “encoder,” similar to the one we describe in §2.1) will produce the latent transcription consisting of the phone/tone sequence, while the reconstructor (“decoder”) will attempt to reconstruct the input given that latent representation. The advantage of this model is that it does not require any ground-truth annotations, as it can rely on the reconstruction loss: the more accurate the latent representations, the better the reconstruction will be.

The figure on the right represents a simplistic visualization of the model, using an example from Maasai. The hypothesized tonal inventory conflates a rising with a level tone, and incorrectly reconstructs the F0 contour of the second syllable.



This model will be coupled with contextual information from across the corpus and with phonological contrast predictor methods, along with our methods from §3.

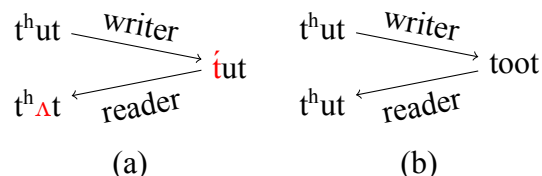
In such approaches, we can also use the existing phonological cross-linguistic databases, such as Phoible or Panphon, to provide an important source of validation for combining various features into structured inventories and for determining which combinations of hypothesized phonemes are admissible, given the existing inventories on well-studied languages.

### 3 Orthography

In this section, we describe our approach to the challenge of unwritten languages, which may comprise as many as half of the world’s languages [1]. The existence of orthographic conventions and written materials, especially for pedagogical purposes, is one of the criteria commonly used to assess language vitality [109]. While the model proposed in the previous section will transcribe audio to IPA, whether the language has a standard writing system or not, orthographic transcription may also be desirable, notably when communities specifically request it, either for use in text messages or social media or for community-wide pedagogical purposes. We propose to explore models that can assist with the creation of an orthography that can then be plugged into the transcription model of the previous section. Those suggestions would be designed to help linguists assist communities in the development of usable and easily learnable orthographic conventions that would allow them to contribute to community-driven language documentation.

We propose an autoencoder setup, in which two models are trained jointly: a “writer” which converts phoneme strings (in IPA) to graphemes, and a “reader” going in the other direction. The models would be trained together on phoneme strings only, and training would balance reconstruction error (that is, how well the reader can reconstruct the phonemes given to the writer, or how well the writer can reconstruct the graphemes given to the reader) against a measure of writing-system complexity (to prevent the model from creating an “orthography” that is just a copy of the IPA).

For example, the figure at right shows a writer-reader setup; in (a), the hypothesized grapheme string uses two different letters for the allophones /t/ and /t<sup>h</sup>/, and incorrectly reconstructs /u/ as /ʌ/, whereas in (b) the grapheme string is more parsimonious and the reconstruction is correct.



Reconstruction error is a commonly-used loss function (e.g., [134]), but measuring writing-system complexity will require more innovation. We want to minimize the amount of information in

written representations; for example, the writer need not learn to distinguish the English allophones /t/ and /t<sup>h</sup>/ because the reader can reliably recover this distinction from context. But a naive measure of written information might lead the model to compress language in unnatural ways. We want to encourage *locality*, which means that, all other things being equal, the mapping between phonemes and graphemes should be order-preserving and rely on as little context as possible.

The reader and writer will both be neural sequence-to-sequence models, but the exact type of model must be determined experimentally. The choice of model is one way of incorporating the biases described above. On the one hand, it should be simple in order to encourage locality. On the other hand, it should be powerful enough to model the phonological and morphological processes that we want the orthography to abstract away from, and even to model some syntax and lexical distinctions – for example, in Spanish, *sí* ‘yes’ and *si* ‘if’ are pronounced nearly the same, but the writer model would need to use syntactic context to generate the right spelling. In our work on Romansh Tuatschin [139], we found that native speakers did incorporate such additional graphemic distinctions spontaneously in the spelling of shorter, high-frequency words (but were more likely to adopt less specific spelling conventions for longer, lower frequency words).

Since the data would consist only of phoneme strings, the grapheme strings must be a latent variable, and the training objective function must sum over all possible grapheme strings, or at least over a random sample of them. One possibility would be for the writer to be a biLSTM-CRF [72], a model commonly used for sequence labeling that computes a distribution over grapheme sequences in the form of a weighted finite automaton. Since it would be intractable to run the reader on all possible grapheme sequences, we could approximate by computing a distribution over graphemes at each position and then use this distribution to calculate an expected grapheme-embedding. The reader would then run on the sequence of expected grapheme-embeddings.

**Multiple levels** It would be easy to give the writer access to the original speech signal, if available. The hope would be to help the orthography model compensate for errors or uncertainties in the ASR model. It would also be easy to give the model access to morphological or even syntactic analyses, which would encourage it to give *atom* and *atomic* related spellings because of their related morphology, even though their pronunciations are in fact different.

We also want to train the writer-reader together with a translation model. This should help the model to abstract away from phones more reliably; for example, if the same word is realized in slightly different ways by different speakers, the fact that they translate to the same word would help the model to map both realizations to the same spelling.

**Multiple languages** We propose to design the model so that it can also be trained on data from other languages that are written in the same script (but possibly not related) and languages that are closely related (but possibly written in a different script). The former is important so that the model can take into account common grapheme-phoneme relationships; for example, Arabic-script ﺏ corresponds to IPA /b/ (or something close to it) across many unrelated languages. The latter (related language, possibly different script) is important so that the model can potentially learn phonological rules that are shared between the two languages; for example, /v/ and /w/ are allophones in both Hindi and Urdu and also written using the same grapheme in their respective scripts.

**Tools** If one (or more) of the methods proposed in this part is successful, it will be released as open-source software. It will take as input a corpus of IPA-transcribed speech and possibly some parallel IPA-orthography data, and produce a model that can be used to convert new data from IPA to orthography. The latter will take the form of an ELAN plug-in using the LEXAN API.



## 4 Morphology

The previous two sections of our proposal revolve around transcription, which by broad consensus is the most important bottleneck in language documentation. But transcription is just the beginning of the linguist’s work; analyzing the transcribed data is the next step, which can also be sped up and aided by computational tools. As above, our proposed two tools will operate at different levels. One will perform glossing (lemmatization and morphological analysis) at the utterance level. Another tool will operate at the corpus level to perform morphological system induction.

### 4.1 Automatic Glossing

**Tool and modeling approach** Glossing can be viewed as the task of joint lemmatization and morphological analysis. Morphological analysis in turn involves the segmentation of words into morphemes and their labelling with meaningful tags that represent grammatical information.

Manual segmentation and morphological analysis is still a time- and labor-consuming work, as it requires linguistic training. Therefore, language documentation projects typically have a gap between the amount of material recorded and archived and the amount of data that is thoroughly analyzed with morphological segmentation and glosses [121]. **This gap can be filled using automatic approaches, which could at least accelerate the annotation process by providing high-quality first-pass annotations.** Previous approaches to automatic gloss generation include manual rule crafting and deep rule-based analysis [27, 123], treating the glossing task as a classification problem focusing only on the morphological tags [97] and requiring a lexicon for stems [118], and using models based on Conditional Random Fields (CRF) integrated with translation and POS-tagging information [91]. Other methods have combined machine labeling and active learning for creating IGT [111, 25]. In contrast, we will build models that will rely on modern neural systems for the automatic glossing task, without requiring any additional components or making unrealistic assumptions regarding data or NLP tool availability for low-resource languages.

**Multiple levels** Linguistic collections may contain rich information that can be beneficial for gloss generation. In line with the main theme of our proposal, we will focus on incorporating information from translations. As Figure 3 shows, the stems/lemmas in the analysis are often implicit in the translation, while the grammatical tags could be derived from the segments in the transcription. The information from the translation can ground the gloss generation, and allow a system that properly takes it into account to generalize to produce lemmas or stems unseen during training.

Morpheme Accuracy				
Model	Lezgian Tsez Arapaho			
Baseline	13%	25%	18%	
Neural Model	35%	35%	56%	
+ translations	<b>53%</b>	<b>40%</b>	<b>58%</b>	

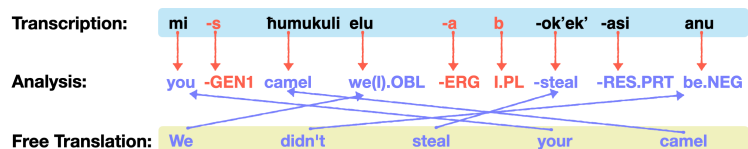


Figure 3: A Tsez IGT example, showing how combining information from both the transcription and the translation can aid in deriving the information in the analysis.

In a pilot study [149], we incorporated this additional translation information using multi-source neural models that encode both the utterance’s transcription and translation, in order to produce glosses. These models were conceptually very similar to those described in §2.1, additionally incorporating inductive biases such as length control suitable for the glossing task. We tested our approach on three morphologi-

cally rich low-resource languages: Arapaho, Tsez, and Lezgian, with training data varying from a few hundred to a few thousand sentences. Incorporating information from the translations significantly improved performance compared to both neural and non-neural baselines (see Table with highlighted results from our pilot study). Also, training a single model for Tsez and Lezgian (two genetically related languages in the Caucasus) enabled cross-lingual transfer and improved results.

This pilot study provides very encouraging indications that our proposed automatic glossing tool can become accurate enough to be able to help linguists with analysis, but we propose additional improvements to the model. **First, we will explore the combination of the multi-source model and a multi-task model by taking both word-level transcription and translation as input, to generate morphologically segmented transcription and morpheme-level glosses. In addition, we will modify the architecture to relax the requirement for transcriptions and translations in the input, in a manner similar to the one proposed in §2.1.** Based on our pilot study findings, we will also investigate adding additional structural biases for attention (such as encouraging monotonic attention [43, 12, 40] and coverage loss [133, 92]).

**Multiple languages** Unfortunately, very small amounts of data are glossed in each language, and quite often the glosses do not follow the same conventions. On a more positive note, projects such as ODIN [144, 145] have managed to aggregate several IGT examples in hundreds of languages through web-scraping [62]. In order to leverage all available data and train multilingually, additional components are required. We propose the following steps, which will allow us to utilize cross-lingual learning for glossing:

1. Develop an automatic tool that will normalize the available IGT data to the Leipzig Glossing Rules standard [32] This will require encoding the Leipzig glossing rules in a machine-readable format and identifying, at a minimum, the most common variations over the data (such as using any of P, P1, PL as a plural marker). This tool will also function as an automatic checker of whether the linguist’s annotation are consistent and adhere to the Leipzig glossing rules.
2. “Standardize” all ODIN data using our normalization tool.
3. Map the morphological tags used in ODIN’s IGT to those used by large multilingual collections with morphological analysis used by the NLP community, namely the Universal Dependencies (UD) corpora [108] and the Universal Morphology (UniMorph) project [127, 76].

Combining data from all available sources (ODIN, UD, UniMorph, the linguists’ annotations), we will be able to train genuinely multilingual models. Such work has already been done within each of these collections. For instance, Kondratyuk and Straka trained a single model on 75 languages for parsing UD [77], while PI Anastasopoulos trained multilingual inflection models on UniMorph [18]. Both show significant improvements for low-resource languages. However, no work so far has combined information across all these data sources. We believe that this will substantially improve performance in low-resource settings and make the models usable for language documentation.

## 4.2 Morphology Induction

**Tool and Methodology** This corpus-level tool will help the linguist tackle the task of unsupervised morphology induction (discovery of paradigms and classes) straight from data.

NLP researchers have devoted significant amounts of time trying to automatically reproduce the linguist’s work. Most approaches have focused on inflection classes, typically first constructing possible inflection tables and then clustering them as a second step. Older works typically only relied on orthographic patterns [99], while systems integrate a semantics view of the words through

subword features (such as the MorphoChains system [107], which can be extended to further abstract over spelling differences for similar phonemes [29]) or word embeddings [124]. More recent works integrate neural inflection models and cluster using embeddings [59] or using edit-tree operations [73].

Taking a more linguistics-oriented view, our approach will be based on the work of Beniamine et al [28], which uses sets of paradigmatic alternations, defined as form alternations between individual cells (e.g., 1SG.PRES: -o and 2SG.PRES: -as for verb A vs. 1SG.PRES: -o and 2SG.PRES: -s for verb B), to group words into *microclasses* (words with the exact same alternations). They then cluster microclasses into incrementally more abstract *macroclasses* based on similarities across microclasses. They use minimum description length (MDL) as a cut-off for macroclass clustering: when the overall description length of the macroclass system stops decreasing (or starts increasing), the optimal set of inflection classes representing the inflectional system is considered attained.

At the core of the work of Beniamine et al. is that a grammar description not only has various granularities (micro- and macroclasses), but each description needs to be as parsimonious as possible, in order to be useful. We will encode these intuitions as part of our modeling approach, and adapt their technique to utilize multiple levels of representation and cross-lingual information.

**Multiple levels** Beniamine et al.’s technique relies on knowledge of feature-form pairs and inventories of complete paradigms. We will relax these two requirements to match the reality of endangered languages, by extending their technique.

First, we will use translations and syntactic distributions to inform the inference of feature-form pairs: instead of using pre-defined features, we will identify forms that are translated consistently and assign them either the features from the translation language as a proxy or some abstract unique feature identifier. Since the language that is being analyzed might in fact have more or fewer featural distinctions than the language used in the translation, those features can be refined by using additional distributional properties derived from the syntactic contexts of the forms.

Second, we will generalize this technique for data directly collected from corpora, which will not contain full paradigms for all words. The techniques developed for form induction above could suggest candidate forms for incomplete paradigms by cross-referencing observable pair-wise alternation patterns over partially attested paradigms. Next, we will integrate this technique with embedding-based approaches such as those of Erdmann et al. [59] and neural feature extractors.

Third, we will extend our models to incorporate information from other annotation tiers. One option will be to use translations in a manner similar to the multi-source models described in the previous sections. Another option will be to take advantage of acoustic and phonological information (that can be extracted from raw audio using methods outlined in §2), which has been ignored, to our knowledge, by all recent approaches. To give a simplistic example from English, consider the word “address”. Simply relying on orthographic information, a model might be inclined to incorrectly classify all its occurrences in a single class; instead, both syntactic distributions over a corpus as well as stress information can guide a model to correctly classify the occurrences in two classes: as a verb (stress in the ultimate syllable) or as a noun (stress in the penultimate syllable). Also, in a language like San Juan Q’ahije Chatino and other Oto-Manguean languages, tone can encode morphological information such as person or number [45].

Last, we will explore another orthogonal research direction of treating the morphological analysis as a latent variable, under the *dual learning* framework we have already discussed in sections §2 and §3. In this framework, we will combine the automatic glossing tools (analyzers) with their

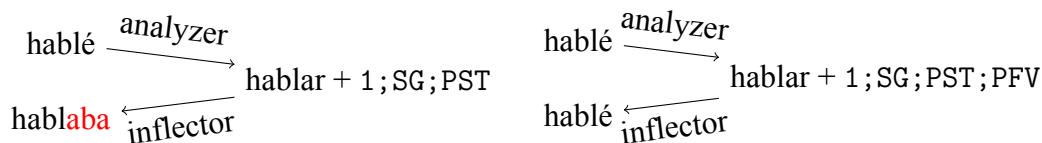


Figure 4: Morphology induction using a reconstruction loss, using a Spanish example. (a) The hypothesized grammar does not include values for aspect (perfective, imperfective), and the inflector produces the wrong form. (b) The correct model includes all required information.

inverse tools (inflectors) which receive the analysis and attempt to reproduce the original forms. A correct induction of the appropriate number of inflection classes and morphological features will be both as parsimonious as possible and guide the analyzers and the inflectors to perform less errors. A visualization of this schema is outlined in Figure 4.

## 5 Case Studies

This project will aim to be as universally applicable as possible, but to confirm the suitability of our tools for endangered languages, we plan to test them on (at least) three typologically diverse languages. We list here the languages that we will work on, along with the available data in them and the case studies we will perform on them. Any annotations created in these case studies will be released publicly in a format suitable both for computational and linguistics research, and will be archived with the Pangloss collection (see Data Management Plan).

### 5.1 Mapudungun

Mapudungun (iso 639-3: *arn*) is an indigenous language of the Americas, with an estimated 100–200 thousand speakers in Chile and 27–60 thousand speakers in Argentina [151, 41–3]. It is an (isolate) Araucanian language and is classified as definitely endangered by UNESCO [104]. Although its noun morphology is relatively simple, its verb morphology is highly agglutinative and complex, with some analyses providing as many as 36 verb suffix slots [122].

**Resources** Recently, PI Anastasopoulos was part of an effort coordinated by CMU LTI’s Latin America programs that released a large resource in Mapudungun.<sup>2</sup> The resource comprises 142 hours of spoken Mapudungun that was recorded during the AVENUE project [82] from 2001 to 2005, transcribed and translated into Spanish. The corpus covers three dialects of Mapudungun: about 110 hours of Nguluche, 20 hours of Lafkenche and 10 hours of Pewenche.

**Output** The available resource includes the core requirements of our project: speech, transcriptions, and translations. As such, we can evaluate all approaches that use translations to better model speech, orthography, or morphology. In order to evaluate morphology, we plan to create and release a small Mapudungun treebank and a collection of verb paradigms.

### 5.2 Yongning Na

Yongning Na is a Sino-Tibetan language of Southwest China, spoken in and around the plain of Yongning in Yunnan province (Glottolog: *yong1270*, ISO 639-3: *nru*; for short, the language is called “Na” below) classified as critically endangered by UNESCO [104]. Na is tonal, with three tonal levels (High, Mid and Low) and a total of seven tone labels. The total number of speakers is

<sup>2</sup>Project page: <http://tts.speech.cs.cmu.edu/mapudungun/>.

estimated to be below 50,000, but cross-dialect variation within Na is high [48], so that the number of people to whom the Na dialect of the Yongning plain is intelligible is likely to be much lower.

**Resources** Resources include a reference grammar [85], a book-length description and analysis of the tone system [95], and a Na-English-Chinese-French dictionary [94]. Alexis Michaud has been documenting Na since 2006, in the process collecting resources that include high-quality audio that has been substantially phonemically transcribed and is also partially glossed and translated to English and French. The corpus is available through the Pangloss collection [93] and consists of around 100 narratives, constituting 11 hours of speech from one speaker in the form of traditional stories, and spontaneous narratives about life, family and customs [95].

**Output** The corpus has been used for some of the few studies on low-resource tonal language transcription [47, 9]. As such, it allows us to directly compare our proposed methods with previous approaches. The partial availability of translations and the fact that Na is tonal make this collection the ideal testbed for all our case studies. For instance, we can compare our tonal inventory discovery approaches with the analysis of Michaud [95]. In addition, we will coordinate with Michaud to study the morphology of Na, utilizing our morphological class system discovery methods.

### 5.3 Griko

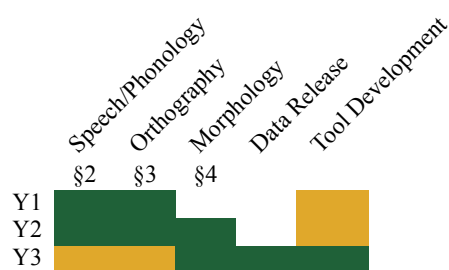
Griko is a Greek (Indo-European) dialect spoken in southern Italy, in the Grecia Salentina area southeast of Lecce. Referred to as *Italiot Greek* along with another endangered Italo-Greek variety spoken in Calabria, it is considered severely endangered by UNESCO [104]. Griko is only partially mutually intelligible with modern Greek, and unlike other Greek dialects, it uses the Latin alphabet. Fewer than 20,000 people (mostly people over 60 years old) are believed to be native speakers [70, 49]; unfortunately, this number is quite likely an overestimate [37].

**Resources** Resources in Griko are very scarce. The first grammar of the language was composed by the German scholar Gerhard Rohlfs [117], followed by others [74]. Recently, PI Anastasopoulos released a corpus of Griko narratives [21]: it contains 114 narratives originally collected by Vito Domenico Palumbo (1854–1928), the most noted Griko scholar [112, 113]. The narratives were further annotated with translations in Italian, and partly annotated with part-of-speech tags. In addition, PI Anastasopoulos has been part of the creation of the only Griko *speech* corpus available online [80], consisting of about 20 minutes of speech in Griko, along with transcriptions, morphosyntactic tags, glosses, and text translations into Italian.

**Output** The availability of speech, transcriptions, analysis, and translations allows us to perform speech recognition, orthography, and morphology case studies. As part of this project, we will create additional morphological analyses from the Griko narratives corpus and study the inflectional classes of Griko verbs. We will also create a Griko treebank and a verb paradigms collection.

## 6 Work Plan

The work in this proposal will be carried out by PIs Anastasopoulos, Walther, and Chiang, who will be jointly in charge of general direction and implementation. Walther will provide direction from a linguistics perspective as she has significant field experience in language documentation, while Anastasopoulos and Chiang will provide the engineering and





computer science perspective. This project takes advantage of the unique abilities of each PI, and duties will be split among PIs and graduate RAs. All members will collaborate closely on all project aspects, meeting regularly to decide general research directions. In order to mitigate risks of specific elements moving slower than expected, we will also be flexible in the timing below.

A graduate student at GMU, co-advised by Anastopoulos and Walther, will be in charge of implementing the core algorithms in the domain of Morphology. A graduate student at ND, advised by Chiang, will be in charge of implementing the core algorithm in the domain of Orthography. Both RAs will be involved with Speech-related methods implementation, as well as with the development of the ELAN plugins. We will coordinate via monthly videoconference calls, and (after the COVID-19 threat subsides) we will have annual project meetings at either Notre Dame or GMU.

**Preliminary Studies** We have already begun a significant amount of preliminary work in the direction of this project. Anastopoulos has released the Griko and Mapudungun corpora that we will work on. His work with Chiang included the pilot studies incorporating translation into speech recognition. Recently, Anastopoulos carried out the pilot study on automatic glossing, which will be presented at COLING 2020.

**Year 1** We will develop the first iteration of our multi-level ASR system (§2.1), including the incorporation of tone recognition. Concurrently, we will use already-available Allosaurus ASR models to begin investigations of inventory discovery and orthography studies (§3). This will involve the necessary infrastructure for dual learning experiments, which will facilitate research on all subcomponents of our project. In the second part of the year we will also get started with AI Helper toolkit development, by first creating plugins that use the existing Allosaurus models.

**Year 2** We will continue research on speech and orthography, focusing particularly in tone recognition and inventory discovery (§2.2). We will also start research related to morphology, by first focusing on automatic glossing (§4.1). In the second part of year 2 we will additionally aim to finalize the ASR and orthography plugins for our AI Helper toolkit.

**Year 3** We will continue to refine the methods developed in the previous years, in order to further improve the underlying methods, with an additional focus on cross-lingual learning. We will continue extending unsupervised morphology induction systems to use translations and phonology (§4.2). We will also put a large effort into providing high-quality transcriptions for untranscribed Na audio files, and we will release our findings on Griko and Mapudungun inflection classes. Early in the year, we also plan to do an official public release of the complete software package, and collaborate with potential early-adopting users of our AI Helper toolbox to shape the direction of our development and build a user base.

## 7 Intellectual Merit

The intellectual merit of this proposal lies in the development of novel methods and computational tools that are better suited for the endangered language documentation pipeline, as they explicitly incorporate different linguistic representation levels and cross-linguistic information.

We will incorporate limited translation information in order to improve automatic speech recognition and morphological analysis. We will also model various levels of linguistic representations (phone, phoneme, grapheme, translation) as latent variables in neural models, which can be extracted in an interpretable manner. We will additionally attempt to uncover the phoneme and tone inventories of a language directly from data and model subtle allophonic differences in order to

suggest an orthographic system. Last, we will investigate models that induce the morphological system of a language based only on small amounts of potentially translated data.

Each of these difficult research questions not only requires an answer to create better computational tools for linguists, but also because they are interesting in their own right, academically and for other applications.

## 8 Broader Impacts

This project will produce software tools that will simplify and accelerate the work of documentary linguists, either in the field or at home. Where possible, they will be implemented as ELAN plug-ins to make them as easy as possible to integrate into existing documentation workflows. All software created will be released under open-source licenses. Additionally, all automatically-generated annotations produced in our case studies (§5) will be released publicly under open licenses and archived with the Pangloss collection.

**ICLDC Workshop** We will propose to organize a workshop at ICLDC 2023 (the prime venue for language documentation), to publicize and showcase our tools. Additionally, this will allow us to receive important feedback from linguists in order to further improve the tools.

**Broadening Participation** The project will fund two PhD students, and given the importance of diversity to the project, we will especially strive to recruit a diverse team of PhD students. We will also involve undergraduate or early-stage graduate students, and ensure that the PhD students involved are assigned some supervisory duties, facilitating their development as educators. We are especially considering involving undergraduate students through the programs run by the GMU Office of Student Scholarship Creative Activities and Research (OSCAR).

**Inter-disciplinarity** This project will further promote the co-operation of the linguistics and computer science communities. We will disseminate results through papers in both linguistics conferences and journals (e.g., ICLDC) and NLP conferences and journals (e.g., ACL, EMNLP), as well as ComputEL, the ACL workshop for computational approaches for endangered languages.

## 9 Results from Prior NSF Support

**RI: Small: Language Induction meets Language Documentation: Leveraging bilingual aligned audio for learning and preserving languages.** Award 1464553, PIs D. Chiang and S. Bird. Total \$470k over four years. Publications: 1 journal, 15 conference, 7 workshop papers. **Intellectual Merit:** Developed new state-of-the-art models for diverse tasks on low-resource and endangered languages: speech-to-translation alignment [55, 19, 14], unsupervised word-spotting [20], joint speech transcription and translation [16, 15], speech transcription of tonal languages [8], bilingual lexicon/embedding induction [56, 4, 5, 6, 57, 7], part-of-speech tagging [53, 21], dependency parsing [54]. **Broader Impacts:** Supported several PhD students including Antonios Anastasopoulos at Notre Dame. Developed an open-source educational deep learning toolkit Penne [39], software and workflows for collecting bilingual audio [35, 34], speech transcription [96], collaborative language documentation [30], and speech elicitation [31]. Created datasets of audio with translations [148] and non-native text with translations [88].